PriorRG: Prior-Guided Contrastive Pre-training and Coarse-to-Fine Decoding for Chest X-ray Report Generation

Kang Liu[1,2,3], **Zhuoqi Ma**[1,2,3], **Zikang Fang**[1], **Yunan Li**[1,2,3], **Kun Xie**[1,2,3], **Qiguang Miao**[1,2,3*]

[1]School of Computer Science and Technology, Xidian University, Xi'an, Shaanxi 710071, China
[2]Xi'an Key Laboratory of Big Data and Intelligent Vision, Xi'an, Shaanxi 710071, China
[3]Key Laboratory of Collaborative Intelligence Systems, Ministry of Education, Xidian University, Xi'an 710071, China
{kangliu, 22009200766}@stu.xidian.edu.cn, {zhuoqima, yunanli, xiekun, qgmiao}@xidian.edu.cn

Abstract

Chest X-ray report generation aims to reduce radiologists' workload by automatically producing high-quality preliminary reports. A critical yet underexplored aspect of this task is the effective use of patient-specific prior knowledge—including clinical context (e.g., symptoms, medical history) and the most recent prior image—which radiologists routinely rely on for diagnostic reasoning. Most existing methods generate reports from single images, neglecting this essential prior information and thus failing to capture diagnostic intent or disease progression. To bridge this gap, we propose **PriorRG**, a novel chest X-ray report generation framework that emulates real-world clinical workflows via a two-stage training pipeline. In Stage 1, we introduce a prior-guided contrastive pre-training scheme that leverages clinical context to guide spatiotemporal feature extraction, allowing the model to align more closely with the intrinsic spatiotemporal semantics in radiology reports. In Stage 2, we present a prior-aware coarse-to-fine decoding for report generation that progressively integrates patient-specific prior knowledge with the vision encoder's hidden states. This decoding allows the model to align with diagnostic focus and track disease progression, thereby enhancing the clinical accuracy and fluency of the generated reports. Extensive experiments on MIMIC-CXR and MIMIC-ABN datasets demonstrate that PriorRG outperforms state-of-the-art methods, achieving a 3.6% BLEU-4 and 3.8% F1 score improvement on MIMIC-CXR, and a 5.9% BLEU-1 gain on MIMIC-ABN.

Code — https://github.com/mk-runner/PriorRG
Extended version — https://arxiv.org/abs/2508.05353

1 Introduction

Radiology report generation (RRG) (Wang et al. 2025; Mei et al. 2024) leverages AI techniques to automatically interpret medical images—such as chest X-rays (Miao et al. 2025; Zhang et al. 2025), CT scans (Hamamci, Er, and Menze 2024; Li et al. 2025a; Izhar et al. 2025), and pathological slides (Guo et al. 2024; Chen et al. 2024)—and generate structured textual descriptions of clinically relevant findings. This automation supports radiologists by providing standardized preliminary reports, thereby improving diagnostic efficiency and consistency.

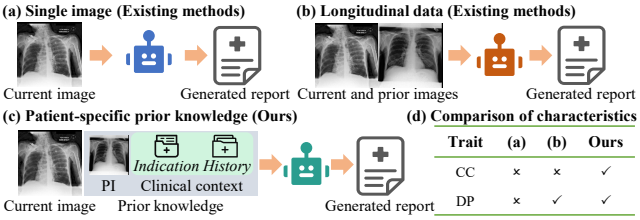


Figure 1: (a-c) show the workflow of existing methods and our approach. (d) summarizes their key properties. Longitudinal data includes both current and prior images. Patient-specific prior knowledge comprises the prior image (PI), *indication*, and medical *history*, which may be partially missing. "CC" and "DP" indicate whether a method models clinical context and disease progression, respectively.

While recent RRG approaches have achieved remarkable progress, most models operate on isolated images (see Figure 1(a)) and overlook essential patient-specific prior knowledge, including both clinical context (i.e., the *indication* and *history* sections) and the most recent prior image. Such prior knowledge plays a critical role in real-world clinical reasoning by enabling personalized interpretation and tracking of disease progression. However, existing methods (Li et al. 2023; Liu et al. 2024a; Xiao et al. 2025) largely ignore these factors, limiting their ability to produce context-aware, disease progression-oriented reports.

To capture the clinical context, (Nguyen et al. 2023) treats *indications* as auxiliary input while ignoring the noise in the data. SEI (Liu et al. 2024b) eliminates invalid or corrupted characters via preprocessing and employs a cross-modal fusion network to incorporate *indications*, thereby generating accurate findings. Similarly, for (Mei et al. 2025). However, they fail to consider longitudinal information, often resulting in hallucinations when describing disease progression. To combat this issue, existing methods incorporate prior images (Figure 1(b)) and employ techniques such as report prefilling (Zhu et al. 2023), intra-modality similarity constraints (Liu et al. 2025b), and group causal transformers (Wang, Du, and Yu 2025) to model temporal visual changes. Yet, they overlook clinical context, limiting their capacity to generate

*Corresponding author

personalized and context-aware reports. This gap leads to our central research question: **Can we jointly model temporal visual changes and clinical context for improved cross-modal alignment and report generation?**

To address this challenge, we propose PriorRG, a novel chest X-ray report generation framework that mirrors real-world radiology workflows via a two-stage training pipeline. Stage 1 introduces a prior-guided contrastive pre-training scheme that simulates diagnostic reasoning by leveraging clinical context to guide spatiotemporal feature extraction, achieving better alignment with context-aware, disease progression-oriented radiology reports. Stage 2 presents a prior-aware coarse-to-fine decoding for report generation. We first devise an attention-enhanced layer fusion network to derive hierarchical visual representations from the vision encoder's hidden states. Motivated by principles of visual cognition, PriorRG progressively integrates clinical context, spatiotemporal information, and hierarchical visual cues in a coarse-to-fine manner. This design equips the report generator with rich, context-aware, disease progression-oriented representations, thereby enhancing both the clinical efficacy and linguistic quality of the generated reports. Comprehensive experiments on the MIMIC-CXR (Johnson et al. 2019) and MIMIC-ABN (Ni et al. 2020) datasets demonstrate that our PriorRG significantly outperforms recent state-of-the-art methods in both medical image-text retrieval and radiology report generation. Our main contributions are:

- We propose PriorRG, which integrates patient-specific prior knowledge to generate context-aware and disease progression-oriented reports.
- We introduce a prior-guided contrastive pre-training scheme that mirrors diagnostic reasoning by using clinical context to guide spatiotemporal features extraction, thereby improving alignment with report semantics and boosting medical image-text retrieval performance.
- We present a prior-aware coarse-to-fine decoding that incrementally integrates clinical context, disease progression patterns, and hierarchical visual cues, enhancing the clinical accuracy and fluency of generated reports.

2 Related Work

Radiology report generation (RRG). Unlike generic image captioning (Zeng et al. 2023; Hu and Li 2024), RRG aims to generate detailed clinical descriptions for medical images (Liu et al. 2025a). Recent advances have explored various techniques to improve accuracy, including knowledge graphs integration (Yin et al. 2025), contrastive learning (Liu et al. 2024b; Li et al. 2025b), retrieval-augmented methods (Jeong et al. 2024; Liu et al. 2024c), memory alignment (Chen et al. 2021; Shen et al. 2024), human preference optimization (Zhou et al. 2024b; Xiao et al. 2025), and LLM-based methods (Wang et al. 2023b; Liu et al. 2024a). However, most existing methods rely solely on single-view images and overlook patient-specific prior knowledge—an essential component in real-world clinical decision-making. This limitation hinders their ability to capture clinical intent and monitor disease progression. To address this gap, we propose PriorRG that integrates prior knowledge via coarse-to-fine decoding to generate context-aware, disease progression-oriented reports.

RRG via prior knowledge. Several studies (Nguyen et al. 2023; Liu et al. 2024b; Mei et al. 2025) leverage *indications* to generate personalized reports, yet overlook disease progression—often leading to hallucinated descriptions of lesion changes. Recent efforts introduce longitudinal modeling via intra-modality similarity constraints (Liu et al. 2025b), report pre-filling (Zhu et al. 2023), or the group causal transformer (Wang, Du, and Yu 2025) to capture temporal changes. While promising, they still underexploit individual clinical context, limiting their ability to infer patients' diagnostic intent. To bridge these gaps, we propose a prior-aware coarse-to-fine decoding that progressively integrates clinical context, disease progression patterns, and hierarchical visual cues, resulting in more accurate, fluent, and context-aware radiology reports.

Medical vision-language pre-training (MVLP). MVLP aims to learn joint visual-textual representations to support downstream tasks such as image-text retrieval (Wang et al. 2022b; Zhang et al. 2023a; Bannur et al. 2023) and report generation (Jin et al. 2024; Liu et al. 2024b, 2025a). GLoRIA (Huang et al. 2021) and MGCA (Wang et al. 2022a) enhance representations using multi-level cross-modal alignment, while SEI (Liu et al. 2024b) links images to structured clinical entities. MedCLIP (Wang et al. 2022b) scales training via decoupled image-text matching based on the semantic similarity. To integrate domain knowledge, ARL (Chen, Li, and Wan 2022) and KAD (Zhang et al. 2023b) utilize UMLS for semantic alignment and zero-shot disease classification. Beyond static image-text pairs, BioViL-T (Bannur et al. 2023) and MLRG (Liu et al. 2025a) leverage longitudinal data to model disease progression. However, clinical context—crucial for diagnostic reasoning—remains largely underexplored. To address this, we introduce a prior-guided contrastive pre-training scheme that explicitly encodes clinical context and tracks disease progression, thereby improving cross-modal alignment.

3 Method

3.1 Problem Statement

Figure 2 illustrates the architecture of our PriorRG. The input includes a current image x_i^{cur}, a most recent prior image x_i^{pri} (which may be absent), an indication z_i (potentially missing), and a medical *history* h_i (possibly unavailable). We refer to z_i and h_i collectively as the clinical context for the i^{th} sample. A key challenge lies in effectively leveraging this potentially incomplete prior information—x_i^{pri}, z_i, and h_i—to provide the model with personalized clinical background and disease progression cues. Our goal is to generate context-aware, disease progression-oriented radiology reports $\hat{y}_i$ by integrating all available inputs.

3.2 Stage 1: Prior-Guided Contrastive Pre-training

Visual features extraction. Following (Liu et al. 2025a), we employ the pre-trained vision encoder RAD-DINO (Pérez-García et al. 2024) to extract visual features from chest X-

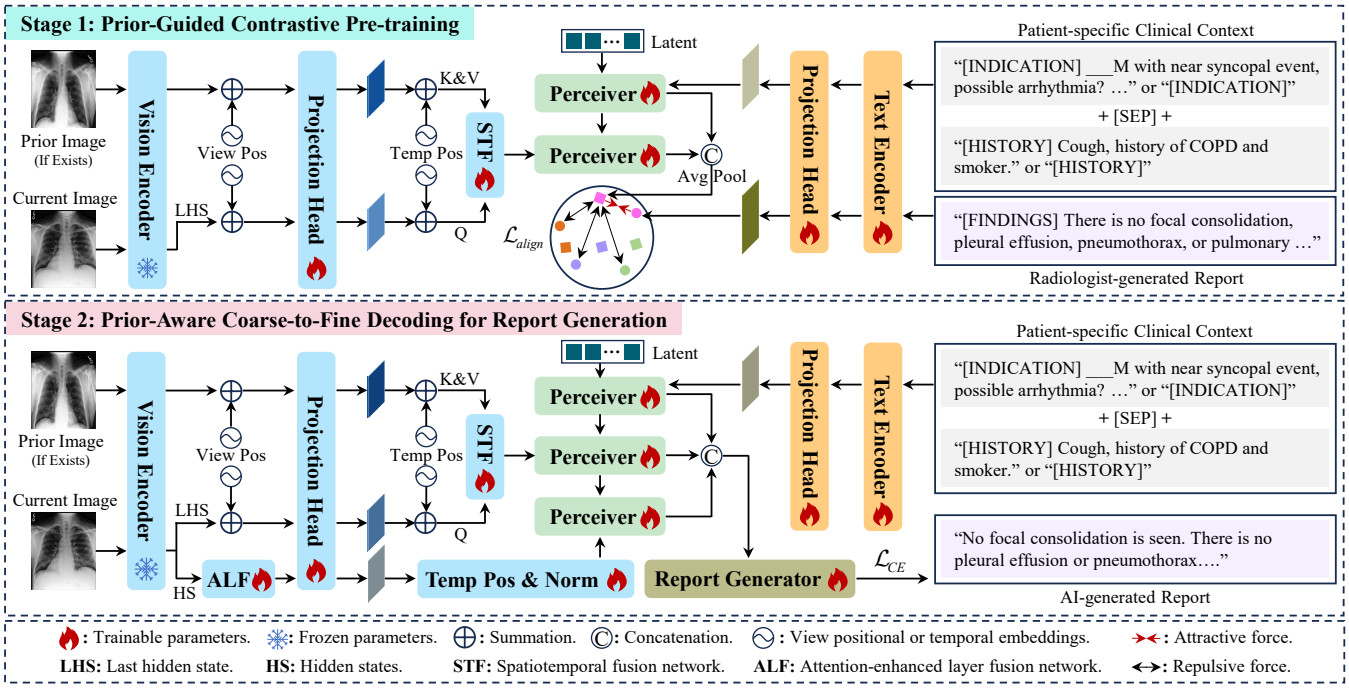


Figure 2: Overview of our PriorRG with a two-stage training pipeline, which consists of a vision encoder (RAD-DINO (Pérez-García et al. 2024)), a text encoder (CXR-BERT (Boecking et al. 2022)), and a report generator (DistilGPT2 (Sanh et al. 2020)).

rays. Radiographic view positions (e.g., AP vs. PA) significantly affect image appearance—for instance, the cardiac silhouette appears enlarged in AP views. To account for such projection-specific variations, we introduce a learnable view-position embedding and fuse it with the extracted visual features, enhancing the model's robustness to view-dependent discrepancies. The augmented features are then projected into a unified d-dimensional embedding space via a projection head. We denote the resulting visual features as $\boldsymbol{V} \in \mathbb{R}^{M\times s\times d}$, where s is the sequence length, and B represents batch size.

Textual features extraction. We adopt the pre-trained text encoder CXR-BERT (Boecking et al. 2022), followed by a projection head. Specifically, the last hidden state from CXR-BERT is passed through the projection head to obtain textual features, denoted as $\boldsymbol{T} \in \mathbb{R}^{B\times p\times d}$, where p is the number of tokens. To ensure consistency across different types of input, we prepend special tokens—"[INDICATION]", "[HISTORY]", and "[FINDINGS]"—to the *indication*, medical *history*, and radiology reports, respectively (see Figure 2). This design facilitates type-aware feature extraction while gracefully handling missing fields (e.g., absent *indication* or medical *history*), enabling a unified and robust encoding process.

Spatiotemporal fusion network (STF). While the current image alone can support report generation, relying solely on it may lead to hallucinations—particularly producing disease progression description (e.g., "*As compared to the previous radiograph, the patient has received a new right internal jugular vein catheter.*"). To mitigate this, we employ a ViT-style spatiotemporal fusion network (Liu et al. 2025a; Dosovitskiy et al. 2021) to model disease progression between current and prior images. The fusion process proceeds as follows: First, we inject temporal embeddings into both current and prior visual features to encode chronological relationships. Each STF block is then formulated as:

$$\boldsymbol{V}^{st}_{ca} = \mathrm{LN}(\boldsymbol{V}^{cur} + \mathrm{CA}(\mathrm{LN}(\boldsymbol{V}^{cur}), \mathrm{LN}(\boldsymbol{V}^{pri}))), \quad (1)$$

$$\boldsymbol{V}^{st} = \mathrm{LN}(\boldsymbol{V}^{st}_{ca} + \mathrm{FFN}(\boldsymbol{V}^{st}_{ca})), \quad (2)$$

where $\boldsymbol{V}^{cur}$ and $\boldsymbol{V}^{pri}$ denote visual features of current and prior images, respectively. $\mathrm{LN}(\cdot)$ and $\mathrm{FFN}(\cdot)$ represent layer normalization and feed-forward network. $\mathrm{CA}(\mathrm{Q}, \mathrm{K\&V})$ denotes the cross-attention module used to capture inter-temporal dependencies. The number of STF blocks is empirically set to 3. For samples without a prior image, we directly treat $\boldsymbol{V}^{cur}$ as the spatiotemporal features $\boldsymbol{V}^{st} \in \mathbb{R}^{B\times s\times d}$.

Instance-wise cross-modal alignment. In clinical practice, physicians typically begin with an initial clinical assessment based on a patient's symptoms (i.e., *indications*) and medical *history*. Radiologists then integrate this clinical context with prior and current images to comprehensively evaluate lesion characteristics and progression. To simulate this diagnostic workflow, we incrementally incorporate clinical context $\boldsymbol{T}^c$ and spatiotemporal visual features $\bar{\boldsymbol{V}}^{st}$ using the Perceiver architecture (Jaegle et al. 2021), formulated as:

$$\bar{\boldsymbol{T}}^c = \mathrm{Perceiver}(\boldsymbol{E}^{lat}, \boldsymbol{T}^c), \quad (3)$$

$$\bar{\boldsymbol{V}}^{st} = \mathrm{Perceiver}(\bar{\boldsymbol{T}}^c, \boldsymbol{V}^{st}), \quad (4)$$

where $\mathrm{Perceiver}(\boldsymbol{P}, \boldsymbol{Q})$ is a modality-agnostic architecture that compresses input $\boldsymbol{Q}$ into a compact, learnable latent embedding $\boldsymbol{P}$ using the cross-attention module. Here, $\boldsymbol{E}^{lat} \in$

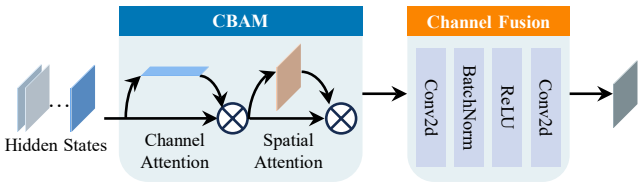


Figure 3: Overview of attention-enhanced layer fusion network (ALF), which leverages CBAM (Woo et al. 2018) to extract hierarchical visual features.

$\mathbb{R}^{B\times N\times d}$ refers to the learnable latent embedding, where N indicates the number of latents. $\bar{\boldsymbol{T}}^c \in \mathbb{R}^{B\times N\times d}$ denotes condensed clinical context feature, and $\bar{\boldsymbol{V}}^{st} \in \mathbb{R}^{B\times N\times d}$ signifies clinically informed spatiotemporal feature.

To ensure the consistency between image-report pairs, we employ the instance-wise cross-modal alignment (Wang et al. 2022a; Liu et al. 2025a) to enhance multimodal representations. More concretely, we concatenate $\bar{\boldsymbol{T}}^c$ and $\bar{\boldsymbol{V}}^{st}$ along the sequence dimension, apply global average pooling, and perform L2 normalization to obtain global visual features $\boldsymbol{V}^g \in \mathbb{R}^{B\times d}$. The image-to-report similarity logits $\boldsymbol{p}^{i2r} \in \mathbb{R}^{B\times B}$ are then computed as:

$$\boldsymbol{p}_i^{i2r} = \frac{\exp\left(\boldsymbol{V}_i^g \cdot (\boldsymbol{T}_i^g)^T/\tau\right)}{\sum_{j=1}^{B}\exp\left(\boldsymbol{V}_i^g \cdot (\boldsymbol{T}_j^g)^T/\tau\right)}, \tag{5}$$

where $\boldsymbol{T}_i^g \in \mathbb{R}^{B\times d}$ denotes the global textual features for radiologist-generated reports. τ is the temperature parameter. Similarly, the report-to-image similarity logits are described as $\boldsymbol{p}^{r2i} \in \mathbb{R}^{B\times B}$. To account for the possibility of multi-view images during a visit, we treat all image-report pairs associated with the same visit as positive pairs, resulting in multiple positive pairs. Consequently, the ground-truth matching label matrix $\boldsymbol{q} \in \mathbb{R}^{B\times B}$ is defined as:

$$\boldsymbol{q}_{i,j} = \frac{\mathbb{I}_{\text{equal}}\left(y_i, y_j\right)}{\sum_{k=1}^{B}\mathbb{I}_{\text{equal}}\left(y_i, y_k\right)}, \tag{6}$$

where $\mathbb{I}_{\text{equal}}(y_i, y_j)$ is an indicator function that equals 1 if the i^{th} and j^{th} samples share the same report (i.e., $y_i = y_j$), and 0 otherwise. The instance-wise cross-modal alignment loss is defined as the cross-entropy loss between ground-truth matching label matrix $\boldsymbol{q}$ and similarity logits $\boldsymbol{p}$:

$$\mathcal{L}_{align} = -\frac{1}{2B}\sum_{k=1}^{B}\left(\boldsymbol{q}_k \log \boldsymbol{p}_k^{i2r} + \boldsymbol{q}_k \log \boldsymbol{p}_k^{r2i}\right). \tag{7}$$

To summarize, Stage 1 optimizes the alignment loss $\mathcal{L}_{align}$, which leverages clinical context to guide the extraction of spatiotemporal features, enhancing cross-modal alignment and boosting medical image-text retrieval performance.

3.3 Stage 2: Prior-Aware Coarse-to-Fine Decoding for Report Generation

Stage 2 consists of two main components: (1) an attention-enhanced hierarchical fusion network for constructing hierarchical visual semantics, and (2) a coarse-to-fine decoding that progressively integrates prior knowledge with hierarchical visual semantics to guide report generation.

Attention-enhanced layer fusion network. Previous methods (Liang et al. 2024; Liu et al. 2025a) rely solely on the vision encoder's last hidden state, overlooking low-level details such as lesion morphology. To mitigate this, we design an attention-enhanced layer fusion network based on CBAM (Woo et al. 2018) (see Figure 3). CBAM's channel and spatial attention are applied to each encoder layer to highlight diagnostically relevant features. The refined features are fused through a Conv2D projector that preserves both spatial and semantic information. A projection head with temporal positional embedding and layer normalization then yields hierarchical visual representations $\boldsymbol{V}^{hier} \in \mathbb{R}^{B\times s\times d}$ capturing multi-level cues.

Prior-aware coarse-to-fine decoding. Motivated by principles of visual cognition and radiologists' diagnostic workflow, we propose a prior-aware coarse-to-fine decoding. Specifically, we first extract the condensed clinical context feature $\bar{\boldsymbol{T}}^c$ and the clinically informed spatiotemporal feature $\bar{\boldsymbol{V}}^{st}$ via Equations (3) and (4). These features are derived from the last hidden state of the text and vision encoders, and thus encode high-level semantics of patient clinical background and disease progression. We treat them as coarse-grained priors that offer a holistic overview for guiding subsequent fine-level decoding. We then enhance $\bar{\boldsymbol{V}}^{st}$ with hierarchical visual representations $\boldsymbol{V}^{hier}$ using the Perceiver architecture (Jaegle et al. 2021), formulated as:

$$\bar{\boldsymbol{V}}^{hier} = \text{Perceiver}(\bar{\boldsymbol{V}}^{st}, \boldsymbol{V}^{hier}). \tag{8}$$

The resulting $\bar{\boldsymbol{V}}^{hier}$ further refines and enriches fine-grained representations. Finally, we concatenate $\bar{\boldsymbol{T}}^c$, $\bar{\boldsymbol{V}}^{st}$, and $\bar{\boldsymbol{V}}^{hier}$ along the sequence dimension before feeding them into the report generator, thereby producing context-aware, disease progression-oriented reports.

Report generation. We adopt the pre-trained DistilGPT2 (Sanh et al. 2020) to generate free-text reports. Stage 2 is trained to minimize the cross-entropy loss between the AI-generated report $\hat{y}_i$ and the radiologist-generated report y_i. The objective function is formulated as:

$$\mathcal{L}_{CE}^i = -\sum_{k=1}^{K}\log p(\hat{y}_i^k \mid \hat{y}_i^{<k}, x_i^{cur}, x_i^{pri}, z_i, h_i), \tag{9}$$

where K denotes the maximum number of generated tokens. $\hat{y}_i^{<k}$ represents the tokens generated before step k.

4 Experiments

4.1 Experimental Settings

Datasets. (1) MIMIC-CXR (Johnson et al. 2019) is a large-scale, publicly available dataset of chest X-rays paired with free-text radiology reports. For each case, we organize data by "study id" and retrieve the most recent prior image when available. **(2) MIMIC-ABN** (Ni et al. 2020) is a curated subset of MIMIC-CXR that focuses exclusively on abnormal findings in reports. Following previous studies (Chen et al.

Dataset	Method	Venue	NLG Metrics ↑						CE Metrics ↑		
			B-1	B-2	B-3	B-4	MTR	R-L	P	R	F1
M-CXR	KiUT	CVPR'23	0.393	0.243	0.159	0.113	0.160	0.285	0.371	0.318	0.321
	METransformer	CVPR'23	0.386	0.250	0.169	0.124	0.152	0.291	0.364	0.309	0.311
	CoFE	ECCV'24	-	-	-	0.125	0.176	0.304	0.489	0.370	0.405
	DCG	MM'24	0.397	0.258	0.166	0.126	0.162	0.295	0.441	0.414	0.404
	MAN	AAAI'24	0.396	0.244	0.162	0.115	0.151	0.274	0.411	0.398	0.389
	R2GenGPT	Meta-Radio'23	0.411	0.267	0.186	0.134	0.160	0.297	0.392	0.387	0.389
	Med-LLM	MM'24	-	-	-	0.128	0.161	0.289	0.412	0.373	0.395
	R2-LLM	AAAI'24	0.402	0.262	0.180	0.128	0.175	0.291	0.465	0.482	0.473
	SEI	MICCAI'24	0.382	0.247	0.177	0.135	0.158	0.299	0.523	0.410	0.460
	HERGen	ECCV'24	0.395	0.248	0.169	0.122	0.156	0.285	-	-	-
	MPO	AAAI'25	**0.416**	0.269	0.191	0.139	0.162	0.309	0.436	0.376	0.353
	PriorRG (Ours)	-	0.412	**0.290**	**0.220**	**0.175**	**0.189**	**0.324**	**0.541**	**0.485**	**0.511**
	△(%) ↑	-	-0.4	+2.1	+2.9	+3.6	+1.3	+1.5	+1.8	+0.3	+3.8
M-ABN	CMN$^{\diamondsuit}$	ACL'21	0.256	0.147	0.095	0.066	0.110	0.230	0.466	0.454	0.460
	SEI$^{\diamondsuit}$	MICCAI'24	0.267	0.157	0.104	0.073	0.114	0.231	0.466	0.408	0.435
	PriorRG (Ours)	-	**0.326**	**0.201**	**0.139**	**0.102**	**0.140**	**0.242**	**0.467**	**0.476**	**0.471**
	△(%) ↑	-	+5.9	+4.4	+3.5	+2.9	+2.6	+1.1	+0.1	+2.2	+1.1

Table 1: Comparison of report generation performance with SOTA methods on MIMIC-CXR (M-CXR) and MIMIC-ABN (M-ABN) datasets. △ indicates the performance difference between PriorRG and the best baseline. ◊ denotes reproduced results; others are cited from the original papers. The **best** and second-best values are in **bold** and underlined, respectively.

2020; Liu et al. 2024b; Xiao et al. 2025), we treat only the "Findings" section as the radiologist-generated report, discarding empty or non-informative entries. All experiments adhere to the official dataset partitions. Dataset statistics are presented in Appendix Table 11.

Evaluation metrics. Following protocols in (Wang et al. 2023a; Liang et al. 2024; Liu et al. 2024a), we evaluate the quality of AI-generated reports using both natural language generation (NLG) and clinical efficacy (CE) metrics. For NLG metrics, which measure linguistic similarities between AI- and radiologist-generated reports, we use BLEU-n (B-n), METEOR (MTR), and ROUGE-L (R-L). For CE metrics, we assess clinical relevance by computing the micro-average Precision (P), Recall (R), and F1-score (F1) across 14 observations annotated by CheXpert (Irvin et al. 2019).

Implementation details. The unified feature dimension is set to $d = 768$, the number of latents to $N = 128$, the maximum generation length to $K = 100$ tokens, and the beam size to 3 during inference. We apply early stopping and learning rate scheduling for training stability. Additional implementation details are in Appendix Section A.

4.2 Main Results

Baseline approaches. We compare our PriorRG against 12 state-of-the-art (SOTA) methods, categorized into eight groups: knowledge graph-based methods (KiUT (Huang, Zhang, and Zhang 2023) and METransformer (Wang et al. 2023a)), contrastive learning framework (CoFE (Li et al. 2025b)), retrieval-augmented method (DCG (Liang et al. 2024)), memory alignment approaches (CMN (Chen et al. 2021) and MAN (Shen et al. 2024)), LLM-based approaches (R2GenGPT (Wang et al. 2023b), Med-LLM (Liu et al. 2024d), and R2-LLM (Liu et al. 2024a)), clinical context-driven model (SEI (Liu et al. 2024b)), longitudinal data-based method (HERGen (Wang, Du, and Yu 2025)), and human preference optimization (MPO (Xiao et al. 2025)).

Comparison with SOTA methods. Table 1 presents a comparison between our PriorRG and SOTA methods on MIMIC-CXR (Johnson et al. 2019) and MIMIC-ABN (Ni et al. 2020) datasets. The results show that our PriorRG consistently surpasses existing methods in both linguistic quality and clinical accuracy. On the MIMIC-CXR dataset, PriorRG achieves a notable 3.6% improvement in B-4, indicating better matching accuracy in longer n-gram sequences. Additionally, gains in MTR (+1.3%) and R-L (+1.5%) highlight improvements in lexical diversity, synonym matching, and semantic consistency. In terms of CE metrics, PriorRG boosts the F1 by 3.8%, demonstrating greater reliability in clinical accuracy. On the MIMIC-ABN dataset, PriorRG achieves a 5.9% increase in B-1 and a 1.1% improvement in F1, emphasizing its effectiveness in describing abnormal findings. These improvements stem from our proposed prior-aware coarse-to-fine decoding, which enables the model to recognize individual clinical contexts and capture disease progression. Appendix presents extra results (Table 5) and comparisons with MLRG (Tables 6–8).

Clinical accuracy of 14 observations. Appendix Table 10 presents the clinical accuracy of 14 observations extracted by the CheXpert (Irvin et al. 2019). Our PriorRG outperforms the clinically informed SEI model (Liu et al. 2024b) on 13 of 14 observations in terms of F1-score, highlighting the superior clinical accuracy of generated reports.

Evaluation with large language model. We assess clinical consistency and expert alignment of generated reports

Model	Stage 1	Stage 2			NLG Metrics ↑						CE Metrics ↑		
		CC	PI	Hidden States	B-1	B-2	B-3	B-4	MTR	R-L	P	R	F1
(a)	✓	✗	✗	✗	0.355	0.219	0.149	0.108	0.154	0.262	0.521	0.431	0.472
(b)	✓	✗	✗	✓	0.357	0.216	0.143	0.102	0.154	0.255	0.513	0.480	0.496
(c)	✓	✓	✗	✗	0.405	0.283	0.214	0.170	0.186	0.320	0.517	0.460	0.487
(d)	✓	✓	✗	✓	0.404	0.285	0.217	0.173	0.187	**0.328**	0.540	0.477	0.507
(e)	✓	✓	✓	✗	0.400	0.282	0.214	0.171	0.186	0.323	0.539	0.465	0.499
(f)	✗	✓	✓	✓	0.387	0.271	0.206	0.165	0.180	0.320	0.526	0.408	0.459
PriorRG	✓	✓	✓	✓	**0.412**	**0.290**	**0.220**	**0.175**	**0.189**	0.324	**0.541**	**0.485**	**0.511**

Table 2: Ablation study on MIMIC-CXR dataset. "CC" and "PI" denote clinical context and prior image, respectively.

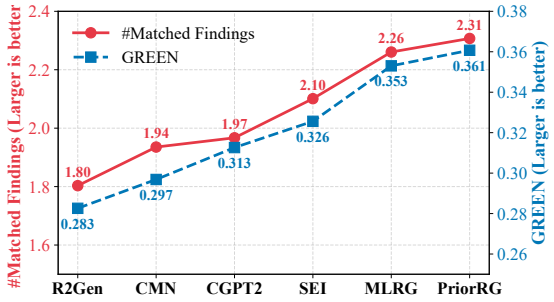


Figure 4: Expert-aligned assessment on the MIMIC-CXR dataset. "#Matched Findings" indicates the average matched findings between AI- and radiologist-generated reports.

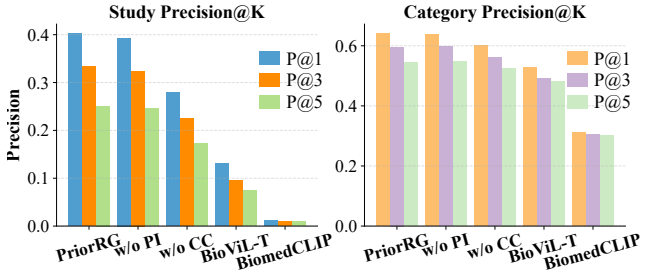


Figure 5: Medical image-text retrieval results on the MIMIC-5x200 dataset. "PI" and "CC" denote prior images and clinical context, respectively.

using two metrics: "#Matched Findings" and the GREEN score (Ostmeier et al. 2024). The GREEN score combines clinical significance errors with "#Matched Findings," offering a robust evaluation aligned with expert preferences. Both metrics are derived via the pre-trained GREEN-RadLlama2-7B model. We benchmark PriorRG against CMN (Chen et al. 2021), CGPT2 (Nicolson, Dowling, and Koopman 2023), SEI (Liu et al. 2024b), and MLRG (Liu et al. 2025a). Figure 4 shows that PriorRG significantly outperforms all baselines across both metrics, confirming its ability to generate clinically reliable, expert-aligned reports.

4.3 Ablation Study

Effect of prior-guided contrastive pre-training (Stage 1). We evaluate Stage 1 using a medical image-text retrieval task on the MIMIC-5x200 dataset (Johnson et al. 2019). Following the protocols in (Zhang et al. 2022; Zhou et al. 2024a), we construct the MIMIC-5x200 dataset by sampling 200 test set instances for each of five common diseases—*Atelectasis*, *Cardiomegaly*, *Consolidation*, *Edema*, and *Pleural Effusion*—from MIMIC-CXR, resulting in a total of 1,000 samples. Given a query image, we retrieve the top-K most similar reports from this dataset and evaluate performance using Category-Precision@K (Cat-P@K) and Study-Precision@K (Stu-P@K). Cat-P@K measures whether retrieved reports belong to the same disease category as the query image, while Stu-P@K evaluates whether retrieved reports originate from the same study. As illustrated in Figure 5, PriorRG surpasses BiomedCLIP (Zhang et al. 2023a), BioViL-T (Bannur et al. 2023), and ablated variants across both metrics. These results indicate that Stage 1 effectively encodes prior knowledge into semantically rich multimodal representations. Furthermore, Table 2 shows that removing Stage 1 (variant (f)) leads to significant drops in both clinical accuracy and language quality, confirming its central role in improving downstream report generation.

Effect of prior-aware coarse-to-fine decoding for report generation (Stage 2). As shown in Table 2, PriorRG significantly outperforms the baseline (a), which omits Stage 2, across both NLG and CE metrics, highlighting the critical role of Stage 2 in improving overall report quality.

Effect of patient-specific prior knowledge. Compared to variant (b) in Table 2, which excludes patient-specific prior knowledge, PriorRG shows a marked improvement in both NLG and CE metrics. This result reveals the positive impact of such prior knowledge on chest X-ray report generation. Additional evidence can be found in Table 4.

Effect of the clinical context (CC). In the image-text retrieval task (Figure 5), PriorRG with CC outperforms its counterpart without CC, improving cross-modal alignment. In the report generation task, adding CC to variant (a) (see Table 2) yields consistent gains across both NLG and CE metrics. This result suggests that CC enhances the model's ability to capture individual information, improving the quality of generated reports.

Effect of prior images (PI). In the image-text retrieval task (Figure 5), PriorRG with PI outperforms its variant without PI in the Stu-P@K metric, demonstrating its ability to model lesion progression through temporal image com-

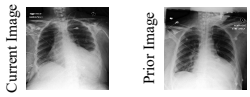
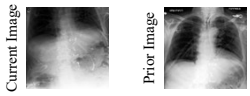
Input Data	Radiologist-generated Report	AI-generated Report by SEI	AI-generated Report by PriorRG (Ours)
 "[INDICATION][SEP][HISTORY] ___-year-old woman with **respiratory failure**. Evaluate fluid status."	Comparison is made to the prior study from ___. The feeding tube, left IJ catheter and endotracheal tube are unchanged in position. There is persistent cardiomegaly. There is unchanged left retrocardiac opacity. There are no signs for overt pulmonary edema. There is a small right-sided pleural effusion as well. Overall, these findings are stable.	in comparison with the earlier study of this date, there has been placement of an endotracheal tube with its tip approximately 5 cm above the carina. otherwise, little change.	Comparison is made to prior study from ___. The **endotracheal tube, feeding tube, and left IJ central line are unchanged in position**. There is **unchanged cardiomegaly**. There is **a persistent left retrocardiac opacity** and left-sided pleural effusion which is stable. There are no signs for overt pulmonary edema. No pneumothoraces are identified.
 "[INDICATION] **Recent Dobbhoff tube placement**.[SEP][HISTORY]"	As compared to the previous radiograph, the Dobbhoff tube shows now normal course. The tip projects over the middle parts of the stomach. No complications, notably no pneumothorax. Otherwise, the image is unchanged.	as compared to the previous radiograph, the dobbhoff tube has been repositioned. the course of the tube is unremarkable, the tip is not included on the image. otherwise, there is no relevant change.	As compared to the previous radiograph, the patient has received a Dobbhoff catheter. The course of the catheter is unremarkable, the tip of the catheter projects over the middle parts of the stomach. There is no evidence of complications, notably no pneumothorax. Otherwise, the radiograph is unchanged.

Figure 6: Qualitative comparison of reports from the baseline (Liu et al. 2024b), our proposed PriorRG, and radiologists. Sentences in the radiologist-generated report are color-coded to match corresponding descriptions in the AI-generated reports. In PriorRG's output, correct disease progression patterns and failure descriptions are marked in **bold** and underlined, respectively.

Model	NLG Metrics ↑			CE Metrics ↑		
	B-2	B-4	MTR	P	R	F1
LastOnly	0.278	0.168	0.185	**0.554**	0.466	0.506
Fine2coarse	0.284	0.172	0.187	0.549	0.482	**0.514**
PriorRG	**0.290**	**0.175**	**0.189**	0.541	**0.485**	0.511

Table 3: Comparison of different progressive fusion strategies for report generation on the MIMIC-CXR dataset.

Metric	w/ PI	w/o PI	w/ CC	w/o CC	w/ PK	w/o PK
B-2 ↑	**0.289**	0.288	**0.294**	0.139	**0.294**	0.140
F1 ↑	**0.516**	0.514	**0.508**	0.492	**0.512**	0.497

Table 4: Report generation performance on the MIMIC-CXR dataset, grouped by presence or absence CC (clinical context), PI (prior image), or PK (prior knowledge).

parison. In the RRG task (Table 2), comparisons such as (c) vs. (e) and PriorRG vs. (d) show that incorporating PI consistently improves both NLG and CE metrics. These results highlight the value of prior images in enhancing study-level retrieval and improving report generation quality.

Effect of hierarchical visual features. PriorRG's gain over variant (e) in Table 2 confirms that incorporating hierarchical visual features enhances report quality by capturing low-level details and high-level semantic cues.

Impact of progressive fusion strategies on report generation. Table 3 compares three progressive fusion strategies for the RRG task. The **LastOnly** variant utilizes only the final fused representation $\bar{V}^{hier}$, discarding intermediate representations. The **Fine2coarse** variant is structurally identical to PriorRG, except that it reverses the order of spatiotemporal and hierarchical features integration. Results show that PriorRG consistently outperforms both variants across all NLG metrics, demonstrating the effectiveness of coarse-to-fine fusion. While Fine2coarse achieves a slightly higher F1-score, it falls behind PriorRG in linguistic fluency. The LastOnly variant performs worst, emphasizing the importance of retaining intermediate representations to support richer, more coherent report generation.

Qualitative analysis. Figure 6 presents a comparative visualization of generated reports on the MIMIC-CXR dataset. In AI-generated reports, more diverse colors indicate broader coverage of clinical findings, while longer bars reflect more accurate and detailed descriptions. The results show that: (1) PriorRG generates high-quality drafts requiring minimal revision—for example, in Case 1, only a brief note on *pleural effusion* is added by the radiologist. (2) PriorRG captures disease progression effectively, as seen in the correct description of *unchanged cardiomegaly* in Case 1. (3) PriorRG responds to clinical context—e.g., in Case 2, the model correctly addresses concern about *Dobbhoff tube placement* and identifies potential complications. These examples demonstrate PriorRG's ability to track disease progression and understand clinical intent. Additional examples and failure case analysis are provided in Appendix B.

Conclusion

We introduced **PriorRG**, a chest X-ray report generation framework that leverages clinical priors to enhance both image-text alignment and report quality. Specifically, we proposed a prior-guided contrastive pre-training scheme that mirrors diagnostic reasoning by using clinical context to guide spatiotemporal features extraction. Next, we presented a prior-aware coarse-to-fine decoding that progressively integrates clinical context, disease progression patterns, and hierarchical visual cues, enhancing the clinical accuracy and fluency of generated reports. Extensive experiments validated the effectiveness of our PriorRG on medical image-text retrieval and report generation. Future work will explore the organ-aware diagnosis framework (Gu et al. 2024) to further enhance interpretability.

Acknowledgments

The work was jointly supported by the National Natural Science Foundations of China [grant number: 62272364], the Provincial Key Research and Development Program of Shaanxi [grant number: 2024GH-ZDXM-47], the Fundamental Research Funds for the Central Universities and the Innovation Fund of Xidian University [grant number: YJSJ25012].

References

Bannur, S.; Hyland, S.; Liu, Q.; Perez-Garcia, F.; Ilse, M.; Castro, D. C.; Boecking, B.; Sharma, H.; Bouzid, K.; Thieme, A.; et al. 2023. Learning to exploit temporal structure for biomedical vision-language processing. In *CVPR*, 15016–15027.

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

Chen, P.; Li, H.; Zhu, C.; Zheng, S.; Shui, Z.; and Yang, L. 2024. WsiCaption: Multiple Instance Generation of Pathology Reports for Gigapixel Whole-Slide Images. In *MICCAI*, 546–556.

Chen, Z.; Li, G.; and Wan, X. 2022. Align, reason and learn: Enhancing medical vision-and-language pre-training with knowledge. In *ACM MM*, 5152–5161.

Chen, Z.; Shen, Y.; Song, Y.; and Wan, X. 2021. Cross-modal Memory Networks for Radiology Report Generation. In *ACL*, volume 1, 5904–5914.

Chen, Z.; Song, Y.; Chang, T.-H.; and Wan, X. 2020. Generating Radiology Reports via Memory-driven Transformer. In *EMNLP*, 1439–1449.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.

Gu, T.; Liu, D.; Li, Z.; and Cai, W. 2024. Complex Organ Mask Guided Radiology Report Generation. In *WACV*, 7995–8004.

Guo, Z.; Ma, J.; Xu, Y.; Wang, Y.; Wang, L.; and Chen, H. 2024. Histgen: Histopathology report generation via local-global feature encoding and cross-modal context interaction. In *MICCAI*, 189–199.

Hamamci, I. E.; Er, S.; and Menze, B. 2024. Ct2rep: Automated radiology report generation for 3d medical imaging. In *MICCAI*, 476–486.

Hu, J.; and Li, Z. 2024. Distilled Cross-Combination Transformer for Image Captioning with Dual Refined Visual Features. In *ACM MM*, 4465–4474.

Huang, S.-C.; Shen, L.; Lungren, M. P.; and Yeung, S. 2021. GLoRIA: A Multimodal Global-Local Representation Learning Framework for Label-Efficient Medical Image Recognition. In *ICCV*, 3942–3951.

Huang, Z.; Zhang, X.; and Zhang, S. 2023. KiUT: Knowledge-injected U-Transformer for Radiology Report Generation. In *CVPR*, 19809–19818.

Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghgoo, B.; Ball, R.; Shpanskaya, K.; Seekins, J.; Mong, D. A.; Halabi, S. S.; Sandberg, J. K.; Jones, R.; Larson, D. B.; Langlotz, C. P.; Patel, B. N.; Lungren, M. P.; and Ng, A. Y. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI*, volume 33, 590–597.

Izhar, A.; Japar, N.; Idris, N.; and Dang, T. 2025. MicarVL-MoE: A Modern Gated Cross-Aligned Vision-Language Mixture of Experts Model for Medical Image Captioning and Report Generation. arXiv:2504.20343.

Jaegle, A.; Gimeno, F.; Brock, A.; Vinyals, O.; Zisserman, A.; and Carreira, J. 2021. Perceiver: General Perception with Iterative Attention. In *ICML*, volume 139, 4651–4664.

Jeong, J.; Tian, K.; Li, A.; Hartung, S.; Adithan, S.; Behzadi, F.; Calle, J.; Osayande, D.; Pohlen, M.; and Rajpurkar, P. 2024. Multimodal image-text matching improves retrieval-based chest x-ray report generation. In *MIDL*, 978–990.

Jin, H.; Che, H.; Lin, Y.; and Chen, H. 2024. PromptMRG: Diagnosis-Driven Prompts for Medical Report Generation. In *AAAI*, volume 38, 2607–2615.

Johnson, A. E. W.; Pollard, T. J.; Greenbaum, N. R.; Lungren, M. P.; ying Deng, C.; Peng, Y.; Lu, Z.; Mark, R. G.; Berkowitz, S. J.; and Horng, S. 2019. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. arXiv:1901.07042.

Li, C.-Y.; Chang, K.-J.; Yang, C.-F.; Wu, H.-Y.; Chen, W.; Bansal, H.; Chen, L.; Yang, Y.-P.; Chen, Y.-C.; Chen, S.-P.; et al. 2025a. Towards a holistic framework for multimodal LLM in 3D brain CT radiology report generation. *Nature Communications*, 16(1): 2258.

Li, M.; Lin, B.; Chen, Z.; Lin, H.; Liang, X.; and Chang, X. 2023. Dynamic Graph Enhanced Contrastive Learning for Chest X-Ray Report Generation. In *CVPR*, 3334–3343.

Li, M.; Lin, H.; Qiu, L.; Liang, X.; Chen, L.; Elsaddik, A.; and Chang, X. 2025b. Contrastive Learning with Counterfactual Explanations for Radiology Report Generation. In *ECCV*, 162–180.

Liang, X.; Zhang, Y.; Wang, D.; Zhong, H.; Li, R.; and Wang, Q. 2024. Divide and Conquer: Isolating Normal-Abnormal Attributes in Knowledge Graph-Enhanced Radiology Report Generation. In *ACM MM*, 4967–4975.

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

Liu, K.; Ma, Z.; Kang, X.; Li, Y.; Xie, K.; Jiao, Z.; and Miao, Q. 2025a. Enhanced Contrastive Learning with Multi-view Longitudinal Data for Chest X-ray Report Generation. In *CVPR*, 10348–10359.

Liu, K.; Ma, Z.; Kang, X.; Zhong, Z.; Jiao, Z.; Baird, G.; Bai, H.; and Miao, Q. 2024b. Structural Entities Extraction and Patient Indications Incorporation for Chest X-Ray Report Generation. In *MICCAI*, 433–443.

Liu, K.; Ma, Z.; Liu, M.; Jiao, Z.; Kang, X.; Miao, Q.; and Xie, K. 2024c. Factual Serialization Enhancement: A Key Innovation for Chest X-ray Report Generation. arXiv:2405.09586.

Liu, R.; Li, M.; Zhao, S.; Chen, L.; Chang, X.; and Yao, L. 2024d. In-Context Learning for Zero-shot Medical Report Generation. In *ACM MM*, 8721–8730.

Liu, T.; Wang, J.; Hu, Y.; Li, M.; Yi, J.; Chang, X.; Gao, J.; and Yin, B. 2025b. HC-LLM: Historical-Constrained Large Language Models for Radiology Report Generation. In *AAAI*, volume 39, 5595–5603.

Mei, X.; Mao, R.; Cai, X.; Yang, L.; and Cambria, E. 2024. Medical Report Generation via Multimodal Spatio-Temporal Fusion. In *ACM MM*, 4699–4708.

Mei, X.; Yang, L.; Gao, D.; Cai, X.; Han, J.; and Liu, T. 2025. Adaptive Medical Topic Learning for Enhanced Fine-grained Cross-modal Alignment in Medical Report Generation. *IEEE Transactions on Multimedia*, 1–12.

Miao, Q.; Liu, K.; Ma, Z.; Li, Y.; Kang, X.; Liu, R.; Liu, T.; Xie, K.; and Jiao, Z. 2025. EVOKE: Elevating Chest X-ray Report Generation via Multi-View Contrastive Learning and Patient-Specific Knowledge. arXiv:2411.10224.

Nguyen, D.; Chen, C.; He, H.; and Tan, C. 2023. Pragmatic Radiology Report Generation. In *ML4H*, volume 225, 385–402.

Ni, J.; Hsu, C.; Gentili, A.; and McAuley, J. J. 2020. Learning Visual-Semantic Embeddings for Reporting Abnormal Findings on Chest X-rays. In *EMNLP*, 1954–1960.

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

Ostmeier, S.; Xu, J.; Chen, Z.; Varma, M.; Blankemeier, L.; Bluethgen, C.; Md, A. E. M.; Moseley, M.; Langlotz, C.; Chaudhari, A. S.; and Delbrouck, J.-B. 2024. GREEN: Generative Radiology Report Evaluation and Error Notation. In *EMNLP*, 374–390.

Pérez-García, F.; Sharma, H.; Bond-Taylor, S.; Bouzid, K.; Salvatelli, V.; Ilse, M.; Bannur, S.; Castro, D. C.; Schwaighofer, A.; Lungren, M. P.; Wetscherek, M.; Codella, N.; Hyland, S. L.; Alvarez-Valle, J.; and Oktay, O. 2024. RAD-DINO: Exploring Scalable Medical Image Encoders Beyond Text Supervision. arXiv:2401.10815.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108.

Shen, H.; Pei, M.; Liu, J.; and Tian, Z. 2024. Automatic Radiology Reports Generation via Memory Alignment Network. In *AAAI*, volume 38, 4776–4783.

Wang, F.; Du, S.; and Yu, L. 2025. HERGen: Elevating Radiology Report Generation with Longitudinal Data. In *ECCV*, 183–200.

Wang, F.; Zhou, Y.; Wang, S.; Vardhanabhuti, V.; and Yu, L. 2022a. Multi-granularity cross-modal alignment for generalized medical visual representation learning. In *NeurIPS*, volume 35, 33536–33549.

Wang, Z.; Liu, L.; Wang, L.; and Zhou, L. 2023a. METransformer: Radiology Report Generation by Transformer with Multiple Learnable Expert Tokens. In *CVPR*, 11558–11567.

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

Wang, Z.; Sun, Y.; Li, Z.; Yang, X.; Chen, F.; and Liao, H. 2025. LLM-RG4: Flexible and Factual Radiology Report Generation across Diverse Input Contexts. In *AAAI*, volume 39, 8250–8258.

Wang, Z.; Wu, Z.; Agarwal, D.; and Sun, J. 2022b. MedCLIP: Contrastive Learning from Unpaired Medical Images and Text. In *EMNLP*, 3876–3887.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. CBAM: Convolutional Block Attention Module. In *ECCV*, 3–19.

Xiao, T.; Shi, L.; Liu, P.; Wang, Z.; and Bai, C. 2025. Radiology Report Generation via Multi-objective Preference Optimization. In *AAAI*, volume 39, 8664–8672.

Yin, H.; Zhou, S.; Wang, P.; Wu, Z.; and Hao, Y. 2025. KIA: Knowledge-Guided Implicit Vision-Language Alignment for Chest X-Ray Report Generation. In *COLING*, 4096–4108.

Zeng, Z.; Zhang, H.; Lu, R.; Wang, D.; Chen, B.; and Wang, Z. 2023. Conzic: Controllable zero-shot image captioning by sampling-based polishing. In *CVPR*, 23465–23476.

Zhang, S.; Xu, Y.; Usuyama, N.; Xu, H.; Bagga, J.; Tinn, R.; Preston, S.; Rao, R.; Wei, M.; Valluri, N.; Wong, C.; Tupini, A.; Wang, Y.; Mazzola, M.; Shukla, S.; Liden, L.; Gao, J.; Crabtree, A.; Piening, B.; Bifulco, C.; Lungren, M. P.; Naumann, T.; Wang, S.; and Poon, H. 2023a. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv:2303.00915.

Zhang, X.; Meng, Z.; Lever, J.; and Ho, E. S. 2025. Libra: Leveraging temporal images for biomedical radiology analysis. In *ACL*, 17275–17303.

Zhang, X.; Wu, C.; Zhang, Y.; Xie, W.; and Wang, Y. 2023b. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nature Communications*, 14(1): 4542.

Zhang, Y.; Jiang, H.; Miura, Y.; Manning, C. D.; and Langlotz, C. P. 2022. Contrastive Learning of Medical Visual Representations from Paired Images and Text. In *ML4H*, volume 182, 2–25.

Zhou, Y.; Faith, T. L. H.; Xu, Y.; Leng, S.; Xu, X.; Liu, Y.; and Goh, R. S. M. 2024a. BenchX: A Unified Benchmark Framework for Medical Vision-Language Pretraining on Chest X-Rays. In *NeurIPS*, volume 37, 6625–6647.

Zhou, Z.; Shi, M.; Wei, M.; Alabi, O.; Yue, Z.; and Vercauteren, T. 2024b. Large Model driven Radiology Report Generation with Clinical Quality Reinforcement Learning. arXiv:2403.06728.

Zhu, Q.; Mathai, T. S.; Mukherjee, P.; Peng, Y.; Summers, R. M.; and Lu, Z. 2023. Utilizing Longitudinal Chest X-Rays and Reports to Pre-fill Radiology Reports. In *MICCAI*, 189–198.