

EVOKE: Elevating Chest X-ray Report Generation via Multi-View Contrastive Learning and Patient-Specific Knowledge

Qiguang Miao *Senior Member, IEEE*, Kang Liu, Zhuoqi Ma, Yunan Li, Xiaolu Kang, Ruixuan Liu, Tianyi Liu, Kun Xie, and Zhicheng Jiao *Member, IEEE*

Abstract—Radiology reports are crucial for planning treatment strategies and facilitating effective doctor-patient communication. However, the manual creation of these reports places a significant burden on radiologists. While automatic radiology report generation presents a promising solution, existing methods often rely on single-view radiographs, which constrain diagnostic accuracy. To address this challenge, we propose EVOKE, a novel chest X-ray report generation framework that incorporates multi-view contrastive learning and patient-specific knowledge. Specifically, we introduce a multi-view contrastive learning method that enhances visual representation by aligning multi-view radiographs with their corresponding report. After that, we present a knowledge-guided report generation module that integrates available patient-specific indications (e.g., symptom descriptions) to trigger the production of accurate and coherent radiology reports. To support research in multi-view report generation, we construct Multi-view CXR and Two-view CXR datasets using publicly available sources. Our proposed EVOKE surpasses recent state-of-the-art methods across multiple datasets, achieving a 2.9% F_1 RadGraph improvement on MIMIC-CXR, a 7.3% BLEU-1 improvement on MIMIC-ABN, a 3.1% BLEU-4 improvement on Multi-view CXR, and an 8.2% $F_{1,mic-14}$ CheXbert improvement on Two-view CXR.

This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.

The work was jointly supported by the National Science and Technology Major Project under Grant No. 2022ZD0117103, the National Natural Science Foundations of China under Grant No. 62272364, the provincial Key Research and Development Program of Shaanxi under Grant No. 2024GH-ZDXM-47, the Research Project on Higher Education Teaching Reform of Shaanxi Province under Grant No. 23JG003, and the Fundamental Research Funds for the Central Universities under Grant No. ZYTS24090. (Corresponding authors: Kang Liu and Zhuoqi Ma)

Qiguang Miao, Kang Liu, Zhuoqi Ma, Yunan Li, Xiaolu Kang, and Kun Xie are with the School of Computer Science and Technology, Xidian University, Xi'an, Shaanxi 710071, China (e-mail: {qgmiao, zhuoqima, yunanli, xiekun}@xidian.edu.cn; {kangliu, 22031212472}@stu.xidian.edu.cn).

Ruixuan Liu is with the Department of Orthopedics, Shanghai Key Laboratory for Prevention and Treatment of Bone and Joint Diseases, Shanghai Institute of Traumatology and Orthopedics, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China (e-mail: osteoliu@163.com).

Tianyi Liu is with the Department of Neurology, The First Affiliated Hospital of Harbin Medical University, Harbin, Heilongjiang 150001, China (e-mail: liutiyi92@163.com).

Zhicheng Jiao is with the Department of Diagnostic Imaging, Brown University, Providence, RI 02903-4923, USA (e-mail: zhicheng-jiao@brown.edu).

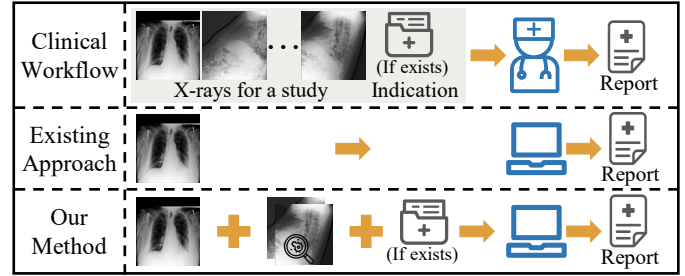


Fig. 1. A comparison between existing methods and our proposed approach for chest X-ray report generation reveals that existing methods rely on single-view images, whereas our approach leverages multi-view radiographs and patient-specific indications.

Index Terms—Chest X-ray report generation, multi-view contrastive learning, patient-specific knowledge.

I. INTRODUCTION

RADIOLOGY reports, crafted by experienced radiologists, meticulously document imaging findings from examinations such as X-rays, PET scans, and CTs, detailing abnormalities and initial diagnostic conclusions. These reports deliver vital imaging insights that empower physicians to develop efficient patient treatment strategies [1], [2]. However, the manual writing process is time-consuming and demands substantial expertise, posing challenges in meeting the increasing demands of modern healthcare [3], particularly in regions with limited medical resources.

Automatic chest X-ray report generation aims to produce detailed and accurate free-text reports based on multi-view radiographs, helping radiologists improve diagnostic efficiency and consistency by providing high-quality draft reports. Limitations like X-ray equipment constraints and the complexity of human anatomical structures can prevent a single-view radiograph from achieving optimal imaging quality and adequately displaying the overall anatomical structure. As a result, multi-view imaging examinations, such as postero-anterior (PA), antero-posterior (AP), and lateral views, are essential for accurate diagnostics and personalized treatment. In existing public datasets (e.g., MIMIC-CXR [4] and MIMIC-ABN [5]), each study includes a variable number of radiographs, an associated radiology report, and a patient-specific *INDICATION* (which

describe the patient's symptoms and may sometimes be missing). In the clinical workflow, radiologists typically select key X-rays as the primary diagnostic basis, supplementing them with additional X-rays and the patient's clinical symptoms (i.e., *INDICATION*) to conduct a comprehensive analysis and generate a radiology report, as shown in Fig. 1. However, existing methods [6]–[10] primarily generate reports based on single-view images, neglecting the detailed anatomical information available in multi-view images and the patient's clinical symptoms. This limitation often results in inaccuracies and inconsistencies in the generated reports.

To address this challenge and emulate radiologists' clinical practices, we propose a new framework named **EVOKE**, which incorporates multi-view images and patient-specific knowledge (i.e., *INDICATION*) into the chest X-ray report generation process. More concretely, we present a multi-view contrastive learning method that utilizes a multi-view fusion module to integrate a variable number of X-rays from the same study. This approach enhances visual representation by maximizing semantic correspondences both among multi-view images within a study and between these images and their corresponding report. Additionally, we introduce a knowledge-guided report generation module that effectively integrates available patient-specific *INDICATION* and employs a transition bridge network to mitigate embedding space discrepancies caused by the presence or absence of *INDICATION*. Together, these components facilitate the generation of accurate and coherent radiology reports. We evaluate our proposed EVOKE on the MIMIC-CXR [4], MIMIC-ABN [5], [11], and our curated Multi-view CXR and Two-view CXR datasets. Extensive experiments confirm that EVOKE outperforms recent state-of-the-art methods. Our key contributions are as follows:

- We introduce the EVOKE framework for chest X-ray report generation, which effectively integrates multi-view images and patient-specific knowledge to enhance the accuracy of radiology reports.
- We propose a multi-view contrastive learning method that improves visual representations and flexibly handles a variable number of images per study.
- We present a knowledge-guided report generation module that incorporates patient-specific *INDICATION* and mitigates embedding space discrepancies caused by the presence or absence of *INDICATION*.
- We curate Multi-view CXR and Two-view CXR datasets from two public sources, ensuring that each study includes multiple radiographs. These datasets support research on multi-view report generation, particularly in scenarios with variable or two-view setups.

II. RELATED WORK

A. Image Captioning

Image captioning [12], [13] and radiology report generation [14]–[16] both translate visual information into textual descriptions. The encoder-decoder framework has significantly advanced image captioning by organizing visual features into coherent text, benefiting radiology report generation. However, unlike image captioning, which typically generates concise and

general descriptions, radiology report generation [17] demands detailed, domain-specific, and diagnostically relevant content.

B. Medical Image Analysis Meets Multi-view Learning

Multi-view learning [18], [19] seeks to capture shared and complementary information from diverse perspectives of the same scene, thereby enhancing overall comprehension. In representation learning, [20] develop a multi-view semantic embedding extracted from X-ray reports, CT reports, and visual traits to enhance the clinical accuracy of chest X-ray diagnosis. REFERS [21] introduces a view fusion module that integrates fixed-view visual features, enhancing cross-modal alignment. In medical report generation, FMVP [22] incorporates single radiographs and auxiliary inputs (i.e., disease tags and medical concepts) as multi-view information, using two shared and continuous cross-attention mechanisms [23] to assist in generating informative reports. Given that the IU X-ray dataset [24] primarily comprises studies with two-view radiographs, several approaches [25], [26] have been developed to handle two-view report generation. These methods enhance the clinical efficacy of generated reports and offer valuable insights for advancing multi-view report generation. However, they often struggle with the varying number of views encountered in real-world scenarios. To tackle this issue, we introduce a multi-view contrastive learning method that flexibly accommodates a variable number of images per study.

C. Medical Image Analysis Meets Contrastive Learning

Contrastive learning is pivotal in medical image analysis, particularly in tasks related to medical visual representation learning [27], [28] and report generation [29], [30]. It aligns radiographs with their corresponding reports by minimizing the distance between positive pairs (radiographs and associated reports) while maximizing the distance between negative pairs (radiographs and unrelated reports). Significant advancements have been made in this area. For instance, MedCLIP [31] adopts a semantic matching loss for global alignment between decoupled radiographs and reports. ARL [32] regards knowledge as an intermediary to facilitate semantic alignment between radiographs and their reports. MGCA [27] introduces multi-grained cross-modal alignment at the instance, pathological region, and disease levels for generalized medical visual representation. SEI [33] utilizes global and local cross-modal alignment between radiographs and factual serialization in reports to generate radiology reports. While these methods effectively align radiographs with their reports by extracting visual features from single-view radiographs, they often overlook the benefits of incorporating multi-view information from the same study. In response, our work introduces multi-view contrastive learning, integrating multi-view visual features to improve cross-modal alignment.

III. METHOD

The overall architecture of our proposed EVOKE, shown in Fig. 2, adopts a two-stage training strategy. In Stage 1, we propose a multi-view contrastive learning method that

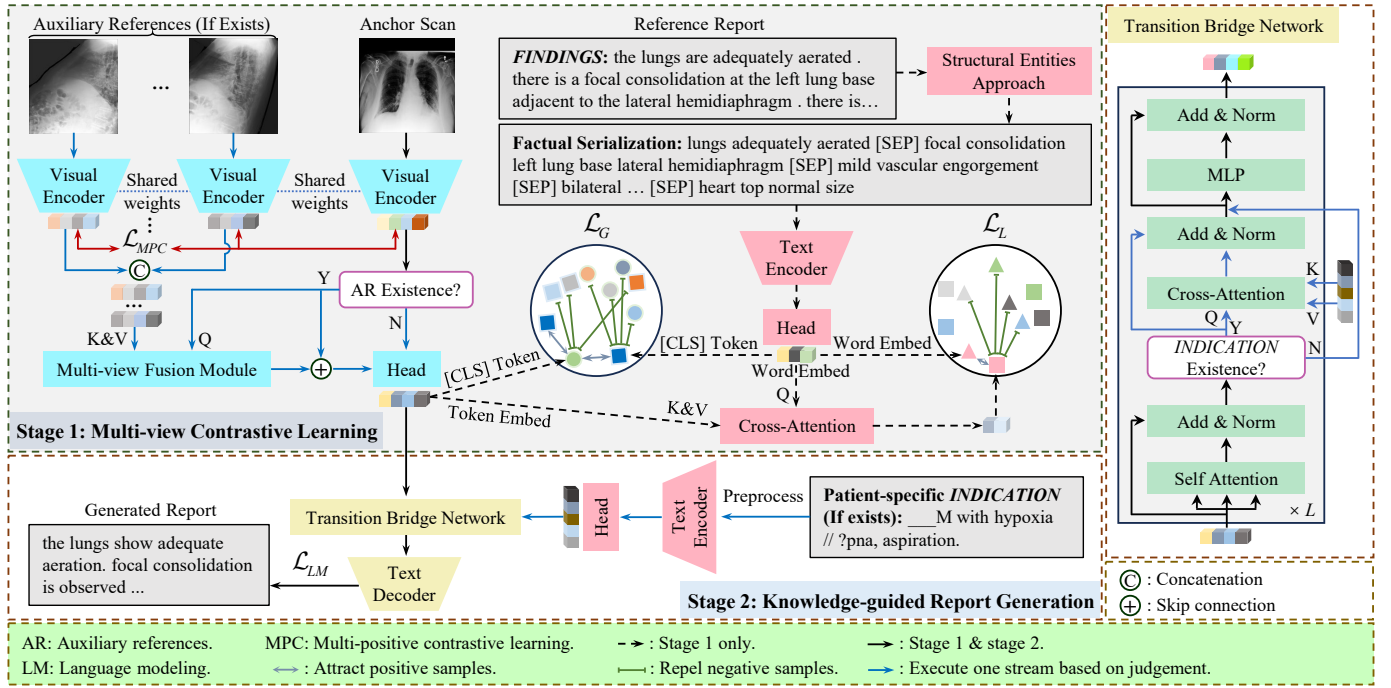


Fig. 2. Illustration of our proposed EVOKE, which comprises a visual encoder, a text encoder, and a text generator. EVOKE employs a two-stage training strategy: Multi-view contrastive learning for representation learning (Stage 1) and knowledge-guided report generation (Stage 2). The model inference is performed solely using Stage 2.

integrates multi-view images to enhance visual representations. In Stage 2, we present a knowledge-guided report generation module that incorporates patient-specific *INDICATION*, equipping the text generator with relevant background information to produce accurate and coherent radiology reports.

A. Problem Formulation

Let $D_{tr} = \{(x_{i,k}, x_{i,\setminus k}, z_i, r_i)\}_{i=1}^n$ be the training set, where n denotes the number of studies and $k \in [1, m_i]$. The i^{th} study includes m_i radiographs (views), a patient-specific *INDICATION* z_i (which may be absent), and a corresponding report r_i . To align with radiologists' clinical practices, we categorize the multi-view images within each study into an anchor scan $x_{i,k}$, serving as the primary diagnostic basis, and auxiliary references $x_{i,\setminus k} = \{x_{i,j} \mid j \neq k, 1 \leq j \leq m_i\}$, which provide additional context and may be empty. The *INDICATION* z_i describes the patient's clinical symptoms; while not directly influencing diagnostic interpretation, it offers valuable contextual insights for radiograph analysis. Our objective is to learn a function $F_\theta(\cdot)$ that maps the input $(x_{i,k}, x_{i,\setminus k}, z_i)$ to the corresponding report r_i using the training set D_{tr} .

B. Multi-view Contrastive Learning Method

We introduce a multi-view contrastive learning method that enhances visual representations by utilizing semantic correspondences both among multi-view radiographs within a study and between these radiographs and their corresponding report. First, we employ multi-positive contrastive learning to align multi-view radiographs, improving visual feature consistency across them. Afterward, we develop a multi-view fusion module to integrate varying numbers of radiographs per

study, generating fused visual features for subsequent cross-modal alignment. Finally, we apply contrastive learning at both instance-wise and token-wise levels, maximizing agreements between these radiographs and their corresponding report.

Visual Features Extraction. We use ResNet101 [34], pre-trained on ImageNet, as the visual encoder. The feature maps from the last convolutional layer of ResNet101 are regarded as the visual features of radiographs, formulated as $V \in \mathbb{R}^{M \times p \times d_1}$, where $M = \sum_{i=1}^B m_i$ denotes the total number of radiographs in the batch. Here, B is the batch size, p represents the dimensions of the feature map, and d_1 corresponds to the number of channels.

Textual Features Extraction. Inspired by [35], [36], we first employ the structural entities approach [36] to extract factual serialization from reports, mitigating noise for subsequent cross-modal alignment. Factual serialization, shown in Fig. 2, is a concise sentence constructed from clinically relevant keyword groups derived from reports. This serialization is then fed into a six-layer pre-trained SciBERT [37] to produce textual features $T \in \mathbb{R}^{B \times k \times d_2}$. Here, k denotes the number of textual tokens and d_2 represents the dimensionality of each token. It is important to note that the textual features T are exclusively utilized in Stage 1.

Multi-positive Contrastive Learning between Multi-view Radiographs. To capture semantic correspondences between multi-view radiographs within the same study, we adopt multi-positive contrastive learning [38]. This method aligns radiographs from the same study while differentiating them from those in other studies, thereby enhancing visual features consistency. We first excluded studies containing only a single-view radiograph from the batch (Notably, these studies are still

used for subsequent cross-modal alignment). For each anchor scan $x_{i,a}$, we compute the contrastive categorical distribution $\mathbf{q}$ to quantify the similarity between $x_{i,a}$ and all other scans $x_{\setminus a}$. Here, $x_{\setminus a}$ comprises both auxiliary references of $x_{i,a}$ and radiographs from other studies. We denote the global visual features, after ℓ_2 normalization, as $\mathbf{v}_{i,a}$ for anchor scans and $\mathbf{v}_{\setminus a}$ for other scans. The $\mathbf{q} \in \mathbb{R}^{K \times (K-1)}$ is defined as:

$$\mathbf{q}_i = \frac{\exp(\mathbf{v}_{i,a} \cdot \mathbf{v}_{\setminus a}^T / \tau_1)}{\sum_{j=1}^{K-1} \exp(\mathbf{v}_{i,a} \cdot (\mathbf{v}_{\setminus a}^j)^T / \tau_1)}, \quad (1)$$

where $K = \sum_{j=1, m_j > 1}^B m_j$ denotes the total number of multi-view radiographs in a batch, and $\tau_1 \in \mathcal{R}^+$ is the temperature parameter. We then construct the ground-truth categorical distribution $\mathbf{p} \in \mathbb{R}^{K \times (K-1)}$, formulated as:

$$\mathbf{p}_i = \frac{\mathbb{I}_{\text{match}}(x_{i,a}, x_{\setminus a})}{\sum_{j=1}^{K-1} \mathbb{I}_{\text{match}}(x_{i,a}, x_{\setminus a}^j)}, \quad (2)$$

$\mathbb{I}_{\text{match}}(\cdot, \cdot)$ is an indicator function that determines whether two radiographs belong to the same study. Finally, the multi-positive contrastive loss is calculated using the cross-entropy between $\mathbf{q}$ and $\mathbf{p}$, denoted as:

$$\mathcal{L}_{MPC} = -\frac{1}{K} \sum_{i=1}^K \mathbf{p}_i \log \mathbf{q}_i. \quad (3)$$

Despite the varying number of radiographs per study, the different number of non-zero elements in $\mathbf{p}_i$ account for this variability. Applying cross-entropy between $\mathbf{q}_i$ and $\mathbf{p}_i$ pushes the multi-view radiographs in the i^{th} study closer together in the embedding space.

Multi-view Fusion Module. Aligning multi-view radiographs with their corresponding reports necessitates effective integration of multi-view visual features. However, the varying number of images per study complicates simple concatenation methods, leading to inconsistent channel dimensions and hindering cross-modal alignment. While averaging and maximization schemes can ensure channel consistency, they sacrifice detailed information. To tackle this challenge, we propose a multi-view fusion module that flexibly integrates a variable number of multi-view images using the scaled dot-product cross-attention mechanism [23], denoted as $\text{ATTN}(Q, K \& V)$. In this module, the anchor scan $x_{i,a}$ acts as queries, while auxiliary references $x_{i,\setminus a}$ serve as both keys and values. Skip connection [34] and layer normalization (LN) are then applied to generate the fused visual features for subsequent cross-modal alignment, formulated as:

$$\tilde{\mathbf{V}}_{i,a} = \text{LN}(\mathbf{V}_{i,a} + \text{ATTN}(\mathbf{V}_{i,a}, \mathbf{V}_{i,\setminus a})) \quad (4)$$

where $\tilde{\mathbf{V}} = \{\tilde{\mathbf{V}}_{i,a} | 1 \leq i \leq B\} \in \mathbb{R}^{B \times p \times d_1}$ ensures consistent channel dimensions across studies while preserving information from both anchor scans and auxiliary references. Notably, we instruct the module to focus on processing one study at a time, flexibly accommodating the variability in the number of radiographs per study.

Instance-wise Alignment Loss. Drawing inspiration from [27], [28], [36], we propose the instance-wise alignment loss

to maximize agreements between multi-view radiographs and their corresponding report in the embedding space. We begin by projecting the fused visual features $\tilde{\mathbf{V}}$ and textual features $\mathbf{T}$ into a unified embedding space d using the two-layer convolutional projection head. The resulting embeddings are denoted as $\tilde{\mathbf{V}} \in \mathbb{R}^{B \times p \times d}$ for visual features and $\tilde{\mathbf{T}} \in \mathbb{R}^{B \times k \times d}$ for textual features. Subsequently, we define the visual-textual feature pair from the same study as a positive pair. To estimate the similarity between multi-view radiographs and their corresponding report, we compute the image-to-text contrastive categorical distribution $\mathbf{q}^{v2t} \in \mathbb{R}^{B \times B}$, given by:

$$\mathbf{q}^{v2t} = \frac{\exp(\bar{\mathbf{v}} \cdot \bar{\mathbf{t}}^T / \tau_2)}{\sum_{j=1}^B \exp(\bar{\mathbf{v}} \cdot (\bar{\mathbf{t}}^j)^T / \tau_2)}, \quad (5)$$

where $\bar{\mathbf{v}}$ and $\bar{\mathbf{t}}$ are global visual and textual features processed by ℓ_2 normalization. Similarly, we compute the symmetric text-to-image contrastive categorical distribution $\mathbf{q}^{t2v}$. The global ground-truth categorical distribution $\mathbf{p}^g$ is defined as:

$$\mathbf{p}_{i,j}^g = \frac{\mathbb{I}_{\text{equal}}(r_i, r_j)}{\sum_{k=1}^B \mathbb{I}_{\text{equal}}(r_i, r_k)}, \quad (6)$$

where $\mathbf{p}^g \in \mathbb{R}^{B \times B}$, and $\mathbb{I}_{\text{equal}}(\cdot, \cdot)$ denotes an indicator function that determines whether two reports are identical. The instance-wise alignment loss is expressed as:

$$\mathcal{L}_G = -\frac{1}{B} \sum_{i=1}^B (\mathbf{p}_i^g \log \mathbf{q}_i^{v2t} + \mathbf{p}_i^g \log \mathbf{q}_i^{t2v}). \quad (7)$$

Our method differs from prior works [27], [28] in two key ways: We utilize factual serialization from reports for alignment instead of complete reports, and we employ multi-positive contrastive loss [38] to learn semantic correspondences between multi-view radiographs and reports, rather than single-positive contrastive loss [39].

Token-wise Alignment Loss. Inspired by [27], [36], we employ the token-wise alignment loss $\mathcal{L}_L$ [36] to learn fine-grained visual features. This loss is achieved through single-positive contrastive learning between token embeddings of visual and textual data. In conclusion, the overall training objective in Stage 1 is formulated as:

$$\mathcal{L}_{\text{pretrain}} = \mathcal{L}_{MPC} + \mathcal{L}_G + \mathcal{L}_L. \quad (8)$$

C. Knowledge-guided Report Generation

Indication Features Extraction. Patient-specific *INDICATION*, as illustrated in Fig. 2, is collected before examinations and describes the patient's clinical symptoms. Although not directly tied to the diagnostic interpretation, it provides valuable contextual information that enhances the model's ability to analyze radiographs. However, de-identification in public datasets often introduces noise into *INDICATION*, such as *---*, *@*, and *-year-old*. We preprocess the data following the approach in [33], which involves noise removal and standardization of gender expressions. The cleaned indications are then fed into the text encoder, initialized with the pre-trained model from Stage 1, to extract indication features.

Split	MIMIC-CXR			MIMIC-ABN			Multi-view CXR			Two-view CXR		
	#Img	#Rpt	%Ind	#Img	#Rpt	%Ind	#Img	#Rpt	%Ind	#Img	#Rpt	%Ind
Train	269,239	150,957	66.4	69,526	34,763	64.6	220,978	100,505	67.1	165,056	82,528	67.9
Val	2,113	1,182	65.4	526	263	62.7	2,299	1,057	71.2	1,778	889	72.1
Test	3,852	2,343	57.3	756	378	56.3	3,947	1,805	67.6	3,000	1,500	68.9

TABLE I

STATISTICS FOR THE TRAINING, VALIDATION, AND TEST SETS ACROSS THE FOUR DATASETS. “#IMG” AND “#RPT” DENOTE THE NUMBER OF RADIOGRAPHS AND REPORTS, RESPECTIVELY. “%IND” REPRESENTS THE RATIO OF *INDICATION*.

Dataset	Method	Year	Input Size	NLG Metrics $\uparrow$						CE Metrics $\uparrow$		
				B-1	B-2	B-3	B-4	MTR	R-L	RG	CX5	CX14
Comparison with single-view methods												
MIMIC-CXR	R2Gen [40]	2020	224	0.353	0.218	0.145	0.103	0.142	0.277	0.207	0.340	0.340
	CMN [25]	2021	224	0.353	0.218	0.148	0.106	0.142	0.278	0.220	0.461	0.391
	CGPT2 [41]	2023	384	0.393	0.248	0.171	0.127	0.155	0.286	-	-	0.442
	MET [42]	2023	-	0.386	0.250	0.169	0.124	0.152	0.291	-	-	0.311
	KiUT [43]	2023	224	0.393	0.243	0.159	0.113	0.160	0.285	-	-	0.321
	SA [35]	2023	256	-	0.184	-	-	-	-	0.228	-	0.394
	FMVP [22]	2023	224	0.389	0.236	0.156	0.108	0.150	0.284	-	-	0.336
	SEI-1 [33]	2024	224	-	0.247	-	0.135	0.158	0.299	0.249	0.542	0.460
	MAN [29]	2024	224	0.396	0.244	0.162	0.115	0.151	0.274	-	-	0.389
	PMRG [44]	2024	224	<u>0.398</u>	-	-	0.112	0.157	0.268	-	-	0.476
	Med-LLM [45]	2024	224	-	-	-	0.128	0.161	0.289	-	-	0.395
	HERGen [46]	2024	384	0.395	0.248	0.169	0.122	0.156	0.285	-	-	-
	EVOKE(Ours)	-	224	0.395	<u>0.262</u>	<u>0.190</u>	<u>0.147</u>	<u>0.167</u>	<u>0.311</u>	0.276	0.557	0.499
	EVOKE(Ours)	-	384	0.408	0.271	0.197	0.151	0.171	0.313	0.278	0.578	0.517
$\Delta \uparrow$	-	-	-	+1.0%	+2.1%	+2.6%	+1.6%	+1.0%	+1.4%	+2.9%	+3.6%	+4.1%
MIMIC-ABN	R2Gen [#] [40]	2020	224	0.253	0.144	0.092	0.063	0.106	0.229	0.179	0.501	0.442
	CMN [#] [25]	2021	224	0.256	0.147	0.095	0.066	0.110	0.230	0.183	0.528	0.460
	EVOKE(Ours)	-	224	<u>0.310</u>	<u>0.185</u>	<u>0.125</u>	<u>0.090</u>	<u>0.127</u>	<u>0.246</u>	<u>0.214</u>	<u>0.535</u>	<u>0.482</u>
	EVOKE(Ours)	-	384	0.329	0.196	0.131	0.093	0.134	0.255	0.220	0.545	0.503
	$\Delta \uparrow$	-	-	-	+7.3%	+4.9%	+3.6%	+2.7%	+2.4%	+2.5%	+3.7%	+1.7%
Multi-view CXR	R2Gen [#] [40]	2020	224	0.359	0.225	0.155	0.114	0.143	0.297	0.255	0.431	0.384
	CMN [#] [25]	2021	224	0.404	0.252	0.170	0.122	0.160	0.311	0.279	0.475	0.416
	EVOKE(Ours)	-	224	<u>0.413</u>	<u>0.274</u>	<u>0.199</u>	<u>0.152</u>	<u>0.174</u>	<u>0.335</u>	<u>0.328</u>	<u>0.515</u>	<u>0.487</u>
	EVOKE(Ours)	-	384	0.415	0.276	0.200	0.153	0.177	0.336	0.329	0.557	0.508
	$\Delta \uparrow$	-	-	-	+1.1%	+2.4%	+3.0%	+3.1%	+1.7%	+2.5%	+5.0%	+8.2%
Comparison with two-view methods												
Two-view CXR	R2Gen [#] [40]	2020	224	0.346	0.219	0.153	0.113	0.141	0.302	0.267	0.413	0.400
	CMN [#] [25]	2021	224	0.387	0.241	0.166	0.122	0.151	0.310	0.268	0.437	0.425
	EVOKE(Ours)	-	224	<u>0.393</u>	<u>0.256</u>	<u>0.184</u>	<u>0.140</u>	<u>0.165</u>	<u>0.322</u>	0.304	<u>0.531</u>	<u>0.501</u>
	EVOKE(Ours)	-	384	0.411	0.270	0.195	0.150	0.172	0.326	<u>0.302</u>	0.547	0.507
	$\Delta \uparrow$	-	-	-	+2.4%	+2.9%	+2.9%	+2.8%	+2.1%	+1.6%	+3.6%	+11.0%

TABLE II

COMPARISONS WITH SOTA METHODS ON MIMIC-CXR, MIMIC-ABN, MULTI-VIEW CXR, AND TWO-VIEW CXR DATASETS. Δ DENOTES THE PERFORMANCE DIFFERENCE BETWEEN EVOKE AND THE BEST PEER METHODS. [#] INDICATES RESULTS REPRODUCED USING OFFICIAL CODES, WHILE OTHER RESULTS ARE CITED FROM THE ORIGINAL WORK. THE BEST AND SECOND-BEST VALUES ARE IN **BOLD** AND UNDERLINED.

Report Generation. As shown in Table I, patient-specific *INDICATION* (i.e., knowledge) may sometimes be absent in certain studies. Consequently, some existing methods either entirely ignore this data [29], [40], [42] or underutilize its potential [33], [47], [48]. To address this limitation, we develop a transition bridge network, illustrated in Fig. 2. This network integrates the available *INDICATION* via a cross-attention mechanism and utilizes the transition “bridge” to handle missing *INDICATION*, thereby reducing embedding space inconsistencies. By uniformly processing cases with and without *INDICATION*, the network supplies patient-specific knowledge to the text generator, facilitating the generation of more accurate and coherent reports. The text generator,

implemented as a memory-driven transformer [40], produces radiology reports in an autoregressive manner, optimized by minimizing the negative log-likelihood as follows:

$$\mathcal{L}_{LM_i} = - \sum_{t=1}^M \log P \left(\tilde{w}_i^t | x_i, z_i, \tilde{w}_i^{j < t} \right), \quad (9)$$

where $\tilde{w}_i$ denotes predicted words generated by the text generator. x_i denotes the multi-view radiographs for the i^{th} study, and z_i corresponds to the *INDICATION*. Here, M indicates the maximum number of generated tokens.

IV. EXPERIMENTS

Model	M/S	Stage 1				Stage 2	NLG Metrics $\uparrow$				CE Metrics $\uparrow$		
		F/R	MPC	G	L		B-2	B-4	MTR	R-L	RG	CX5	CX14
R2Gen	S	-	$\times$	$\times$	$\times$	$\times$ ($\times$)	0.218	0.103	0.142	0.277	0.207	0.340	0.340
(a)	M	-	$\times$	$\times$	$\times$	$\checkmark$ ($\checkmark$)	0.247	0.134	0.158	0.296	0.252	0.545	0.480
(b)	S	F	$\times$	$\checkmark$	$\checkmark$	$\times$ ($\times$)	0.213	0.102	0.142	0.278	0.233	0.512	0.430
(c)	S	F	$\times$	$\checkmark$	$\checkmark$	$\checkmark$ ($\checkmark$)	0.242	0.130	0.158	0.297	0.256	0.548	0.484
(d)	M	-	$\checkmark$	$\times$	$\times$	$\checkmark$ ($\checkmark$)	0.261	0.146	0.165	0.309	0.270	0.551	0.482
(e)	M	F	$\checkmark$	$\checkmark$	$\times$	$\checkmark$ ($\checkmark$)	0.263	0.145	0.166	0.305	0.268	0.540	0.480
(f)	M	F	$\checkmark$	$\times$	$\checkmark$	$\checkmark$ ($\checkmark$)	0.246	0.136	0.157	0.305	0.255	0.513	0.450
(g)	M	F	$\times$	$\checkmark$	$\checkmark$	$\checkmark$ ($\checkmark$)	0.252	0.142	0.162	0.307	0.266	0.536	0.477
(h)	M	F	$\checkmark$	$\checkmark$	$\checkmark$	$\times$ ($\times$)	0.226	0.108	0.148	0.279	0.230	0.539	0.477
(i)	M	R	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$ ($\checkmark$)	0.258	0.142	0.164	0.307	0.266	0.522	0.470
(j)	M	F	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$ ($\times$)	0.259	0.145	0.164	0.312	0.272	0.548	0.488
EVOKE	M	F	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$ ($\checkmark$)	0.262	0.147	0.167	0.311	0.276	0.557	0.499

TABLE III

ABLATION STUDY ON MIMIC-CXR WITH A RESOLUTION OF 224^2 . “M/S” INDICATES A MULTI-VIEW OR SINGLE-VIEW METHOD. “F/R” DENOTES CROSS-MODAL ALIGNMENT USING EITHER FACT SERIALIZATION OR REPORT. “MPC”, “G”, AND “L” INDICATE MULTI-POSITIVE CONTRASTIVE LOSS, INSTANCE-WISE ALIGNMENT LOSS (G), AND TOKEN-WISE ALIGNMENT LOSS (L), RESPECTIVELY. ‘IND (“BRIDGE”)’ DENOTES WHETHER INDICATION IS USED IN STAGE 2 AND IF THE TRANSITION “BRIDGE” IS ADDED FOR HANDLING MISSING INDICATION.

A. Experimental Settings

Datasets. We evaluate our approach on the following datasets: **1) MIMIC-CXR** [4] is a large-scale chest X-ray dataset with free-text radiology reports and varying numbers of radiographs per study. **2) MIMIC-ABN** [5], a subset of MIMIC-CXR, focuses on abnormal sentences within radiology reports. **3) Our curated Multi-view CXR** aggregates studies with multiple views from both the MIMIC-CXR and IU X-ray [24] datasets. **4) Our curated Two-view CXR** is a variant of Multi-view CXR that includes exactly two views per study, with other settings unchanged. Following established practices [29], [40], [42], we treat the *FINDINGS* section of each radiology report as the reference report, with MIMIC-ABN containing only abnormal sentences from this section. MIMIC-CXR and MIMIC-ABN datasets adhere to their official splits. For Multi-view CXR, we follow the MIMIC-CXR split for its MIMIC-CXR component and baseline split [26], [29], [40] for its IU X-ray component. Detailed statistics for these datasets are summarized in Table I.

Evaluation Metrics. We evaluate model performance using both natural language generation (NLG) and clinical efficacy (CE) metrics. NLG metrics measure the lexical similarities between generated and reference reports, including BLEU- n (B- n , where $n \in \{1, 2, 3, 4\}$), METEOR (MTR), and ROUGE-L (R-L). CE metrics evaluate the clinical correctness of generated reports, including F_1 RadGraph (RG) [49], $F_{1,\text{mic-5}}$ CheXbert (CX5), and $F_{1,\text{mic-14}}$ CheXbert (CX14) [50]. RG metric quantifies the overlap of clinical entities and their relationships, aligning more closely with radiologists’ evaluations than B-3 and CX14 metrics [51]. CX5 and CX14 metrics assess the model’s ability to describe 5 and 14 clinical observations (e.g., *Edema*, *Pneumothorax*, and *Cardiomegaly*) by computing the micro-averaged F_1 score for multi-label classification. These metrics are calculated using the pycocoevalcap [52], radgraph [49], and flchexbert [50] libraries, respectively.

Implementation Details. **1) MIMIC-CXR:** In stage 1, we train the model for 50 epochs using the AdamW optimizer with a learning rate of $5e-5$. In stage 2, we use the RAdam

optimizer for 50 epochs, with learning rates of $5e-6$ for the pre-trained model and $5e-5$ for other parameters. **2) MIMIC-ABN and Multi-view CXR:** As most images in these datasets are derived from MIMIC-CXR [4], Stage 1 is skipped, and the pre-trained model from MIMIC-CXR is fine-tuned using the RAdam optimizer for 50 epochs. Learning rates are set to $5e-6$ for MIMIC-ABN [5] and $5e-7$ for Multi-view CXR. For fairness, all baselines are evaluated using models pre-trained on MIMIC-CXR and fine-tuned before assessment. **3) Two-view CXR:** The same optimizer, learning rate, and epoch settings as MIMIC-CXR are applied. Peer methods are implemented according to their official code, employing concatenation for visual feature fusion. **4) Common settings:** For EVOKE with a resolution of 224^2 , radiograph pre-processing follows R2Gen [40]. For 384^2 resolution, images are resized to 448^2 , randomly cropped to 384^2 , rotated by 5 degrees, and normalized. We exclude studies with empty or clinically insignificant reference reports (e.g., containing only “*Portable AP upright chest film --- at 09:31 is submitted.*”). Stage 1 involves 242M parameters, while Stage 2 has 362M parameters. Experiments at 224^2 resolution are conducted on an NVIDIA 3090 GPU (24GB), and those at 384^2 resolution are performed on an NVIDIA A100 GPU (48GB). The batch size is set to 32. τ_1 , τ_2 , and the number of transition network blocks are set to 0.5, 0.5, and 1, respectively. The maximum number of generated tokens M is fixed at 100 across all datasets. The optimal model is selected based on the combined scores of B-4, RG, and CX14 on the validation set, with results reported on the test set.

B. Results and Analyses

Comparison with State-of-the-art Methods. We compare our proposed EVOKE with 12 state-of-the-art (SOTA) methods: R2Gen [40], CMN [25], CGPT2 [41], MET [42], KiUT [43], SA [35], FMVP [22], SEI [33], MAN [29], PMRG [44], Med-LLM [45], and HERGen [46]. Results are presented in Table II, where “*Input Size*” indicates the image resolution used by the visual encoder (e.g., 224 refers to a 224^2

Observation	MIMIC-CXR						Two-view CXR							
	%	R2Gen			EVOKE (Ours)			%	R2Gen			EVOKE (Ours)		
		P	R	F ₁	P	R	F ₁		P	R	F ₁	P	R	F ₁
ECM	10.2	0.377	0.353	0.365	0.396	0.345	0.369	9.8	0.371	0.280	0.319	0.351	0.300	0.324
Cardiomegaly	14.7	0.567	0.347	0.431	0.628	0.566	0.595	14.1	0.590	0.409	0.483	0.578	0.567	0.573
Lung Opacity	13.3	0.457	0.231	0.307	0.578	0.352	0.437	13.1	0.614	0.069	0.124	0.567	0.335	0.421
Lung Lesion	2.6	0.481	0.046	0.084	0.217	0.036	0.061	2.8	0.500	0.012	0.023	0.357	0.060	0.102
Edema	8.3	0.445	0.254	0.323	0.514	0.447	0.478	5.8	0.523	0.259	0.346	0.475	0.333	0.392
Consolidation	3.3	0.188	0.096	0.127	0.240	0.154	0.188	3.5	0.300	0.115	0.167	0.279	0.163	0.206
Pneumonia	4.6	0.238	0.204	0.220	0.264	0.186	0.218	3.7	0.276	0.071	0.113	0.349	0.268	0.303
Atelectasis	11.1	0.421	0.249	0.313	0.496	0.519	0.507	10.2	0.463	0.248	0.323	0.483	0.464	0.473
Pneumothorax	1.0	0.333	0.029	0.054	0.291	0.157	0.204	0.8	0.000	0.000	0.000	0.222	0.087	0.125
Pleural Effusion	12.5	0.804	0.196	0.316	0.716	0.669	0.691	10.5	0.787	0.375	0.508	0.709	0.667	0.687
Pleural Other	1.7	0.000	0.000	0.000	0.217	0.083	0.120	1.9	0.500	0.018	0.034	0.091	0.036	0.051
Fracture	2.0	0.000	0.000	0.000	0.071	0.014	0.023	2.3	0.000	0.000	0.000	0.188	0.043	0.070
Support Devices	12.5	0.723	0.489	0.583	0.755	0.713	0.734	9.5	0.706	0.488	0.577	0.722	0.646	0.681
No Finding	2.2	0.130	0.626	0.215	0.223	0.660	0.333	12.0	0.387	0.950	0.550	0.547	0.813	0.654
micro avg	-	0.432	0.280	0.340	0.538	0.465	0.499	-	0.484	0.341	0.400	0.539	0.469	0.501
macro avg	-	0.369	0.223	0.238	0.400	0.350	0.354	-	0.430	0.235	0.255	0.423	0.342	0.362

TABLE IV

CLINICAL ACCURACY OF OUR EVOKE ON THE MIMIC-CXR AND TWO-VIEW CXR DATASETS WITH A RESOLUTION OF 224^2 . ECM STANDS FOR *Enlarged Cardiomeastinum*. P, R, AND F₁ DENOTE THE PRECISION, RECALL, AND F₁ SCORE, RESPECTIVELY.

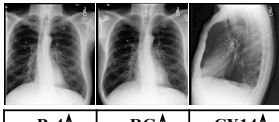
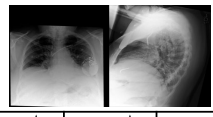
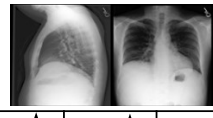
 B-4↑ RG↑ CX14↑ 0.696/0.160 0.857/0.182 0.000/0.000	Reference report: frontal and lateral views of the chest are obtained . the lungs remain hyperinflated , suggesting chronic obstructive pulmonary disease . no focal consolidation , pleural effusion , or evidence of pneumothorax is seen . the cardiac and mediastinal silhouettes are stable and unremarkable . hilar contours are also stable .	Generated report by R2Gen: pa and lateral views of the chest provided . there is no focal consolidation effusion or pneumothorax . the cardiomeastinal silhouette is normal . imaged osseous structures are intact . no free air below the right hemidiaphragm is seen .	Generated report by EVOKE (Ours): frontal and lateral views of the chest were obtained . the lungs are hyperinflated with flattening of the diaphragms , suggesting chronic obstructive pulmonary disease . no focal consolidation , pleural effusion , or evidence of pneumothorax is seen . the cardiac and mediastinal silhouettes are unremarkable .
 B-4↑ RG↑ CX14↑ 0.658/0.000 0.762/0.000 0.667/0.500	Reference report: in comparison with the study of __ , the monitoring and support devices remain in place . diffuse bilateral pulmonary opacifications are essentially unchanged .	Generated report by R2Gen: a single portable ap chest radiograph was obtained . an endotracheal tube terminates in the mid thoracic trachea . a left port-a-cath terminates in the right atrium . a right chest tube terminates in the right apex . a right pleural effusion is small . lung volumes are low . there is no focal consolidation or pneumothorax . cardiac and mediastinal contours are normal .	Generated report by EVOKE (Ours): in comparison with the study of __ , the monitoring and support devices remain in place . diffuse bilateral pulmonary opacifications are again seen , consistent with widespread pneumonia and pulmonary edema .
 B-4↑ RG↑ CX14↑ 0.491/0.125 0.732/0.310 0.857/0.333	Reference report: lung volumes are low compared to the previous study . left-sided aicd device is noted with single lead terminating in unchanged position in the right ventricle . heart size appears at least mildly enlarged . the mediastinal and hilar contours are unremarkable . there is crowding of the bronchovascular structures without overt pulmonary edema . streaky opacities in the lung bases likely reflect areas of atelectasis . no pleural effusion or pneumothorax is present . there are no acute osseous abnormalities .	Generated report by R2Gen: frontal and lateral views of the chest were obtained . dual-lead left-sided pacemaker is again seen with leads extending to the expected positions of the right atrium and right ventricle . there are low lung volumes . there are relatively low lung volumes . there is elevation of the right hemidiaphragm and overlying atelectasis . no definite focal consolidation is seen . there is no pleural effusion or pneumothorax . the cardiac and mediastinal silhouettes are stable .	Generated report by EVOKE (Ours): lung volumes are low . left - sided aicd device is noted with single lead terminating in the right ventricle . heart size is mildly enlarged . mediastinal contours are unremarkable . there is crowding of the bronchovascular structures with likely mild pulmonary vascular congestion . patchy bibasilar airspace opacities likely reflect atelectasis . no pleural effusion or pneumothorax is seen . there are no acute osseous abnormalities .
 B-4↑ RG↑ CX14↑ 0.591/0.245 0.919/0.421 0.000/0.000	Reference report: frontal and lateral views of the chest were obtained . there are relatively low lung volumes . mild elevation of the right hemidiaphragm persists . there is persistent right base atelectasis . no new focal consolidation is seen . there is no pleural effusion or pneumothorax . the cardiac and mediastinal silhouettes are unremarkable .	Generated report by R2Gen: pa and lateral views of the chest provided . lung volumes are low limiting assessment . allowing for this there is no focal consolidation effusion or pneumothorax . the cardiomeastinal silhouette is normal . imaged osseous structures are intact . no free air below the right hemidiaphragm is seen .	Generated report by EVOKE (Ours): frontal and lateral views of the chest were obtained . there are relatively low lung volumes . there is mild elevation of the right hemidiaphragm . no focal consolidation , pleural effusion , or evidence of pneumothorax is seen . the cardiac and mediastinal silhouettes are unremarkable .
 B-4↑ RG↑ CX14↑ 0.100/0.001 0.235/0.074 0.000/0.000	Reference report: the right loculated pleural effusion has decreased in size from __ . there is a little to no remaining effusion on the left and there has been interval resolution of the mild interstitial pulmonary edema . there are no focal consolidations to suggest active infectious process . there is no pneumothorax . the hilar and cardiomeastinal contours are normal . the hemidiaphragms are flattened and the ap diameter is increased .	Generated report by R2Gen: pa and lateral chest views were obtained with patient in upright position . analysis is performed in direct comparison with the next preceding similar study of . the heart size remains unchanged and is within normal limits . unchanged appearance of thoracic aorta . the pulmonary vasculature is not congested . there is now a right-sided port-a-cath system seen to terminate in the mid portion of the svc . no pneumothorax is identified in the apical area . similar as on the preceding examination there is a diffuse interstitial pattern in both lungs with a peripheral distribution in the periphery	Generated report by EVOKE (Ours): frontal and lateral chest radiographs demonstrate a right chest port with the tip terminating in the low svc . there is a small right apical pneumothorax without evidence of tension . post - surgical changes are noted in the right lower lung . there is no focal consolidation or left pleural effusion . the cardiomeastinal silhouette is normal .

Fig. 3. Examples of generated reports on the MIMIC-CXR test set with a resolution of 224^2 . The cell “A/B” represents “EVOKE/R2Gen”. Sentences in the reference report are shown in different colors. Each sentence in generated reports is highlighted in matching colors corresponding to those in the reference report. Failure sentences in EVOKE are underlined.

Metric	w/ MV	w/o MV	w/ Ind	w/o Ind
% (n)	70.7 (2,724)	29.3 (1,128)	57.8 (2,225)	42.2 (1,627)
B-1	0.406	0.369	0.413	0.372
B-2	0.270	0.242	0.280	0.238
B-3	0.196	0.176	0.206	0.169
B-4	0.151	0.136	0.161	0.128
MTR	0.170	0.157	0.174	0.157
R-L	0.316	0.298	0.321	0.298
RG	0.287	0.249	0.299	0.244

TABLE V

BREAKDOWN OF EVOKE’S METRICS ON THE MIMIC-CXR TEST SET USING A RESOLUTION OF 224^2 , CATEGORIZED BY: (I) THE PRESENCE OF MULTI-VIEW RADIOGRAPHS (MV), AND (II) THE INCLUSION OF AN *INDICATION* (IND), IN ALL STUDIES. CX5-F AND CX14-F DENOTE THE FAILURE RATES OF THE CX5 AND CX14 METRICS, RESPECTIVELY.

resolution). EVOKE outperforms all baselines across every metric on four datasets, as evidenced by a 2.9% F_1 RadGraph improvement on MIMIC-CXR, a 7.3% BLEU-1 improvement on MIMIC-ABN, a 3.1% BLEU-4 improvement on Multi-view CXR, and an 8.2% $F_{1, \text{mic-14}}$ CheXbert improvement on Two-view CXR. These results highlight the effectiveness of EVOKE in generating clinically accurate reports. Additionally, at a resolution of 384^2 , EVOKE consistently surpasses its performance at 224^2 resolution on most metrics across these datasets, demonstrating that higher image resolution can significantly enhance the model’s performance.

Effect of Multi-view Contrastive Learning. We evaluate the impact of our multi-view contrastive learning method on model performance, as shown in Table III. Results show positive effects from multi-positive contrastive loss (EVOKE vs. (g)), instance-wise alignment loss (EVOKE vs. (f)), and token-wise alignment loss (EVOKE vs. (e)). EVOKE achieves superior performance across all metrics compared to (a) directly applying Stage 2 for report generation. This demonstrates the effectiveness of our multi-view contrastive learning method in enhancing visual features, thereby laying a solid foundation for Stage 2 to generate accurate and coherent reports based on patient-specific *INDICATION*.

Effect of Knowledge-guided Report Generation. We assess the impact of our knowledge-guided report generation module on model performance, as detailed in Table III. Both single-view ((c) vs. (b)) and multi-view methods (EVOKE vs. (h)) exhibit significant improvements across all metrics, emphasizing the importance of patient-specific *INDICATION* in generating accurate reports. Additionally, EVOKE shows a 3.1% increase in the sum of all metrics compared to (j), confirming the effectiveness of employing the transition “bridge” for handling missing *INDICATION*.

Clinical Accuracy of 14 Observations. Table IV illustrates the clinical accuracy of 14 observations related to thoracic diseases and support devices on the MIMIC-CXR and Two-view CXR datasets. Results reveal that: 1) EVOKE significantly improves most F_1 compared to R2Gen [40]. 2) Although EVOKE is not specifically designed for imbalanced observations, it slightly outperforms the baseline, improving F_1 for *Pleural Other* and *Fracture* by 12% and 2.3% in MIMIC-CXR, and by 1.7% and 7% in Two-view CXR, respectively.

Model Benefits from Multi-view Radiographs and *INDICATION*. Table V shows performance variations on subsets of the MIMIC-CXR test set with and without multi-view radiographs (*INDICATION*). Significant improvements in NLG and RG metrics are observed for studies with multi-view radiographs (*INDICATION*). This suggests that multi-view images offer rich visual information, while patient-specific knowledge provides essential contextual cues, collectively enhancing the accuracy of the generated radiology reports.

Case Study. Fig. 3 presents five examples from the MIMIC-CXR test set to show the quality of our generated reports. A color distribution closely matching the reference reports indicates comprehensive coverage of clinical findings, while longer color bars in generated reports suggest more detailed descriptions. Results reveal that EVOKE generates higher-quality draft reports for radiologists compared to R2Gen [40]. For instance, in case 1, radiologists only needed to add “*hilar contours are also stable*”. However, both EVOKE and R2Gen struggle to accurately capture disease progression due to the lack of consideration for patient temporal data, as seen in case 2, where the term “*unchanged*” is not inferred.

Evaluation via Large Language Models. We evaluate the quality of our generated reports using the GREEN model [53], which employs a fine-tuned LLaMA-2 [54] model to detect clinically significant errors in radiology reports. The GREEN model identifies six types of clinically significant errors: (a) False report of a finding in the candidate; (b) Missing a finding present in the reference; (c) Misidentification of a finding’s anatomic location/position; (d) Misassessment of the severity of a finding; (e) Mentioning a comparison that isn’t in the reference; (f) Omitting a comparison detailing a change from a prior study. Based on the number of matched findings between generated and reference reports (denoted as “#Matched Findings”), the GREEN score is defined as:

$$\text{GREEN} = \frac{\# \text{Matched Findings}}{\# \text{Matched Findings} + \sum_{i=(a)}^{(f)} \# \text{Error}_i}, \quad (10)$$

where $\# \text{Error}_i$ refers to the count of the i^{th} clinically significant error. Fig. 4 compares our EVOKE with R2Gen [40], CMN [25], and CGPT2 [41] on the MIMIC-CXR test set, evaluated using “#Matched Findings” and GREEN score. The result shows that EVOKE outperforms all baselines, indicating its superior capability to generate clinically accurate reports.

Human Evaluations. To ensure objectivity and reliability, we conduct a double-blind human evaluation of radiology report generation. We randomly select 50 cases from the MIMIC-CXR test set, each consisting of a reference report, multi-view radiographs, and three randomly ordered candidate reports generated by R2Gen [40], CMN [25], and our EVOKE-224. Two experienced radiologists independently assess the candidate reports based on two key criteria: 1) Error type assessment. Each candidate report is evaluated for the severity of four potential errors: (I) False prediction of finding, (II) Omission of finding, (III) Incorrect location/position of finding, and (IV) Incorrect severity of finding. The severity is categorized into significant error (SE), insignificant error (IE), or no error (NE). 2) Candidate report overall ranking.

Model	(I)			(II)			(III)			(IV)			Overall Rank		
	SE ↓	IE ↓	NE ↑	SE ↓	IE ↓	NE ↑	SE ↓	IE ↓	NE ↑	SE ↓	IE ↓	NE ↑	1	2	3
R2Gen	24%	18%	58%	54%	38%	8%	24%	2%	74%	32%	32%	36%	22%	46%	32%
CMN	26%	24%	50%	34%	46%	20%	22%	12%	66%	42%	28%	30%	26%	48%	26%
EVOKE (Ours)	16%	14%	70%	24%	44%	32%	10%	14%	76%	12%	34%	54%	68%	18%	14%

TABLE VI

HUMAN EVALUATION RESULTS, INCLUDING SEVERITY ASSESSMENT OF FOUR ERROR TYPES (SE: SIGNIFICANT ERROR, IE: INSIGNIFICANT ERROR, NE: NO ERROR) AND OVERALL RANKING OF CANDIDATE REPORTS (1: BEST, 3: WORST).

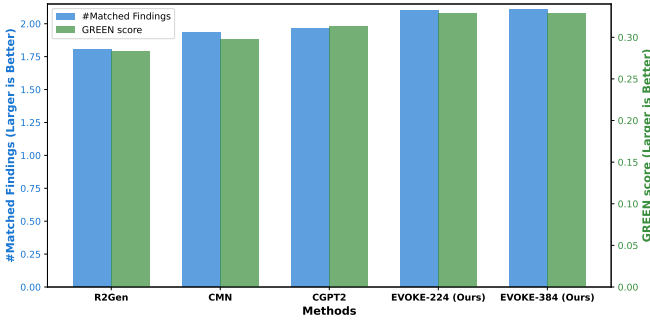


Fig. 4. Comparisons with baselines on the MIMIC-CXR test set, evaluated by “#Matched Findings” and GREEN score. “EVOKE-s” represents a resolution-specific variant of our proposed EVOKE framework.

The radiologists rank the three candidate reports in each case based on overall quality, assigning 1 (best), 2 (second-best), and 3 (worst). Table VI presents the results, showing that our EVOKE consistently outperforms both R2Gen [40] and CMN [25] across all error categories and overall ranking. Specifically, EVOKE achieves the lowest rates of significant errors (SE) and the highest rates of no errors (NE) for all error types (I-IV). Additionally, 68% of EVOKE-generated reports are ranked as the best (1), significantly surpassing R2Gen (22%) and CMN (26%). These results indicate that EVOKE generates more accurate and clinically reliable radiology reports than baseline models.

V. CONCLUSION

In this paper, we proposed EVOKE for chest X-ray report generation by incorporating multi-view images and patient-specific *INDICATION*. We introduced a multi-view contrastive learning method to align multi-view radiographs with their corresponding report, addressing the challenge of handling varying numbers of images per study. Additionally, we presented a knowledge-guided report generation module that integrates available patient-specific *INDICATION*, providing the model with contextual insights. Experiments on MIMIC-CXR, MIMIC-ABN, Multi-view CXR, and Two-view CXR datasets show that incorporating multi-view radiographs, *INDICATION*, and high-resolution radiographs improves the model’s accuracy in describing clinical findings. In the future, we plan to utilize patient temporal data [46], [55] to capture disease progression and enhance model reliability through prediction uncertainty [56].

REFERENCES

- [1] P. Messina *et al.*, “A survey on deep learning and explainability for automatic report generation from medical images,” *ACM Computing Surveys*, vol. 54, no. 10s, pp. 1–40, 2022.
- [2] X. Mei *et al.*, “Phraseaug: An augmented medical report generation model with phrasebook,” *IEEE Transactions on Medical Imaging*, vol. 43, no. 12, pp. 4211–4223, 2024.
- [3] S. Bannur *et al.*, “Maira-2: Grounded radiology report generation,” 2024. [Online]. Available: <https://arxiv.org/abs/2406.04449>
- [4] A. E. W. Johnson *et al.*, “Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs,” 2019. [Online]. Available: <https://arxiv.org/abs/1901.07042>
- [5] J. Ni *et al.*, “Learning visual-semantic embeddings for reporting abnormal findings on chest x-rays,” in *EMNLP*, 2020, pp. 1954–1960.
- [6] Z. Wang *et al.*, “Automated radiographic report generation purely on transformer: A multicriteria supervised approach,” *IEEE Transactions on Medical Imaging*, vol. 41, no. 10, pp. 2803–2813, 2022.
- [7] A. Liu *et al.*, “Multi-grained radiology report generation with sentence-level image-language contrastive learning,” *IEEE Transactions on Medical Imaging*, pp. 1–1, 2024.
- [8] M. Li *et al.*, “Dynamic graph enhanced contrastive learning for chest x-ray report generation,” in *CVPR*, 2023, pp. 3334–3343.
- [9] C. Liu *et al.*, “Bootstrapping large language models for radiology report generation,” in *AAAI*, vol. 38, no. 17, 2024, pp. 18 635–18 643.
- [10] X. Liang *et al.*, “Divide and conquer: Isolating normal-abnormal attributes in knowledge graph-enhanced radiology report generation,” in *ACMMM*, 2024, pp. 4967–4975.
- [11] W. Hou *et al.*, “RECAP: towards precise radiology report generation via dynamic disease progression reasoning,” in *EMNLP*, 2023, pp. 2134–2147.
- [12] J. Li *et al.*, “BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation,” in *ICML*, vol. 162, 17–23 Jul 2022, pp. 12 888–12 900.
- [13] T. Chen, R. ZHANG, and G. Hinton, “Analog bits: Generating discrete data using diffusion models with self-conditioning,” in *ICLR*, 2023.
- [14] Y. Yang *et al.*, “Token-mixer: Bind image and text in one embedding space for medical image reporting,” *IEEE Transactions on Medical Imaging*, vol. 43, no. 11, pp. 4017–4028, 2024.
- [15] S. Bu *et al.*, “Instance-level expert knowledge and aggregate discriminative attention for radiology report generation,” in *CVPR*, June 2024, pp. 14 194–14 204.
- [16] K. Liu *et al.*, “Enhanced contrastive learning with multi-view longitudinal data for chest x-ray report generation,” 2025. [Online]. Available: <https://arxiv.org/abs/2502.20056>
- [17] S. Li *et al.*, “An organ-aware diagnosis framework for radiology report generation,” *IEEE Transactions on Medical Imaging*, vol. 43, no. 12, pp. 4253–4265, 2024.
- [18] C. Xu, D. Tao, and C. Xu, “A survey on multi-view learning,” 2013. [Online]. Available: <https://arxiv.org/abs/1304.5634>
- [19] J. Zhao *et al.*, “Multi-view learning overview: Recent progress and new challenges,” *Information Fusion*, vol. 38, pp. 43–54, 2017.
- [20] A. Paul *et al.*, “Generalized zero-shot chest x-ray diagnosis through trait-guided multi-view semantic embedding with self-training,” *IEEE Transactions on Medical Imaging*, vol. 40, no. 10, pp. 2642–2655, 2021.
- [21] H.-Y. Zhou *et al.*, “Generalized radiograph representation learning via cross-supervision between images and free-text radiology reports,” *Nature Machine Intelligence*, vol. 4, no. 1, p. 32–40, Jan 2022.
- [22] Z. Liu *et al.*, “From observation to concept: A flexible multi-view paradigm for medical report generation,” *IEEE Transactions on Multimedia*, vol. 26, pp. 5987–5995, 2024.
- [23] A. Vaswani *et al.*, “Attention is all you need,” in *NeurIPS*, vol. 30, 2017.

- [24] D. Demner-Fushman *et al.*, “Preparing a collection of radiology examinations for distribution and retrieval,” *Journal of the American Medical Informatics Association*, vol. 23, no. 2, pp. 304–310, 2016.
- [25] Z. Chen *et al.*, “Cross-modal memory networks for radiology report generation,” in *ACL*, vol. 1, 2021, pp. 5904–5914.
- [26] S. Yang *et al.*, “Radiology report generation with a learned knowledge base and multi-modal alignment,” *Medical Image Analysis*, vol. 86, p. 102798, 2023.
- [27] F. Wang *et al.*, “Multi-granularity cross-modal alignment for generalized medical visual representation learning,” in *NeurIPS*, vol. 35, 2022, pp. 33 536–33 549.
- [28] P. Cheng *et al.*, “Prior: Prototype representation joint learning from medical images and reports,” in *ICCV*, 2023, pp. 21 361–21 371.
- [29] H. Shen *et al.*, “Automatic radiology reports generation via memory alignment network,” in *AAAI*, vol. 38, no. 5, 2024, pp. 4776–4783.
- [30] A. Deria *et al.*, “Inverge: Intelligent visual encoder for bridging modalities in report generation,” in *CVPRW*, June 2024, pp. 2028–2038.
- [31] Z. Wang *et al.*, “Medclip: Contrastive learning from unpaired medical images and text,” in *EMNLP*, 2022, pp. 3876–3887.
- [32] Z. Chen, G. Li, and X. Wan, “Align, reason and learn: Enhancing medical vision-and-language pre-training with knowledge,” in *ACMMM*, 2022, pp. 5152–5161.
- [33] K. Liu *et al.*, “Structural entities extraction and patient indications incorporation for chest x-ray report generation,” in *MICCAI*. Cham: Springer Nature Switzerland, 2024, pp. 433–443.
- [34] K. He *et al.*, “Deep residual learning for image recognition,” in *CVPR*, 2016, pp. 770–778.
- [35] B. Yan *et al.*, “Style-aware radiology report generation with radgraph and few-shot prompting,” in *EMNLP*, 2023, pp. 14 676–14 688.
- [36] K. Liu *et al.*, “Factual serialization enhancement: A key innovation for chest x-ray report generation,” 2024. [Online]. Available: <https://arxiv.org/abs/2405.09586>
- [37] I. Beltagy, K. Lo, and A. Cohan, “Scibert: A pretrained language model for scientific text,” in *EMNLP*, 2019, pp. 3615–3620.
- [38] Y. Tian *et al.*, “Stablerep: Synthetic images from text-to-image models make strong visual representation learners,” in *NeurIPS*, vol. 36, 2023, pp. 48 382–48 402.
- [39] A. van den Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” 2019. [Online]. Available: <https://arxiv.org/abs/1807.03748>
- [40] Z. Chen *et al.*, “Generating radiology reports via memory-driven transformer,” in *EMNLP*, 2020, pp. 1439–1449.
- [41] A. Nicolson, J. Dowling, and B. Koopman, “Improving chest x-ray report generation by leveraging warm starting,” *Artificial Intelligence in Medicine*, vol. 144, p. 102633, 2023.
- [42] Z. Wang *et al.*, “Mettransformer: Radiology report generation by transformer with multiple learnable expert tokens,” in *CVPR*, 2023, pp. 11 558–11 567.
- [43] Z. Huang, X. Zhang, and S. Zhang, “Kiut: Knowledge-injected u-transformer for radiology report generation,” in *CVPR*, 2023, pp. 19 809–19 818.
- [44] H. Jin *et al.*, “Promptmrg: Diagnosis-driven prompts for medical report generation,” in *AAAI*, vol. 38, no. 3, Mar 2024, p. 2607–2615.
- [45] R. Liu *et al.*, “In-context learning for zero-shot medical report generation,” in *ACMMM*, 2024, pp. 8721–8730.
- [46] F. Wang, S. Du, and L. Yu, “Hergen: Elevating radiology report generation with longitudinal data,” in *ECCV*. Cham: Springer Nature Switzerland, 2025, pp. 183–200.
- [47] J. Tian *et al.*, “Towards automatic diagnosis from multi-modal medical data,” in *MICCAI/W*, vol. 11797, 2019, pp. 67–74.
- [48] D. Nguyen *et al.*, “Pragmatic radiology report generation,” in *MLAH*, vol. 225, 2023, pp. 385–402.
- [49] S. Jain *et al.*, “Radgraph: Extracting clinical entities and relations from radiology reports,” in *NeurIPS*, vol. 1, 2021.
- [50] A. Smit *et al.*, “Combining automatic labelers and expert annotations for accurate radiology report labeling using bert,” in *EMNLP*, 2020.
- [51] F. Yu *et al.*, “Evaluating progress in automatic chest x-ray radiology report generation,” *Patterns*, vol. 4, no. 9, p. 100802, 2023.
- [52] X. Chen *et al.*, “Microsoft coco captions: Data collection and evaluation server,” 2015. [Online]. Available: <https://arxiv.org/abs/1504.00325>
- [53] S. Ostmeier *et al.*, “GREEN: Generative radiology report evaluation and error notation,” in *EMNLP*. Miami, Florida, USA: Association for Computational Linguistics, nov 2024, pp. 374–390. [Online]. Available: <https://aclanthology.org/2024.findings-emnlp.21>
- [54] H. Touvron *et al.*, “Llama 2: Open foundation and fine-tuned chat models,” 2023. [Online]. Available: <https://arxiv.org/abs/2307.09288>
- [55] W. Hou *et al.*, “ORGAN: observation-guided radiology report generation via tree reasoning,” in *ACL*, 2023, pp. 8108–8122.
- [56] X. Kang *et al.*, “Multi-scale information sharing and selection network with boundary attention for polyp segmentation,” 2024. [Online]. Available: <https://arxiv.org/abs/2405.11151>