Seeing What Matters: Lesion-Aware High-Resolution Patch Discovery and Fusion for Chest X-ray Report Generation

Yingshu Li[*1], Yunyi Liu[*1], Zhenghao Chen[2], Tong Chen[1], Zailong Chen[3], Lingqiao Liu[4], Lei Wang[3], and Luping Zhou[†1]

[1] The University of Sydney, Sydney, NSW 2006, Australia
{yingshu.li,yunyi.liu1,luping.zhou}@sydney.edu.au,
tche2095@uni.sydney.edu.au
[2] The University of Newcastle, Newcastle, NSW 2308, Australia
zhenghao.chen@newcastle.edu.au
[3] University of Wollongong, Wollongong, NSW 2522, Australia
zc881@uowmail.edu.au, leiw@uow.edu.au
[4] The University of Adelaide, Adelaide, SA 5005, Australia
lingqiao.liu@adelaide.edu.au

Abstract. Despite rapid advances in chest X-ray (CXR) foundation models, most radiology report generation (RRG) systems still rely on heavily downsampled inputs (e.g., 256×256) due to the fixed visual token budgets of pretrained vision encoders, suppressing subtle yet clinically important cues present in native-resolution images. However, enabling high-resolution (high-res) perception remains challenging: naïve tiling causes prohibitive token inflation, while global compression suppresses subtle lesions and degrades diagnostic fidelity. Inspired by radiologists' workflow, localizing suspicious regions before detailed high-res assessment. We propose **Le**sion-**A**ware High-**R**esolution **Pa**tch Discovery and Fusion for Chest **X**-ray Reporting (LePaX), the first RRG framework that enables efficient high-res CXR perception (up to 1920×1920) without increasing the vision-token count. LePaX formulates high-res perception as a constrained spatial resolution allocation problem under a fixed token budget and introduces two key components: Learnable Spatial Resolution Allocation (LSRA), which learns a spatial utility map that adaptively allocates limited high-res capacity to diagnostically relevant regions, enabling targeted extraction of high-res patches from native CXRs; and Global–Regional Fusion (GRF), which performs token-preserving region-to-global refinement by projecting high-resolution regional evidence back onto the global feature grid through spatially aligned resolution write-back, avoiding token inflation. Experiments on multiple CXR benchmarks demonstrate that LePaX consistently improves both clinical and linguistic metrics while enabling native-resolution CXR perception with over 10× fewer visual tokens than naïve high-res tiling.

Keywords: Radiology Report Generation · High-Resolution Imaging · Multimodal Large Language Model

[*] Equal contribution. [†] Corresponding author

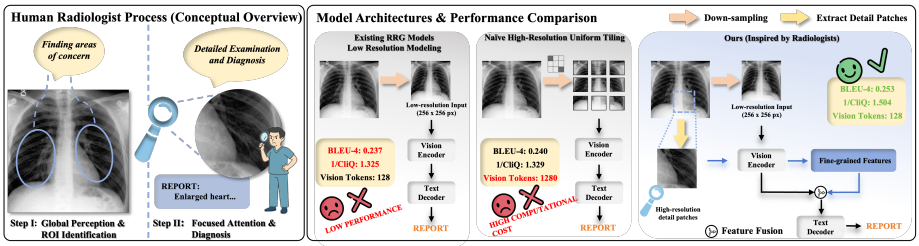


Fig. 1: Left: Radiologists typically follow a two-stage workflow: a global survey to localize suspicious regions, followed by focused examination of these areas to verify specific findings (e.g., enlarged heart). **Right**: Conventional RRG models rely on fixed-size, low-resolution inputs and lack this global–local reasoning process. Our radiologist-inspired framework instead extracts and integrates high-resolution regional details, producing reports with improved clinical fidelity.

1 Introduction

Radiology report generation (RRG) [18, 28, 30, 36, 49, 57] aims to automatically produce clinically faithful diagnostic narratives and reduce radiologists' workload. Recent advances in multimodal large language models (MLLMs) [3, 24, 48] have improved CXR report generation [28, 31, 32, 37, 44, 46]. However, most RRG systems rely on heavily downsampled inputs (e.g., 256×256) to satisfy fixed-size encoder constraints. While resizing simplifies visual encoding, it suppresses subtle but clinically important cues present in native resolution CXRs (raw CXRs are multi-megapixel). This is problematic because diagnostically relevant abnormalities are often small, faint, and spatially sparse. Consequently, findings such as nodules, fine vascular markings, or faint consolidations may be attenuated, weakening lesion localization and increasing the risk of diagnostic omissions.

However, enabling high-resolution perception in MLLMs is computationally challenging. Transformer-based vision encoders represent images as visual tokens whose number grows with spatial resolution, increasing computational cost. A common solution is uniform tiling, where the image is divided into patches, but this strategy is poorly suited for medical imaging. Because abnormalities in CXRs are often sparse and anatomically localized, uniform tiling allocates many visual tokens to non-diagnostic regions, consuming token budgets without improving diagnostic sensitivity (Fig. 1 right).

More fundamentally, such designs fail to reflect how radiologists interpret medical images. Clinical studies show that radiologists typically follow a global-to-local diagnostic workflow: they first perform a global survey to establish anatomical context and identify suspicious regions, and then zoom in for detailed verification [5, 38]. This workflow enables clinicians to allocate visual attention efficiently and integrate local evidence with global context. This raises a central question: *How can radiology-aligned high-resolution perception be modeled efficiently within MLLMs so that systems better reflect real diagnostic workflows?*

Inspired by this diagnostic process, we argue that high-resolution perception in medical imaging should be selective rather than uniform. Instead of processing all regions at the same resolution, models should allocate high-resolution modeling capacity to diagnostically informative areas while preserving global contextual awareness. We therefore **formulate high-resolution perception as a constrained spatial resolution allocation problem**, where the model determines where allocating additional high-resolution modeling capacity yields the greatest diagnostic utility under a fixed visual token budget. By enabling non-uniform high-resolution reasoning, this formulation explicitly mirrors the radiologists' global-to-local diagnostic strategy: global context is first established and then selectively refined with detailed local evidence.

Built on this formulation, we propose **LePaX** (Lesion-Aware High-Resolution Patch Discovery and Fusion), the first RRG framework enabling efficient high-resolution CXR perception (up to 1920×1920) without increasing visual token count. LePaX consists of two jointly optimized components: **Learnable Spatial Resolution Allocation (LSRA)** and **Global–Regional Fusion (GRF)**.

LSRA models the radiologists' "where to zoom" decision. It learns a resolution allocation policy over the global feature grid by predicting a spatial utility map that indicates where additional high-resolution modeling capacity is most beneficial. Under a fixed token budget, LSRA selects a small number of high-resolution patches from the native-resolution image, enabling the model to preserve subtle pathological patterns and fine anatomical structures that would otherwise be suppressed by aggressive downsampling. During training, the allocation policy is jointly guided by weak localization priors (e.g., Grad-CAM [39] priors from disease classifiers) and the end-to-end report generation objective, allowing the model to learn diagnostically meaningful high-resolution allocation strategies.

However, lesion-centric regions alone are insufficient for reliable diagnosis. Radiologists also rely on **global image** to extract secondary cues (e.g., bilateral symmetry, cardiothoracic ratio) that inform diagnostic decisions [13, 23]. To emulate this holistic reasoning process, we introduce GRF, which models the integration of local evidence with global context. GRF GRF performs spatially grounded resolution write-back, injecting high-resolution regional features into corresponding locations within the global feature grid. This spatially grounded fusion enriches global representations with fine-grained diagnostic evidence while preserving anatomical coherence, all without increasing the visual token budget.

Our key contributions are summarized as follows:

- We present **LePaX**, a radiologist-inspired high-resolution RRG framework that formulates lesion-sensitive perception as a *spatial resolution allocation* problem under fixed visual token budgets, explicitly reflecting the global-to-local diagnostic workflow.
- We introduce two jointly optimized modules: (1) **LSRA**, which learns a *resolution allocation policy* to decide where additional high-resolution capacity should be invested, and (2) **GRF**, a spatially grounded *resolution write-back refinement* mechanism that injects high-resolution evidence into the global representation without token explosion.

– Extensive experiments on **MIMIC-CXR**, **IU-Xray**, and **CheXpertPlus** demonstrate consistent gains on both NLG and clinical metrics.

2 Related Work

2.1 Radiology Report Generation

Radiology report generation (RRG) is a challenging task, and recent advances have improved performance [6,15,18,36,45]. Early methods mainly relied on encoder–decoder architectures, while more recent work adopts Transformer-based frameworks [41]. METransformer [45] introduces learnable expert tokens for multi-specialist reasoning. KiUT [18] incorporates structured medical knowledge into a U-shaped Transformer for disease-aware visual–text alignment. DART [36] proposes a disease-aware alignment framework with a self-correcting re-alignment mechanism to improve factual consistency and reduce hallucinations.

The emergence of multimodal large language models (MLLMs) [3, 24, 48] has shifted RRG from traditional encoder–decoder paradigms toward unified vision–language generation [8,25,27,28,31,37,44]. R2GenGPT [46] was among the first to leverage frozen LLMs with efficient visual–language alignment, achieving strong linguistic fluency with minimal training cost. LLaVA-Med [25] extends instruction-tuned LLMs to medical imaging via staged multimodal alignment and instruction tuning. Multi-Phased Supervision [8] adopts curriculum-style training from disease labels to entity–relation reasoning and full report generation. Most recently, MambaXray-VL [44] combines autoregressive visual pretraining and image–text contrastive alignment within an LLaMA-based decoder, achieving strong fluency and clinically grounded reports.

Despite these advances, most frameworks operate on low-resolution CXRs, treating visual tokens as uniformly informative and overlooking the global-to-local reasoning radiologists use to link sparse abnormalities with textual evidence. This limitation motivates our **Lesion-Aware High-Resolution Patch Discovery and Fusion for Chest X-ray Report Generation (LePaX)**, a high-resolution yet efficient framework that preserves fine-grained anatomical details while maintaining global coherence.

2.2 High-resolution Multimodal Large Language Models

Recent multimodal large language models (MLLMs) process high-resolution or irregularly shaped images through tiling, cropping, or feature compression to balance spatial fidelity and efficiency [3,9,24,51,58]. LLaVA-OneVision [24] adopts the Higher AnyRes scheme with multi-crop inputs and interpolation to maintain token budgets; InternVL-1.5 [9] partitions images into 448×448 tiles with a global thumbnail and compresses features via pixel-shuffle; Qwen2.5-VL [3] employs dynamic-resolution ViTs with windowed attention and token merging for efficient native-scale encoding. Although effective for general MLLMs, these tiling and compression strategies often average fine-grained cues, weakening lesion localization and clinical grounding in radiology report generation (RRG).

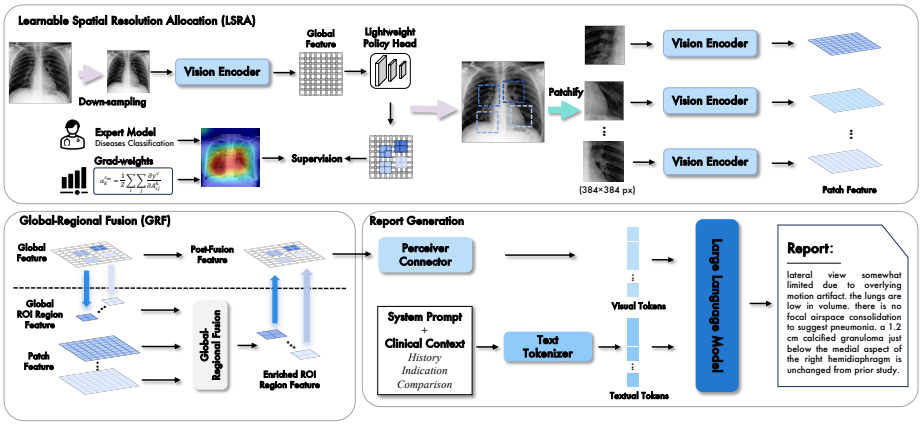


Fig. 2: Overview of the proposed **LePaX** framework. A lightweight policy head predicts a spatial utility map from global features to allocate a fixed number of high-resolution patches under a resolution budget (LSRA). Global and patch features share a vision encoder and are fused via utility-guided Global–Regional Fusion (GRF), which writes high-resolution evidence back to the global feature grid. The fused visual tokens are compressed by a Perceiver connector and passed to an LLM conditioned on clinical prompts to generate grounded radiology reports.

Unlike natural images, where dense spatial coverage is beneficial, medical imaging features *spatially sparse, anatomically localized*, and clinically weighted abnormalities. Uniform tiling or global merging thus introduces redundant, non-diagnostic tokens and suppresses disease-relevant evidence. In contrast, **LePaX** formulates high-resolution modeling as a **learnable spatial resolution allocation (LSRA)** problem. A utility map predicts informative regions where high-resolution patches are allocated, and the resulting regional features are written back to the global grid via a spatially grounded global–regional fusion (GRF) module, enabling detailed lesion modeling without increasing the token budget.

3 Methodology

3.1 Overall Framework

As illustrated in Fig. 2, **LePaX** is a lesion-aware radiology report generation framework that enables adaptive high-resolution perception under a fixed visual token budget. Rather than uniformly processing high-resolution images, LePaX allocates limited high-resolution capacity to diagnostically informative regions. The framework introduces two key mechanisms. **(1) Learnable Spatial Resolution Allocation (LSRA).** LSRA predicts a grid-level *resolution utility map* from global visual features, indicating where additional high-resolution modeling capacity should be allocated. Under a fixed token budget, LSRA selects a small set of spatial locations and extracts high-resolution patches from

high-resolution CXRs (up to 1920×1920 px), preserving subtle pathological cues while avoiding uniform high-resolution processing. **(2) Global–Regional Fusion (GRF).** The global image and selected patches are encoded by a shared backbone to ensure geometric alignment in feature space. GRF performs region-conditioned update within the selected regions of interest (ROIs), injecting high-resolution evidence back into their spatially corresponding global tokens without increasing the visual token budget. This produces fused visual representations $F_g^{\text{fused}} \in \mathbb{R}^{N_v \times D_v}$. The fused representation is then processed by a Perceiver connector and used to condition the LLM for radiology report generation.

Visual Encoder. The global image and selected high-resolution patches are encoded by a shared vision backbone ϕ_{img} to maintain geometric alignment in feature space:

$$F = \phi_{\text{img}}(X) \in \mathbb{R}^{N_v \times D_v}. \tag{1}$$

We denote the global tokens as F_g and the k-th patch tokens as F_p^k.

Perceiver Connector. To align fused visual features with the language model, we employ a Perceiver connector that compresses F_g^{fused}:

$$\begin{aligned} F_c &= \text{Perceiver}(F_g^{\text{fused}}) \\ &= \text{CrossAttn}\big(Q_0,\ K{=}[F_g^{\text{fused}}; Q_0],\ V{=}[F_g^{\text{fused}}; Q_0]\big)\ \in \mathbb{R}^{N_c \times D_c},\ N_c \ll N_v. \end{aligned} \tag{2}$$

where $Q_0 \in \mathbb{R}^{N_c \times D_c}$ denotes learnable latent tokens and $[\cdot\,;\cdot]$ denotes token-wise concatenation.

Text Decoder. Conditioned on F_c and textual prompt C, the LLM autoregressively generates the radiology report $Y = \{y_1, \ldots, y_T\}$:

$$p(y_t \mid F_c, C, y_{<t}) = \text{LLM}(F_c, C, y_{<t}). \tag{3}$$

3.2 Resolution Allocation Formulation

Let $\mathbf{X} \in \mathbb{R}^{H_0 \times W_0 \times 3}$ denote a high-resolution chest X-ray and $\mathbf{X}_{\text{low}}$ its downsampled version used for policy prediction. Given a fixed patch budget B, we formulate report generation as a constrained spatial resolution allocation problem that assigns limited high-resolution modeling capacity to the most informative regions. We define a learnable allocation policy π_θ that selects spatial locations $\mathcal{S} = \{(i_k, j_k)\}_{k=1}^{K}$ with $K \leqslant B$. The objective is

$$\max_{\theta}\ \mathbb{E}_{(\mathbf{X},Y)\sim\mathcal{D}}\left[\log p_\phi(Y \mid \mathbf{X}_{\text{low}}, \pi_\theta(\mathbf{X}_{\text{low}}, \mathbf{X}))\right] \quad \text{s.t.} \quad |\pi_\theta(\mathbf{X}_{\text{low}}, \mathbf{X})| \leqslant B. \tag{4}$$

Here $\pi_\theta(\mathbf{X}_{\text{low}}, \mathbf{X})$ predicts spatial locations from $\mathbf{X}_{\text{low}}$ and crops the corresponding high-resolution patches from $\mathbf{X}$. LSRA parameterizes the allocation policy π_θ, while GRF integrates the selected high-resolution evidence with global representations without increasing the number of visual tokens.

3.3 Learnable Spatial Resolution Allocation (LSRA)

High-resolution CXRs contain subtle pathological cues, yet most vision encoders operate on low-resolution inputs, attenuating fine-grained diagnostic patterns. Meanwhile, clinically relevant findings are often spatially sparse, making uniform high-resolution processing computationally inefficient. To address this, LSRA learns a spatial utility function over the global feature grid, enabling adaptive high-resolution allocation under a fixed resolution budget.

Resolution Utility Learning. Let $F_g \in \mathbb{R}^{C\times H\times W}$ denote the global feature extracted from the low-resolution CXR. LSRA predicts a spatial allocation map

$$Z = g(F_g) \in \mathbb{R}^{1\times H\times W}, \qquad U = \sigma(Z), \tag{5}$$

where U_{ij} represents the utility of allocating additional modeling capacity to location (i, j).

Disease-Aware Spatial Prior. Learning a spatial allocation policy from report supervision alone can be unstable during early optimization. Since most medical datasets lack pixel-level lesion annotations, we introduce a disease-aware spatial bias by deriving a coarse prior using Grad-CAM from a pretrained multi-label disease classifier f_{cls}. Given an input image X, let $A \in \mathbb{R}^{C_a\times H_0\times W_0}$ denote the final convolutional feature map and y^c the logit of class c. The Grad-CAM activation for class c is

$$H^c = \text{ReLU}\left(\sum_{k=1}^{C_a} \alpha_k^c A^k\right), \quad \alpha_k^c = \frac{1}{H_0 W_0}\sum_{i,j}\frac{\partial y^c}{\partial A_{ij}^k}. \tag{6}$$

Class-wise activation maps are aggregated and normalized to obtain a spatial prior M^{cam}, providing weak spatial guidance for the allocation policy. The report generation loss further refines the utility toward task-relevant regions. We implement this prior as a regularization term on the allocation logits:

$$\mathcal{L}_{\text{prior}} = \text{BCEWithLogits}(Z, M^{\text{cam}}). \tag{7}$$

This prior is used only to regularize LSRA during training; at inference, patch locations are selected solely from the learned utility map U, without computing Grad-CAM or invoking any external disease classifier.

Allocation Strategy. Given the utility map $U = \sigma(Z)$, LSRA allocates a fixed high-resolution budget by selecting the top-K spatial locations on the grid.

Training. To enable gradient-based optimization of the allocation policy, we employ a Gumbel-Top-K relaxation with a straight-through estimator (STE). During training, Gumbel noise perturbs the allocation logits to produce stochastic rankings, from which a hard Top-K mask is obtained for patch extraction. $\mathcal{L}_{\text{policy}}$ provides a spatial prior for the allocation logits, while $\mathcal{L}_{\text{RRG}}$ (defined in Eq. 13) supplies additional task-driven signals, encouraging the learned utility map to prioritize regions beneficial for report generation.

Inference. At inference time, LSRA selects the K spatial locations with the highest utilities. To encourage spatial diversity and reduce redundant overlap,

we apply an NMS-like constraint that enforces a minimum grid distance $d_{\min}$:

$$\mathcal{S} = \{(i_k, j_k)\}_{k=1}^{K} = \text{TopK+NMS}(U; K, d_{\min}). \tag{8}$$

Each selected location defines a region-of-interest (ROI) Ω_k on the global feature grid. The ROI center is projected to the original image to obtain pixel coordinates (x_k, y_k), from which a fixed-size patch is extracted:

$$X_k = \text{Crop}\big(X; x_k, y_k, p_w, p_h\big), \qquad p_w = p_h = 384. \tag{9}$$

We emphasize that LSRA is not a Grad-CAM-based crop selector: M^{cam} only regularizes allocation learning during training, whereas inference selects patches solely from the learned utility map $U = \sigma(Z)$ without Grad-CAM or external classifiers. Thus, LSRA acts as an end-to-end adaptive resolution allocation policy for report generation.

3.4 Global-Regional Fusion (GRF)

The global image and selected patches are encoded by a shared backbone, producing F_g and $\{F_p^k\}_{k=1}^{K}$. Rather than expanding the visual token sequence by concatenating patch tokens, GRF performs token-budget-preserving resolution refinement by injecting localized high-resolution evidence back into the global representation. For each selected region Ω_k, we extract the corresponding global tokens $F_g^{\text{roi}} = F_g[\Omega_k]$ and perform a region-conditioned update:

$$\bar{F}_g^{\text{roi}} = F_g^{\text{roi}} + \mathcal{U}\big(F_g^{\text{roi}}, F_p^k\big)\,, \tag{10}$$

where $\mathcal{U}(\cdot)$ integrates high-resolution patch features into the aligned global tokens. In practice, $\mathcal{U}$ is implemented as a Transformer-style cross-attention block

$$\mathcal{U}(F_g^{\text{roi}}, F_p^k) = W_O\,\text{Attn}(Q, K, V) + \text{FFN}\big(\text{LN}(\cdot)\big), \tag{11}$$

with $Q = W_Q(F_g^{\text{roi}}+P_Q), K = W_K(F_p^k+P_K), V = W_V(F_p^k+P_K)$. The refined tokens are written back to their original spatial coordinates $F_g[\Omega_k] \leftarrow \bar{F}_g^{\text{roi}}$, yielding the fused representation F_g^{fused}.

3.5 Enhanced Report Generation

The fused representation F_g^{fused} conditions the LLM together with clinical context C to generate the final report:

$$Y^* = \arg\max_{\hat{Y}\in\mathcal{Y}} \sum_{t=1}^{T} \log p(\hat{y}_t \mid F_g^{\text{fused}}, C, \hat{y}_{<t}). \tag{12}$$

3.6 Training Objective

The model is trained end-to-end with a joint objective:

$$\mathcal{L} = \mathcal{L}_{\text{RRG}} + \lambda \mathcal{L}_{\text{policy}}, \qquad \mathcal{L}_{\text{RRG}} = -\sum_{t=1}^{T} \log p(y_t \mid F_g^{\text{fused}}, C, y_{<t}). \tag{13}$$

$\mathcal{L}_{\text{policy}}$ provides spatial priors, while $\mathcal{L}_{\text{RRG}}$ supplies task-driven signals that encourage the allocation policy to focus on regions beneficial for report generation.

4 Experiments

4.1 Datasets

IU-Xray [12] contains 7,470 images and 3,955 reports and is a standard benchmark for RRG. Following [11], we adopt the same 7:1:2 split for training, validation, and testing, and report results on the official test set.
MIMIC-CXR [22] is a large-scale dataset containing 377,110 images and 227,835 reports from 64,588 patients. Following [11], we use the same split with 270,790 images for training and 3,858 for testing.
CheXpertPlus [7] is a large-scale benchmark for radiology modeling with 223,228 chest X-rays and 187,711 reports. Following CXPMRG-Bench [44], we use the *Findings* section as ground truth and adopt a 7:1:2 split for training, validation, and testing (40,463/5,780/11,562).

4.2 Experimental Settings

Evaluation Metrics. (1) NLG Metrics. Following prior work [11], we evaluate linguistic quality using BLEU [35], ROUGE-L [29], and METEOR [4]. **(2) Clinical Metrics.** For clinical correctness, we adopt widely used metrics: CheXpert Clinical Efficacy [19] evaluates alignment between generated findings and ground truth; RadGraph F_1 [20] measures overlap in entity–relation structures; BERTScore [54] assesses semantic similarity via contextual embeddings; RadCliQ [52] combines multiple factors and correlates with radiologist preference. **(3) LLM-based Metrics.** We further include two LLM-based metrics for human-aligned factuality: GREEN [34] evaluates factual consistency with clinical entities; RateScore [55] measures overall report quality including fluency and diagnostic accuracy.
Implementation Details. Our model builds on BLIP-3 [48] with a SigLIP vision encoder [53], a Perceiver Resampler [2], and a Phi-3-mini$_{3.8B}$ language model [1]. LSRA is implemented as a lightweight convolutional policy head on the global feature grid (e.g., 27×27), using a depthwise 3×3 convolution followed by two 1×1 convolutions to predict a resolution utility map. During training, Grad-CAM maps from a ResNet-34 [14, 39] provide a disease-informed spatial prior to stabilize policy learning (**not used at inference**). Since BLIP-3 is not designed for medical imaging, all components are fine-tuned: LoRA [16]

Table 1: Comparison on MIMIC-CXR and IU-Xray. Blue shading indicates the **Top-3** methods: darkest (1st), medium (2nd), light (3rd). Results marked with † are reported from published literature. *LLM-based models are labeled with their parameter size.*

Dataset	Methods	Publication	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE
MIMIC CXR	R2Gen† [11]	EMNLP 2020	0.353	0.218	0.145	0.103	0.142	-
	R2GenCMN† [10]	ACL-IJCNLP 2021	0.353	0.218	0.148	0.106	0.142	-
	CvT2DistilGPT2† [33]	AIM 2023	0.393	0.248	0.171	0.127	-	0.155
	METransformer† [45]	CVPR 2023	0.386	0.250	0.169	0.124	0.152	0.291
	DCL† [26]	CVPR 2023	-	-	-	0.109	0.150	0.284
	KiUT† [18]	CVPR 2023	0.393	0.243	0.159	0.113	0.160	0.285
	R2GenGPT† (7B) [46]	Meta-Rad 2023	0.411	0.267	0.186	0.134	0.160	0.297
	EKAGen† [6]	CVPR 2024	0.419	0.258	0.170	0.119	0.157	0.287
	Bootstrapping† (14.2B) [31]	AAAI 2024	0.402	0.262	0.180	0.128	0.175	0.291
	Multi-Grained† [30]	TMI 2024	0.406	0.267	0.190	0.141	0.163	0.309
	REVTAF-RRG† [56]	ICCV 2025	0.465	0.318	0.235	0.182	0.199	0.336
	DAMPER† [17]	AAAI 2025	0.402	0.284	0.227	0.193	0.289	0.301
	DATR† [36]	CVPR 2025	0.437	0.279	0.191	0.137	0.175	0.310
	MultiP-R2Gen† (7B) [8]	TMI 2025	0.425	0.279	0.194	0.140	0.167	0.307
	KACL† (8B) [40]	MICCAI 2025	0.414	0.270	0.184	0.136	0.169	0.303
	MambaXray-VL-Large† (7B) [44]	CVPR 2025	0.422	0.268	0.184	0.133	0.167	0.289
	Ours (4B)	ECCV 2026	**0.501**	**0.376**	**0.304**	**0.253**	**0.336**	**0.380**
IU-Xray	R2Gen† [11]	EMNLP 2020	0.470	0.304	0.219	0.165	0.187	0.371
	R2GenCMN† [10]	ACL-IJCNLP 2021	0.475	0.309	0.222	0.170	0.191	0.375
	METransformer† [45]	CVPR 2023	0.483	0.322	0.228	0.172	0.192	0.380
	KiUT† [18]	CVPR 2023	0.525	0.360	0.251	0.185	0.242	0.409
	R2GenGPT† (7B) [46]	Meta-Rad 2023	0.488	0.316	0.228	0.173	0.211	0.377
	EKAGen† [6]	CVPR 2024	0.497	0.339	0.250	0.190	0.210	0.399
	Bootstrapping† (14.2B) [31]	AAAI 2024	0.499	0.323	0.238	0.184	0.208	0.390
	Multi-Grained† [30]	TMI 2024	0.472	0.321	0.234	0.175	0.192	0.379
	REVTAF-RRG† [56]	ICCV 2025	0.420	0.249	0.159	0.107	0.176	0.309
	DAMPER† [17]	AAAI 2025	0.520	0.383	0.300	0.225	0.284	0.397
	DATR† [36]	CVPR 2025	0.486	0.348	0.265	0.208	0.205	0.411
	MultiP-R2Gen† (7B) [8]	TMI 2025	0.515	0.332	0.238	0.178	0.216	0.386
	KACL† (8B) [40]	MICCAI 2025	0.501	0.326	0.244	0.184	0.211	0.385
	MambaXray-VL-Large† (7B) [44]	CVPR 2025	0.491	0.330	0.241	0.185	0.216	0.371
	Ours (4B)	ECCV 2026	**0.531**	**0.393**	**0.324**	**0.235**	**0.316**	0.402

is applied to the vision encoder and language model (r=32, α=64), while the Perceiver Resampler is fully optimized. Training runs for three epochs using AdamW (1×10^{-4} LR, cosine schedule, warmup ratio 0.03) with batch size 16.

4.3 Main Results

We evaluate LePaX on MIMIC-CXR, IU-Xray, and CheXpertPlus using both NLG and clinical metrics.

NLG Metrics. As shown in Table 1 and Table 2, our method consistently outperforms prior approaches across all datasets and NLG metrics. On MIMIC-CXR, it achieves BLEU-4 / METEOR / ROUGE scores of 0.253 / 0.336 / 0.380, surpassing previous encoder–decoder, transformer, and LLM-based methods. On IU-Xray, it reaches 0.235 / 0.316 / 0.402, yielding an average 11% improvement over the previous best. On CheXpertPlus, our model also achieves state-of-the-art performance on both NLG and clinical efficacy metrics, demonstrating strong generalization. Together, these results indicate that our high-resolution patch-guided representation captures subtle disease cues while maintaining coherent global–local semantics, leading to accurate and clinically faithful reports.

Table 2: Experimental results on the CXPMRG-Bench [44]. NLG metrics: BLEU-4, ROUGE, METEOR. CE metrics: Precision, Recall, and F1.

Methods	Publish	BLEU-4	ROUGE	METEOR	Precision	Recall	F1
ORGan [15]	ACL 2023	0.086	0.261	0.135	0.288	0.287	0.277
M2KT [49]	MIA 2021	0.078	0.247	0.101	0.044	0.142	0.058
TIMER [47]	CHIL 2023	0.083	0.254	0.121	0.345	0.238	0.234
CvT2DistilGPT2 [33]	AIM 2023	0.067	0.238	0.118	0.285	0.252	0.246
R2Gen [11]	EMNLP 2020	0.081	0.246	0.113	0.318	0.200	0.181
R2GenCMN [10]	ACL 2021	0.087	0.256	0.127	0.329	0.241	0.231
Zhu et al. [59]	MICCAI 2023	0.074	0.235	0.128	0.217	0.308	0.205
CAMANet [42]	IEEE JBHI 2023	0.083	0.249	0.118	0.328	0.224	0.216
Token-Mixer [50]	IEEE TMI 2023	0.091	0.261	0.135	0.309	0.270	0.288
PromptMRG [21]	AAAI 2024	0.095	0.222	0.121	0.258	0.265	0.281
R2GenGPT [46]	Meta-Rad.2023	0.101	0.266	0.145	0.315	0.244	0.260
R2GenCSR [43]	arXiv 2024	0.100	0.265	0.146	0.315	0.247	0.259
MambaXray-VL-B [44]	CVPR 2025	0.105	0.267	0.149	0.333	0.264	0.273
MambaXray-VL-L [44]	CVPR 2025	0.112	**0.276**	0.157	0.377	0.319	0.335
Ours	ECCV 2026	**0.138**	0.252	**0.227**	**0.420**	**0.368**	**0.363**

Clinical Metrics. As shown in Table 3 and Table 4, our model achieves strong clinical fidelity across evaluation metrics. Since large-scale expert review of thousands of reports is impractical, we adopt RadCliQ and GREEN, human-aligned metrics trained on radiologist ratings that correlate strongly with expert judgments. On MIMIC-CXR, our model achieves the highest RadGraph F_1 (0.346) and BERTScore (0.526), together with a lower RadCliQ (0.665↓), indicating closer alignment with expert reports. It also obtains 0.413 GREEN and 0.627 RateScore on LLM-based evaluations. Under the CheXpert Clinical Efficacy framework, the model reaches 0.595 precision, 0.479 recall, and 0.531 F_1, surpassing recent transformer- and LLM-based baselines.

Table 3: Comparison with state-of-the-art methods on MIMIC-CXR. RG_{F1}: RadGraph F1; BERT: BERTScore; CliQ: RadCliQ; GREEN: clinical consistency score; Rate: human rating score.

Methods	Publication	RG_{F1}(↑)	BERT(↑)	CliQ(↓)	GREEN(↑)	Rate(↑)
R2Gen [11]	EMNLP 2020	0.172	0.406	1.228	0.276	0.526
CvT2DistilGPT2 [33]	AI in Med 2023	0.196	0.374	1.220	0.320	0.527
RaDialog-RG [37]	MIDL 2024	-	0.400	-	-	-
PromptMRG [21]	AAAI 2024	0.190	0.357	1.169	0.287	0.528
R2GenGPT [46]	Meta-Rad 2023	0.187	0.415	1.207	0.300	0.528
KARGEN [28]	MICCAI 2024	0.203	0.421	1.165	0.308	0.533
EKAGen [6]	CVPR 2024	0.199	0.412	1.126	0.256	0.512
MultiP-R2Gen [8]	TMI 2025	0.208	0.425	1.140	-	-
Ours	ECCV 2026	**0.346**	**0.526**	**0.665**	**0.413**	**0.627**

4.4 Ablation Study

Contribution of Each Component. We perform ablation studies on MIMIC-CXR (Table 5) to evaluate each component. The baseline without either module

Table 4: Evaluation of CheXpert clinical efficacy metrics on the MIMIC-CXR dataset.

Methods	Publication	Precision	Recall	F1
R2Gen [11]	EMNLP 2020	0.333	0.273	0.276
R2GenGPT [46]	Meta-Rad 2023	0.392	0.387	0.389
METransformer [45]	CVPR 2023	0.364	0.309	0.311
Multi-Grained [30]	TMI 2024	0.457	0.337	0.330
MambaXray-VL-Large [44]	CVPR 2025	0.371	0.321	0.340
KACL [40]	MICCAI 2025	0.503	0.442	0.469
DAMPER [17]	AAAI 2025	0.512	0.473	0.507
DART [36]	CVPR 2025	0.520	**0.546**	**0.533**
Ours	ECCV 2026	**0.595**	0.479	0.531

achieves the lowest performance, indicating limited ability to capture fine-grained pathological cues.

GRF. Introducing GRF with randomly sampled patches already improves performance (BLEU-4: +0.006, METEOR: +0.013, ROUGE: +0.010) under the same patch budget as LSRA, indicating that regional high-resolution evidence becomes beneficial when fused with global representations.

LSRA. Replacing random sampling with guided region selection further improves performance. Grad-CAM increases B-4 to 0.251, while LSRA achieves the best result (B-4: 0.253) under the same token budget (128) and comparable inference time, demonstrating the advantage of task-driven resolution allocation.

Policy supervision. Training LSRA with report-only supervision leads to less stable optimization, as the report objective provides indirect signals for discrete region allocation. Adding policy supervision improves performance, indicating that report supervision and spatial priors provide complementary guidance.

Fusion strategy. Simply concatenating tokens without GRF degrades performance and increases the input sequence length (e.g., $(1_{\text{global}} + 4_{\text{patches}}) \times 128 = 640$ tokens), resulting in much higher inference cost (6.5s per case).

Full model. The complete model maintains a fixed token budget (128 tokens) while enriching token semantics through GRF achieving the best performance with only a small inference overhead (2.7s per case). As shown in Fig. 5b, GRF increases feature variance and L2 norm, indicating that high-resolution regional evidence enhances token expressiveness without increasing token count.

Overall, these results highlight the complementary roles of LSRA and GRF: LSRA adaptively allocates resolution to informative regions, while GRF integrates regional evidence into global representation in a token-efficient manner.

Clinical Evaluation. We further evaluate clinical reliability using RadGraph F_1, BERTScore, and RadCliQ (Fig. 3(a)). **LePaX** improves all metrics, achieving relative gains of +**8.5%** in RadGraph F_1, +**5.0%** in BERTScore, and +**13.5%** in 1/RadCliQ. These results indicate improved factual correctness and clinical consistency in the generated reports.

Resolution Allocation Efficiency. We analyze the impact of input resolution on report generation (Table 6). Increasing resolution from 384 to 1024 yields only marginal gains while greatly increasing cost, with vision tokens rising from 128 to 1280 and inference time reaching 10.0s per case. In contrast, **LePaX** allocates

Table 5: Ablation study on MIMIC-CXR. GRF: Global–Regional Fusion; Grad-CAM: external localization-based patch selection; LSRA: Learnable Spatial Resolution Allocation. **LSRA Loss**: report-only supervision vs. report + policy supervision.

Modules			NLG Metrics			#Vision Tokens	Inference Time (case)
GRF	Grad-CAM	LSRA	B-4	METEOR	ROUGE		
–	–	–	0.237	0.313	0.350	128	2.3s
✓	–	–	0.243	0.326	0.360	128	2.6s
–	✓	–	0.234	0.318	0.334	∼640	6.5s
–	–	✓	0.235	0.318	0.333	∼640	6.5s
✓	✓	–	0.251	0.334	0.379	128	2.6s
✓	–	✓ (report-only)	0.246	0.329	0.365	128	2.7s
✓	–	✓ (report+policy)	**0.253**	**0.336**	**0.380**	128	2.7s

Table 6: Comparison of uniform resolution scaling and LePaX. Selective high-resolution allocation achieves better report quality under a fixed visual token budget with significantly lower computational cost.

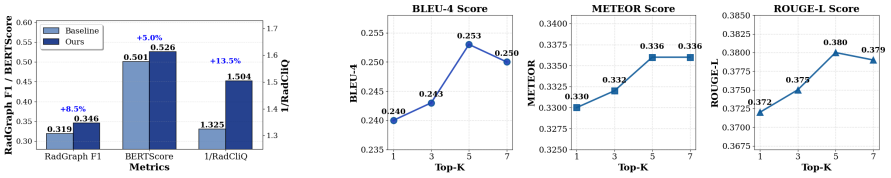
Setting	Report Quality Metrics					Efficiency		
	B-4	ROUGE	RG_{F1}	BERT	1/CliQ	Vision Tokens	Time/Case	GPU Mem
Uniform-384	0.237	0.350	0.319	0.501	1.325	128	2.3s	36GB
Uniform-1024	0.240	0.352	0.320	0.504	1.329	1280	10.0s	65GB
Ours-1024	0.245	0.364	0.331	0.513	1.389	128	2.4s	40GB
Ours-1920	**0.253**	**0.380**	**0.346**	**0.526**	**1.504**	**128**	**2.7s**	**46GB**

high-resolution capacity to informative regions under a fixed token budget. Even with native-resolution images (up to 1920×1920), it achieves better performance with only a small increase in inference time, demonstrating greater efficiency than uniform resolution scaling.

Ablation on Top-K Region Selection. We evaluate different Top-K values (1, 3, 5, 7) for high-resolution patch extraction (Fig. 3(b)). Increasing K improves performance by incorporating more disease-relevant regions, while overly large K introduces redundancy and slightly degrades report quality. **Top-K=5** achieves the best results (B-4: 0.253, METEOR: 0.336, ROUGE-L: 0.380).

4.5 Qualitative Analysis

Fig. 4 compares the baseline and **LePaX** on MIMIC-CXR cases. Grad-CAM heatmaps show that diagnostically relevant regions occupy only a small portion of the image, motivating selective high-resolution patch extraction. While the baseline detects findings such as *pleural effusion* and *tube placement*, its descriptions are often coarse or incomplete. In contrast, **LePaX** integrates lesion-focused details with global context to produce more precise and clinically grounded reports. For example, in **Case I**, the model localizes the *endotracheal tube* relative to the carina and identifies the *orogastric tube below the diaphragm*, consistent with the reference report. Fig.5a further visualizes selected low-resolution (LR) regions, their corresponding high-resolution (HR) patches, and LR→HR fusion attention maps. Red boxes denote LR regions selected for high-resolution



(a) Comparison between the baseline and our LePaX framework across three clinic-related metrics

(b) Ablation study of the Top-K region selection on report generation performance

Fig. 3: Ablation studies on MIMIC-CXR.

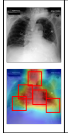
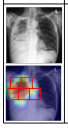
CASE	Reference Report	Baseline	Ours
	endotracheal tube terminates **approximately 5.6 cm** above the carina and is adequately placed. right internal jugular line ends at mid svc. a **feeding tube is seen to course below the diaphragm into the stomach;** however, the distal end is beyond the radiographic view. **left** lower lung opacities reflecting combination of atelectasis and mild pleural effusion is unchanged since.mildly enlarged heart and mediastinal contours are stable.	indwelling support and monitoring devices are in standard position, and cardiomediastinal contours are stable in appearance. persistent left retrocardiac atelectasis andor consolidation with adjacent small left pleural effusion. improving atelectasis at ~~right lung~~ base with residual linear atelectasis remaining. no visible pneumothorax on this semi upright radiograph.	the cardiomediastinal contours are within normal limits, and the right lung is clear without pneumothorax. mild-to-moderate left lower lobe atelectasis with a small **left** pleural effusion remains unchanged from prior. the endotracheal tube **terminates approximately 5 cm** above the carina and is appropriately positioned. a right internal jugular line projects to the lower SVC. **the orogastric tube extends below the diaphragm into the stomach**, with the distal tip not visualized.
	as compared to the prior radiograph, there is no significant interval change. there is a large right-sided pleural collection with an associated air–fluid level, consistent with hydropneumothorax. there is diffuse right pleural thickening with mild volume loss of the right hemithorax, along with fibrotic changes in the right lung parenchyma. the left lung is clear without pleural abnormality. the left heart border appears unremarkable.	compared to the previous study, there is no relevant change. there is a large right-sided pleural effusion with adjacent atelectasis at the right lung base. the left lung remains unchanged. cardiac size is stable. no pneumothorax is identified.	Compared with the prior examination, there is no significant interval change. There is a large right pleural fluid collection with a visible air–fluid level, compatible with hydropneumothorax. There is extensive right pleural thickening with slight reduction in right hemithoracic volume, along with fibrotic changes of the right lung parenchyma. The left lung and pleura are unremarkable. The left cardiac border remains normal.

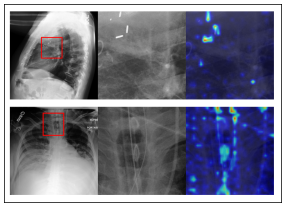
Fig. 4: Qualitative comparison between the baseline and our method on two MIMIC-CXR cases. The main clinical findings are highlighted in different colors for clarity.

processing. In **Case I**, attention highlights subtle post-surgical structures such as surgical clips, while in **Case II** it follows the course of the right internal jugular line in the upper mediastinum. These examples show how GRF refines coarse spatial cues and enhances diagnostically relevant regions, enabling spatially grounded reasoning beyond classifier-driven activation.

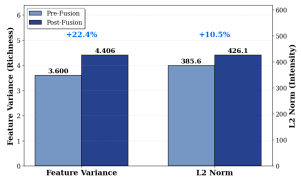
Additionally, Fig. 6 compares LSRA and Grad-CAM region allocation. Red boxes denote LSRA predictions and green boxes indicate Grad-CAM guidance. Across the examples, LSRA consistently captures clinically relevant regions highlighted by Grad-CAM. Cases 2–4 further demonstrate the effect of report-level supervision, where the learned policy better localizes medical devices (Cases 2–3) and pleural effusion (Case 4). These results show that LSRA preserves diagnostically important regions while adapting toward report-oriented spatial relevance.

5 Conclusions

We propose LePaX, a high-resolution RRG framework with two complementary components: LSRA for lesion-focused patch extraction and GRF for hierarchical fusion of regional and global cues. This shifts RRG from coarse, pixel-limited encoding toward clinically guided high-resolution understanding. By combining lesion-focused precision with context-aware interpretation, LePaX achieves consistent gains over prior methods on both lexical and clinical metrics across MIMIC-CXR, IU-Xray, and CheXpertPlus.



(a) Adaptive region selection and LR→HR fusion



(b) Feature statistics before and after GRF

Fig. 5: Analysis of the proposed Global–Regional Fusion (GRF). (a) Visualization of selected regions and LR→HR fusion attention. (b) Feature statistics before and after fusion, showing increased variance and L2 norm after GRF.

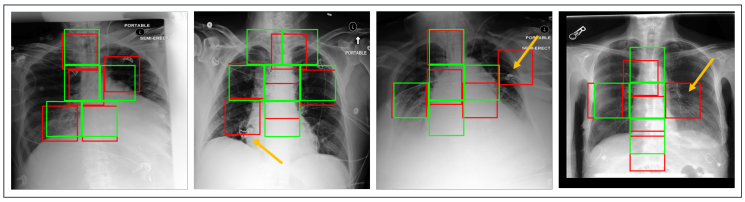


Fig. 6: Qualitative comparison of region allocation. Red boxes denote LSRA predictions and green boxes indicate Grad-CAM guidance. LSRA consistently identifies clinically relevant regions similar to Grad-CAM. Cases 2–4 further show that report-level supervision helps the learned policy better localize medical devices (Cases 2–3) and pleural effusion (Case 4).

References

1. Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A.A., Bach, N., Bahree, A., Bakhtiari, A., Bao, J., Behl, H., et al.: Phi-3 technical report: A highly capable language model locally on your phone. arXiv preprint arXiv:2404.14219 (2024)
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022)
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
4. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
5. Brennan, P.C., Gandomkar, Z., Ekpo, E.U., Tapia, K., Trieu, P.D., Lewis, S.J., Wolfe, J.M., Evans, K.K.: Radiologists can detect the 'gist'of breast cancer before any overt signs of cancer appear. Scientific reports **8**(1), 8717 (2018)
6. Bu, S., Li, T., Yang, Y., Dai, Z.: Instance-level expert knowledge and aggregate discriminative attention for radiology report generation. In: Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14194–14204 (2024)
7. Chambon, P., Delbrouck, J.B., Sounack, T., Huang, S.C., Chen, Z., Varma, M., Truong, S.Q., Chuong, C.T., Langlotz, C.P.: Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats. arXiv preprint arXiv:2405.19538 (2024)
8. Chen, Z., Li, Y., Wang, Z., Gao, P., Barthelemy, J., Zhou, L., Wang, L.: Enhancing radiology report generation via multi-phased supervision. IEEE Transactions on Medical Imaging (2025)
9. Chen, Z., Wang, W., Tian, H., Ye, S., Gao, Z., Cui, E., Tong, W., Hu, K., Luo, J., Ma, Z., et al.: How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. Science China Information Sciences **67**(12), 220101 (2024)
10. Chen, Z., Shen, Y., Song, Y., Wan, X.: Cross-modal memory networks for radiology report generation. arXiv preprint arXiv:2204.13258 (2022)
11. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. arXiv preprint arXiv:2010.16056 (2020)
12. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association **23**(2), 304–310 (2016)
13. Gandomkar, Z., Siviengphanom, S., Ekpo, E.U., Suleiman, M., Taba, S.T., Li, T., Xu, D., Evans, K.K., Lewis, S.J., Wolfe, J.M., et al.: Global processing provides malignancy evidence complementary to the information captured by humans or machines following detailed mammogram inspection. Scientific reports **11**(1), 20122 (2021)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
15. Hou, W., Xu, K., Cheng, Y., Li, W., Liu, J.: Organ: observation-guided radiology report generation via tree reasoning. arXiv preprint arXiv:2306.06466 (2023)
16. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
17. Huang, X., Chen, W., Liu, J., Lu, Q., Luo, X., Shen, L.: Damper: A dual-stage medical report generation framework with coarse-grained mesh alignment and fine-grained hypergraph matching. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 3769–3778 (2025)
18. Huang, Z., Zhang, X., Zhang, S.: Kiut: Knowledge-injected u-transformer for radiology report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19809–19818 (2023)
19. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 590–597 (2019)
20. Jain, S., Agrawal, A., Saporta, A., Truong, S.Q., Duong, D.N., Bui, T., Chambon, P., Zhang, Y., Lungren, M.P., Ng, A.Y., et al.: Radgraph: Extracting clinical entities and relations from radiology reports. arXiv preprint arXiv:2106.14463 (2021)
21. Jin, H., Che, H., Lin, Y., Chen, H.: Promptmrg: Diagnosis-driven prompts for medical report generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 2607–2615 (2024)

22. Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S.: Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019)
23. Kundel, H.L., Nodine, C.F., Conant, E.F., Weinstein, S.P.: Holistic component of image perception in mammogram interpretation: gaze-tracking study. Radiology **242**(2), 396–402 (2007)
24. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. Transactions on Machine Learning Research (2024)
25. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36**, 28541–28564 (2023)
26. Li, M., Lin, B., Chen, Z., Lin, H., Liang, X., Chang, X.: Dynamic graph enhanced contrastive learning for chest x-ray report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3334–3343 (2023)
27. Li, Y., Liu, Y., Wang, Z., Liang, X., Liu, L., Wang, L., Cui, L., Tu, Z., Wang, L., Zhou, L.: A comprehensive study of gpt-4v's multimodal capabilities in medical imaging. medRxiv pp. 2023–11 (2023)
28. Li, Y., Wang, Z., Liu, Y., Wang, L., Liu, L., Zhou, L.: Kargen: Knowledge-enhanced automated radiology report generation using large language models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 382–392. Springer (2024)
29. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
30. Liu, A., Guo, Y., Yong, J.h., Xu, F.: Multi-grained radiology report generation with sentence-level image-language contrastive learning. IEEE Transactions on Medical Imaging (2024)
31. Liu, C., Tian, Y., Chen, W., Song, Y., Zhang, Y.: Bootstrapping large language models for radiology report generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 18635–18643 (2024)
32. Liu, Y., Li, Y., Chen, T., Liu, L., Wang, L., Zhou, L.: Sat-rrg: Llm-guided self-adaptive training for radiology report generation with token-level push-pull optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 35363–35372 (2026)
33. Nicolson, A., Dowling, J., Koopman, B.: Improving chest x-ray report generation by leveraging warm starting. Artificial intelligence in medicine **144**, 102633 (2023)
34. Ostmeier, S., Xu, J., Chen, Z., Varma, M., Blankemeier, L., Bluethgen, C., Md, A., Moseley, M., Langlotz, C., Chaudhari, A., et al.: Green: Generative radiology report evaluation and error notation. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 374–390 (2024)
35. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
36. Park, S.J., Heo, K.S., Shin, D.H., Son, Y.H., Oh, J.H., Kam, T.E.: Dart: Disease-aware image-text alignment and self-correcting re-alignment for trustworthy radiology report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15580–15589 (June 2025)

37. Pellegrini, C., Özsoy, E., Busam, B., Navab, N., Keicher, M.: Radialog: A large vision-language model for radiology report generation and conversational assistance. arXiv preprint arXiv:2311.18681 (2023)
38. Seah, J.C., Tang, C.H., Buchlak, Q.D., Holt, X.G., Wardman, J.B., Aimoldin, A., Esmaili, N., Ahmad, H., Pham, H., Lambert, J.F., et al.: Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study. The Lancet Digital Health **3**(8), e496–e506 (2021)
39. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
40. Sha, Y., Pan, H., Meng, W., Li, K.: Contrastive knowledge-guided large language models for medical report generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 111–120. Springer (2025)
41. Vaswani, A.: Attention is all you need. Advances in Neural Information Processing Systems (2017)
42. Wang, J., Bhalerao, A., Yin, T., See, S., He, Y.: Camanet: class activation map guided attention network for radiology report generation. IEEE Journal of Biomedical and Health Informatics **28**(4), 2199–2210 (2024)
43. Wang, X., Li, Y., Wang, F., Wang, S., Li, C., Jiang, B.: R2gencsr: Retrieving context samples for large language model based x-ray medical report generation. arXiv preprint arXiv:2408.09743 (2024)
44. Wang, X., Wang, F., Li, Y., Ma, Q., Wang, S., Jiang, B., Tang, J.: Cxpmrg-bench: Pre-training and benchmarking for x-ray medical report generation on chexpert plus dataset. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5123–5133 (2025)
45. Wang, Z., Liu, L., Wang, L., Zhou, L.: Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11558–11567 (2023)
46. Wang, Z., Liu, L., Wang, L., Zhou, L.: R2gengpt: Radiology report generation with frozen llms. Meta-Radiology **1**(3), 100033 (2023)
47. Wu, Y., Huang, I.C., Huang, X.: Token imbalance adaptation for radiology report generation. In: Conference on Health, Inference, and Learning. pp. 72–85. PMLR (2023)
48. Xue, L., Shu, M., Awadalla, A., Wang, J., Yan, A., Purushwalkam, S., Zhou, H., Prabhu, V., Dai, Y., Ryoo, M.S., et al.: xgen-mm (blip-3): A family of open large multimodal models. arXiv preprint arXiv:2408.08872 (2024)
49. Yang, S., Wu, X., Ge, S., Zheng, Z., Zhou, S.K., Xiao, L.: Radiology report generation with a learned knowledge base and multi-modal alignment. Medical Image Analysis **86**, 102798 (2023)
50. Yang, Y., Yu, J., Fu, Z., Zhang, K., Yu, T., Wang, X., Jiang, H., Lv, J., Huang, Q., Han, W.: Token-mixer: Bind image and text in one embedding space for medical image reporting. IEEE Transactions on Medical Imaging **43**(11), 4017–4028 (2024)
51. Ye, Q., Xu, H., Ye, J., Yan, M., Hu, A., Liu, H., Qian, Q., Zhang, J., Huang, F.: mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13040–13051 (2024)

52. Yu, F., Endo, M., Krishnan, R., Pan, I., Tsai, A., Reis, E.P., Fonseca, E.K.U.N., Lee, H.M.H., Abad, Z.S.H., Ng, A.Y., et al.: Evaluating progress in automatic chest x-ray radiology report generation. Patterns **4**(9) (2023)
53. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11975–11986 (2023)
54. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675 (2019)
55. Zhao, W., Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Ratescore: A metric for radiology report generation. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 15004–15019 (2024)
56. Zhou, Q., Liang, G., Li, X., Chen, J., Wang, Z., Yao, C., Wu, S.: Learnable retrieval enhanced visual-text alignment and fusion for radiology report generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22529–22538 (2025)
57. Zhou, S., Li, Y., Liu, Y., Liu, L., Wang, L., Zhou, L.: A review of longitudinal radiology report generation: Dataset composition, methods, and performance evaluation. arXiv preprint arXiv:2510.12444 (2025)
58. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)
59. Zhu, Q., Mathai, T.S., Mukherjee, P., Peng, Y., Summers, R.M., Lu, Z.: Utilizing longitudinal chest x-rays and reports to pre-fill radiology reports. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 189–198. Springer (2023)