RefleXNet: Targeted Self-Reflection for Accurate Chest X-ray Reporting

Xin Mei[1,2], Rui Mao[2], Xiaoyan Cai[1*], Libin Yang[1], Erik Cambria[2]

[1]Northwestern Polytechnical University, China
[2]Nanyang Technological University, Singapore
meixin@mail.nwpu.edu.cn, rui.mao@ntu.edu.sg, xiaoyanc@nwpu.edu.cn, libiny@nwpu.edu.cn, cambria@ntu.edu.sg

Abstract

Automated interpretation and reporting of chest X-rays (CXRs) hold significant promise in reducing diagnostic errors and supporting radiologists under heavy clinical workloads. However, existing methods typically rely on global visual features and token-level supervision, limiting their sensitivity to subtle abnormalities and reducing their clinical reliability. To address these challenges, we present Reflective X-ray Network (RefleXNet), which systematically integrates multi-scale visual feature fusion and anatomical relational reasoning with a targeted self-reflective learning strategy. RefleXNet first constructs multi-scale visual representations and captures anatomical context through graph-based relational modeling. Building upon these representations, we introduce a targeted self-reflection strategy that uses clinically guided feedback from generated reports to selectively refine abnormality predictions and their associated region-level visual features. Extensive experiments on MIMIC-CXR demonstrate that RefleXNet consistently outperforms state-of-the-art baselines across clinical factual correctness metrics. Notably, our compact 3B-parameter model surpasses several recent models with over twice the parameter count. Additionally, RefleXNet exhibits strong generalization performance in zero-shot evaluations on IU-Xray compared with leading multimodal language models, highlighting its robustness and clinical effectiveness.

Introduction

Accurate interpretation and reporting of chest X-ray (CXR) images are essential for diagnosing and managing thoracic diseases. Despite the widespread use of CXRs due to their cost-effectiveness and accessibility, manual interpretation remains labor-intensive and heavily dependent on radiologist expertise. Under high clinical workloads, subtle abnormalities are frequently overlooked, potentially leading to diagnostic errors and inter-observer variability. Automated CXR report generation has emerged as a promising approach to address these limitations. Early studies primarily employed convolutional or transformer-based visual encoders combined with sequential language models (Chen et al. 2020; Perera et al. 2025).

Recent methods further incorporated anatomical priors, structured knowledge graphs (Mei et al. 2024), or region-guided modules (Tanida et al. 2023) to enhance accuracy and interpretability. The development of multimodal large language models (MLLMs) (Lin et al. 2025), such as Med-PaLM (Tu et al. 2023) and MAIRA (Hyland et al. 2023), has notably advanced automated CXR analysis. Despite this progress, most existing models rely on token-level supervision, which may not always correspond to clinical accuracy due to diverse reporting styles and a focus on surface-level text features. In addition, many approaches depend on global or patch-level features and often lack explicit anatomical structure modeling, making them less sensitive to subtle, localized findings that are clinically important.

To overcome these limitations, we propose the **Reflective X-ray Network (RefleXNet)**, which systematically integrates multi-scale visual modeling, anatomical relational reasoning, and a targeted self-reflection mechanism for CXR diagnosis and report generation. Our method first constructs rich visual representations by combining region-level and patch-level features, capturing both explicit anatomical structure and fine-grained image context. These representations are further enhanced through graph-based relational reasoning, which models spatial and semantic dependencies among anatomical regions, thereby providing a more robust foundation for abnormality localization and detection. Building on these enriched features, RefleXNet introduces a targeted self-reflection learning strategy that leverages feedback from the content of generated reports for clinically guided, error-driven optimization. Instead of uniformly updating all abnormality predictions, our approach identifies predictions inconsistent with the generated report's clinical findings and selectively optimizes only those. Guided by the region-abnormality mapping, these targeted corrections are propagated to the relevant anatomical regions, enabling self-reflection optimization to focus feedback and refine the most clinically meaningful features. This targeted optimization process stands in contrast to conventional joint training methods, which treat all abnormality categories equally and are more susceptible to data imbalance or overfitting to prevalent findings. By focusing on clinically meaningful corrections, RefleXNet promotes more balanced, reliable learning. Keeping parameters frozen during self-reflection also cuts computation and enables efficient adaptation.

*Corresponding author.

Experimental results demonstrate that RefleXNet achieves superior performance compared to recent state-of-the-art methods across multiple metrics on the MIMIC-CXR dataset. Notably, our compact 3B-parameter model outperforms several larger (7B-parameter) models, underscoring the efficiency and effectiveness of our approach. Additionally, zero-shot evaluation on the IU-Xray dataset demonstrates robust generalization compared to leading multimodal language models.

Our contributions can be summarized as follows: (1) We introduce RefleXNet, a unified framework combining multi-scale visual representations with graph-based relational reasoning for CXR diagnosis. (2) We propose a targeted self-reflective learning strategy that leverages clinically guided, report-level feedback to iteratively refine abnormality predictions and related region representations, significantly enhancing clinical accuracy. (3) Extensive experiments demonstrate that RefleXNet delivers state-of-the-art performance and strong generalizability with significantly fewer parameters, highlighting its clinical effectiveness and computational efficiency.

Related Work

Recent advancements in deep learning and multimodal models have enhanced performance in complex tasks (Wu et al. 2023; Zhang et al. 2025), with notable progress in medical applications (Wu et al. 2026). In particular, Chest X-ray (CXR) interpretation plays a crucial role in assisting clinicians in detecting thoracic abnormalities (McBee et al. 2018). Early CXR analysis, similar to other computer vision tasks (Zhao et al. 2025), focused on relatively simple objectives, such as multi-label disease classification (Rahmat, Ismail, and Aliman 2018). More recent approaches have shifted towards advanced encoder-decoder frameworks for directly generating reports (Sloan et al. 2024), integrating domain-specific knowledge to enhance accuracy. Knowledge graphs have been used to capture relationships between clinical concepts in reports (Mei et al. 2024), while heterogeneous graphs link anatomical structures with findings (Zhang et al. 2020). Region-aware methods, such as RGRG (Tanida et al. 2023), extract anatomical features to better align reports with clinical practices, though they often focus on binary classifications and provide limited abnormality characterization.

The development of large language models (LLMs) and multimodal LLMs (MLLMs) has also led to significant progress in CXR analysis. Early work using LLMs, such as PromptMRG (Jin et al. 2024), provided auxiliary textual supervision or regional descriptions to guide visual grounding. Subsequent approaches, such as Med-PaLM (Tu et al. 2023), fine-tuned models like PaLM-E (Driess et al. 2023) for radiology tasks, achieving broad clinical utility but with high computational costs. To address these challenges, moderate-scale open-source MLLMs, such as LLaVA-Rad (Zambrano Chaves et al. 2025), CheXagent (Chen et al. 2024), and MAIRA (Hyland et al. 2023), have been introduced, employing tailored instruction-following datasets and specialized biomedical vision-language alignment.

Despite these advancements, many existing methods rely on token-level or global supervision, limiting the targeted refinement of specific abnormalities. Our approach addresses this limitation by leveraging structured report feedback for more targeted, region-specific optimization.

Methodology

Our approach, outlined in Figure 1, consists of two major components. The Multi-Scale Graph-Augmented Visual Embedding (MGVE) module extracts comprehensive anatomical features by fusing region-level and patch-level visual cues with region-aware cross-attention, and further incorporates spatial and semantic context through relational graph modeling. Building on these representations, report generation is performed via a targeted self-reflection training strategy, which operates in three stages: (1) pretraining MGVE with an abnormality-aware multi-label classification task, (2) supervised fine-tuning of the combined MGVE and language model for radiology report generation, and (3) targeted refinement of abnormality predictions and region features through feedback from generated reports. This pipeline directly couples prediction refinement with report-level clinical accuracy, supporting both improved abnormality detection and more reliable report generation.

Multi-Scale Graph-Augmented Visual Embedding for Abnormality Detection

Multi-Scale Visual Feature Extraction and Region-Patch Fusion Interpreting chest X-ray (CXR) images is challenging due to complex anatomy and subtle abnormalities. To address this, we adopt a multi-scale visual encoding strategy that combines anatomical region information with detailed image context. Specifically, we extract region-level features that capture anatomical structures and patch-level features that represent local details and broader context. Region-level features are obtained using a Faster R-CNN model pretrained by Hu et al. (Hu et al. 2023), resulting in N_r region representations $\mathbf{V}^r \in R^{N_r \times d_r}$. In parallel, patch-level features are produced by Rad-DINO, a self-supervised vision transformer for medical imaging, which divides each image into $N_p = 1,369$ non-overlapping patches of 37×37 pixels to obtain representations $\mathbf{V}^p \in R^{N_p \times d_p}$.

However, providing all patch features directly to downstream models can lead to prohibitively long input sequences and a loss of anatomical coherence. To overcome this limitation, we introduce a region-aware cross-attention fusion mechanism that enables each anatomical region to selectively aggregate informative cues from all image patches. Formally, region and patch features are projected into a shared latent space using learnable linear transformations:

$$\mathbf{Q} = \mathbf{V}^r \mathbf{W}_Q, \quad \mathbf{K} = \mathbf{V}^p \mathbf{W}_K, \quad \mathbf{V} = \mathbf{V}^p \mathbf{W}_V, \qquad (1)$$

where $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$ are learnable parameters. Cross-attention is then computed by

$$\mathbf{A} = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_a/h}}\right), \qquad (2)$$

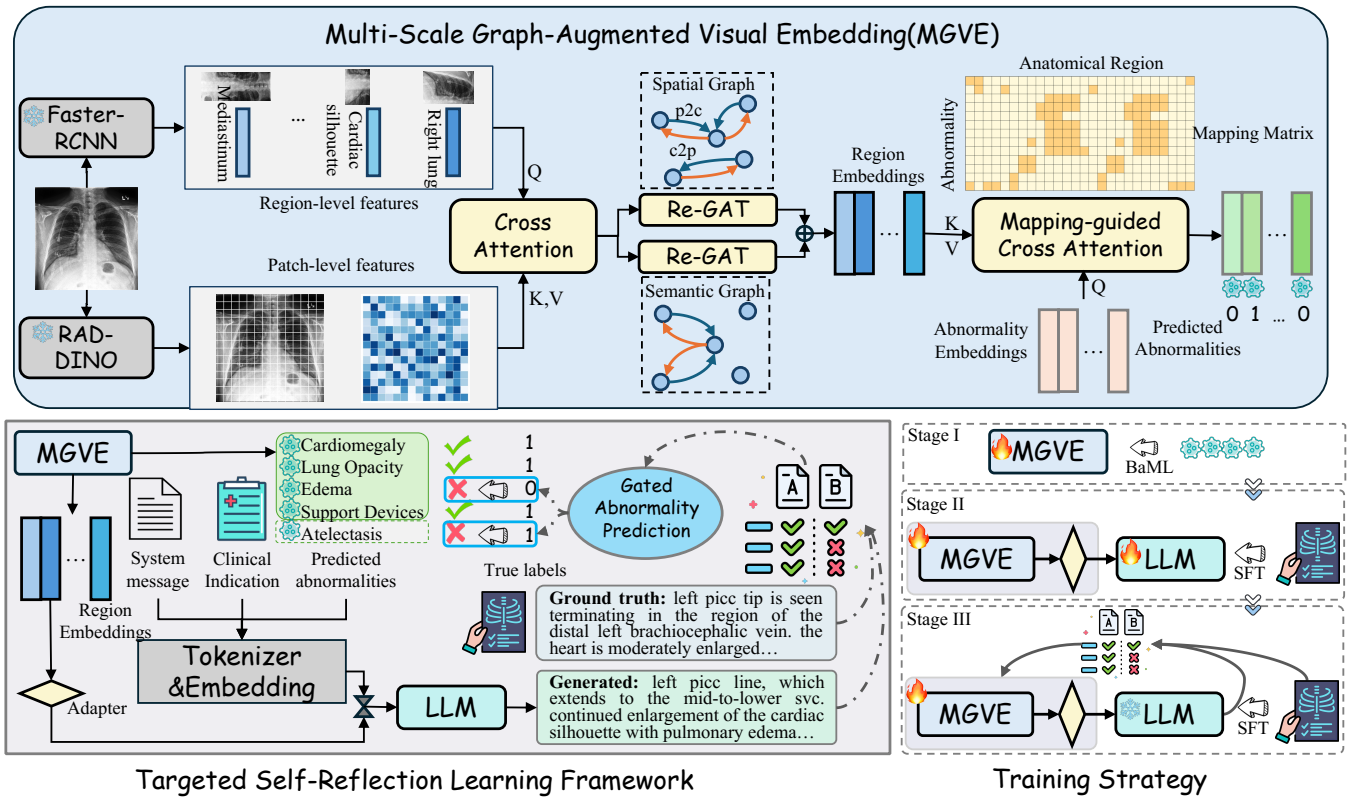


Figure 1: Overview of our propsed RefleXNet.The top panel illustrates the architecture of the MGVE module, which integrates region- and patch-level features using graph-based relational reasoning. The lower left shows the overall targeted self-reflection learning framework, while the lower right details the three-stage training strategy: (I) MGVE pretraining, (II) supervised fine-tuning of MGVE and the language model, and (III) self-reflection optimization.

where h is the number of attention heads. Each region aggregates contextual information from the patches, and the outputs from all heads are concatenated and projected by a linear layer to obtain the final fused region embeddings:

$$\widetilde{\mathbf{V}} = \text{Linear}\left(\text{Concat}_{\text{heads}}(\mathbf{AV})\right). \quad (3)$$

Graph-Augmented Feature Refinement To further enhance the anatomical context captured by the region embeddings obtained in the previous stage, we construct two complementary relation graphs: a *spatial relation graph* and a *semantic relation graph*. These graphs encode different yet mutually reinforcing types of anatomical relationships, enabling refined feature representation through structured relational reasoning. The spatial relation graph explicitly models geometric and positional relationships among the anatomical regions. Following established approaches (Yao et al. 2018; Hu et al. 2023), we define spatial relationships based on bounding-box geometry and positional configurations. Specifically, edges between region pairs are assigned according to spatial criteria, including containment (*inside*, *cover*), significant intersection (*overlap*, determined by intersection-over-union), and directional relations categorized into eight angle-based classes (such as superior, inferior, lateral) when the centroid distance between two regions falls below a predefined threshold relative to image size. Thus, this graph encodes explicit geometric context essential for distinguishing abnormalities in relation to anatomical layout. In addition, the semantic relationship graph captures the hierarchical and functional anatomical relationships defined by the Chest ImaGenome dataset (Wu et al. 2021). Unlike the spatial graph, this semantic graph emphasizes parent-child anatomical structures, such as "lung contains upper lobe" or "upper lobe inside lung." Edges are established between anatomical regions based on these hierarchical semantic relations, thereby providing meaningful anatomical context beyond purely spatial proximity.

Each node in both graphs represents one of the 26 anatomical regions (for example, "aortic arch," "cardiac silhouette," or "left lung"), and edges capture their corresponding spatial or semantic relationships. Subsequently, we employ a relation-aware graph attention network (ReGAT) (Li et al. 2019) independently on both graphs to update the fused region embeddings $\widetilde{\mathbf{V}} \in R^{N_r \times d_r}$. This yields two refined sets of embeddings: spatial-aware region embeddings $\widetilde{\mathbf{V}}_{\text{spa}}$ and semantic-aware region embeddings $\widetilde{\mathbf{V}}_{\text{sem}}$. Finally, we integrate these two complementary embeddings by computing their element-wise average, resulting in the final refined region representations $\mathbf{V}' = (\widetilde{\mathbf{V}}_{\text{sem}} + \widetilde{\mathbf{V}}_{\text{spa}})/2$.

Abnormality-Specific Feature Aggregation and Prediction Our model targets 14 abnormality observations defined by the CheXpert benchmark (Smit et al. 2020), encompassing a broad spectrum of clinically important thoracic conditions. For each abnormality, we associate a set of anatomically relevant regions determined by clinical expertise. This association is encoded in a binary mapping matrix $\mathbf{M} \in 0, 1^{N_a \times N_r}$, where $N_a = 14$ denotes the number of abnormalities and $N_r = 26$ the number of anatomical regions. In contrast to conventional multi-label attention mechanisms that learn free-form associations, our approach leverages an explicit region-abnormality mapping to encode clinical priors, allowing each abnormality prediction to focus only on anatomically relevant regions and thereby promoting more targeted learning. Further details and the complete mapping matrix are provided in the supplementary materials.

Given the refined region embeddings $\mathbf{V}' \in R^{N_r \times d_r}$ for each image and a set of learnable abnormality-specific query vectors $\mathbf{Q} \in R^{N_a \times d_r}$, we employ a parameterized cross-attention mechanism to model their interactions. Specifically, both $\mathbf{Q}$ and $\mathbf{V}'$ are first projected into a shared latent space via learnable linear transformations. The attention score between the i-th abnormality query and the j-th anatomical region is computed as:

$$\mathbf{A}_{i,j} = \begin{cases} \frac{\phi(\mathbf{Q}_i) \cdot \psi(\mathbf{V}'_j)}{\sqrt{d_r}}, & \text{if } \mathbf{M}_{i,j} = 1 \\ -\infty, & \text{otherwise} \end{cases}, \tag{4}$$

where $\phi(\cdot)$ and $\psi(\cdot)$ are learnable linear projections. After softmax normalization over regions, the attention weights aggregate contextualized region features for each abnormality:

$$\mathbf{F}i = \sum_{j=1}^{N_r} \text{softmax}(\mathbf{A}_{i,:})_j \cdot \gamma(\mathbf{V}'_j), \tag{5}$$

where $\gamma(\cdot)$ is an additional value projection. The resulting abnormality-aware features are subsequently passed to a classification head for multi-label abnormality prediction.

Report Generation Reinforced with Self-reflection

Stage I: MGVE Pre-training via Abnormality-aware Multi-label Classification In the first stage, we pre-train MGVE using a multi-label abnormality classification task, where each of the 14 CheXpert abnormality observations is treated as a separate binary classification problem. A key challenge is the severe class imbalance, which can lead the model to focus on frequent findings and overlook rare but clinically significant abnormalities. To mitigate this, we employ a balanced multi-label classification loss (BaML Loss) that adaptively adjusts the weight of each prediction based on both its difficulty and the class imbalance level. BaML Loss for a batch is defined as:

$$\mathcal{L}_{BaML} = \sum_{i=1}^{N_\text{a}} \sum_{j=1}^{B} \tilde{w}_{ij}, \ell_{ij}, \tag{6}$$

where ℓ_{ij} is the binary cross-entropy loss for abnormality class i and sample j:

$$\ell_{ij} = -y_{ij} \ln \sigma(x_{ij}) - (1 - y_{ij}) \ln(1 - \sigma(x_{ij})), \tag{7}$$

with $\sigma(x_{ij})$ denoting the sigmoid activation of the predicted logit x_{ij}, and $y_{ij} \in \{0, 1\}$ the ground-truth label indicating the presence or absence of abnormality i in sample j.

The weight w_{ij} for each prediction incorporates both difficulty and imbalance information. Specifically, the difficulty-aware term is defined as $h_{ij} = \ln(1 + \ell_{ij})$, which assigns higher weights to more challenging predictions. The imbalance-aware term is given by

$$l_{ij} = \begin{cases} \ln(1 + r_i), & \text{if } y_{ij} = 1, \\ 1, & \text{if } y_{ij} = 0, \end{cases} \tag{8}$$

where $r_i = \frac{B - \sum_{j=1}^{B} y_{ij}}{\sum_{j=1}^{B} y_{ij}}$ is the imbalance ratio for abnormality i in the current batch, and l_{ij} increases the weight for positive samples of rare abnormalities. The final per-sample weight is $w_{ij} = h_{ij} \cdot l_{ij}$, and the normalized weight is

$$\tilde{w}ij = \frac{w_{ij}}{\sum_{i=1}^{N_\text{abn}} \sum_{j=1}^{B} w_{ij}}. \tag{9}$$

Stage II: Training with Supervised Fine-tuning In the second stage, we build a vision-language model for chest X-ray report generation by combining the pre-trained MGVE as the visual encoder with Qwen-2.5 as the language decoder. The region embeddings $\mathbf{V}' \in R^{N_r \times d_r}$ produced by MGVE are projected into the language model's semantic space through a lightweight adapter, following the approach used in most MLLMs such as Qwen-VL. The adapter consists of a linear transformation followed by layer normalization:

$$\mathbf{H} = \text{LayerNorm}(\mathbf{V}'\mathbf{W}_\text{adp} + \mathbf{b}_\text{adp}), \tag{10}$$

where $\mathbf{W}_\text{adp}$ and $\mathbf{b}_\text{adp}$ are the adapter's learnable parameters, and $\mathbf{H}$ denotes the adapted visual features.

During training, the language model is provided with the clinical indication, the adapted region embeddings, and the ground truth CheXpert abnormality observations as input prompts. Supervised fine-tuning (SFT) is performed with a token-level generation loss, which encourages the model to produce reports that closely match expert references. The objective for sample j is defined as

$$\mathcal{L}_\text{gen}^{(j)} = -\sum_{t=1}^{T_j} \log P(y_t^{(j)} \mid y_{<t}^{(j)}, \mathbf{X}^{(j)}), \tag{11}$$

where T_j is the length of the reference report for sample j, $y_t^{(j)}$ is the reference token at position t, and $\mathbf{X}^{(j)}$ denotes the combined input features.

Stage III: Continual Training with Self-Reflection After supervised fine-tuning, our vision-language model generates radiology reports conditioned on visual features, clinical indications, and ground-truth abnormality observations. As predicted abnormalities directly influence report quality, we further propose a targeted self-reflection optimization strategy that leverages feedback from the generated reports to iteratively improve abnormality predictions. Specifically, we employ CheXbert to extract abnormality labels from both the generated and ground-truth reports, producing binary label vectors $\mathbf{y}^\text{pred}$ and $\mathbf{y}^\text{gt} \in 0, 1^{N_\text{a}}$.

Method	Size	BLUE-4	RL	F1-Rad	RadCliQ$_0$↓	Micro-F1		Macro-F1	
						14	5	14	5
R2Gen(Chen et al. 2020)	<1B	0.103	0.277	0.196	–	0.228	0.346	0.276	-
KiUT(Huang, Zhang, and Zhang 2023)	<1B	0.113	0.285	–	–	–	–	0.321	-
DCL (Li et al. 2023b)	<1B	0.109	0.284	–	–	–	–	0.373	-
RGRG (Tanida et al. 2023)	<1B	0.126	0.264	–	–	0.447	0.547	–	–
ORGAN (Hou et al. 2023)	<1B	0.123	0.293	–	–	–	–	0.385	-
PromptMRG (Jin et al. 2024)	<1B	0.112	0.268	–	–	0.476	–	–	
Qwen2.5-VL (Bai et al. 2025)	3B	0.012	0.133	0.094	–	0.320	0.313	0.238	0.274
LLaVA (Liu et al. 2023)	7B	0.013	0.132	0.022	–	0.229	0.234	0.154	0.175
LLaVA-Med (Li et al. 2023a)	7B	0.010	0.133	0.065	–	0.272	0.220	0.155	0.166
GPT-4V (OpenAI 2023)	-	0.019	0.132	0.132	–	0.355	0.258	0.204	0.196
CheXagent (Chen et al. 2024)	7B	0.047	0.215	0.205	–	0.393	0.412	0.247	0.345
Med-PaLM S (Tu et al. 2023)	12B	0.104	0.262	0.252	–	0.514	0.565	0.373	0.506
Med-PaLM M (Tu et al. 2023)	84B	0.113	0.273	0.267	–	0.536	0.579	<u>0.398</u>	0.516
Qwen2.5-VL* (Bai et al. 2025)	3B	0.114	0.276	0.263	3.31	0.424	0.453	0.301	0.388
MAIRA (Hyland et al. 2023)	7B	0.142	0.289	<u>0.296</u>	<u>3.10</u>	0.557	0.560	0.386	0.477
LLaVA-Rad (Zambrano Chaves et al. 2025)	7B	0.154	<u>0.306</u>	0.294	3.19	**0.573**	0.574	0.395	0.477
ReflexNet$_{w/o\text{SR}}$	3B	<u>0.158</u>	0.291	0.291	3.15	0.505	<u>0.576</u>	0.393	<u>0.525</u>
ReflexNet	3B	**0.164**	**0.309**	**0.311**	**2.96**	<u>0.570</u>	**0.607**	**0.407**	**0.536**

Table 1: Comparison with baselines on the MIMIC-CXR dataset. The best result for each metric is shown in **bold**, and the second-best is <u>underlined</u>. RL indicates ROUGE-L; F1-Rad denotes F1-RadGraph. Metrics marked with ↓ indicate that lower values are better. An asterisk (*) denotes continued training on MIMIC-CXR. ReflexNet$_{w/o\text{SR}}$ refers to the variant without self-reflection optimization after the second training stage. Qwen2.5-VL results are from our evaluation, GPT-4V from (Zambrano Chaves et al. 2025), and other baselines from original publications. Statistically significant improvements according to paired two-tailed t-tests ($p < 0.01$).

Following standard practice, CheXbert's four-category outputs (positive, negative, uncertain, not mentioned) are binarized, considering only "positive" predictions as positive. By element-wise comparison, an error gate vector $\mathbf{s} \in 0, 1^{N_\text{a}}$ is obtained, where each entry $s_i = 1$ if the predicted abnormality differs from the ground truth, and 0 otherwise.

For each input sample j, we define a gated abnormality prediction loss that selectively focuses optimization only on those abnormality categories identified as incorrect by the generated report:

$$\mathcal{L}_{\text{label}}^{(j)} = \begin{cases} \frac{1}{\sum_{i=1}^{N_\text{a}} s_{ij}} \sum_{i=1}^{N_\text{a}} s_{ij} \cdot \ell_{ij}, & \text{if } \sum_{i=1}^{N_\text{a}} s_{ij} > 0 \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

where ℓ_{ij} denotes the previously defined binary cross-entropy loss for abnormality i in sample j. If all abnormality predictions for sample j are correct, the abnormality prediction loss is zero, and only the standard report generation loss $\mathcal{L}_{\text{gen}}^{(j)}$ is applied. Thus, the total loss for sample j becomes:

$$\mathcal{L}_{\text{total}}^{(j)} = \mathcal{L}_{\text{gen}}^{(j)} + \mathcal{L}_{\text{label}}^{(j)}. \tag{13}$$

Critically, leveraging the previously established region-abnormality mapping, the optimization of abnormality predictions directly provides feedback to refine the corresponding region-level visual representations. By precisely targeting only abnormality categories identified as errors through the generated reports, our self-reflection strategy significantly reduces the risk of overfitting to common or normal findings. Furthermore, this targeted approach maintains report generation as the primary learning objective, thereby promoting clinical accuracy and coherence. To further enhance training efficiency and stability, the language model parameters are frozen during the self-reflection stage, substantially reducing computational overhead.

Experiment

Datasets, Metrics and Implementation

We train and validate our model on the publicly available MIMIC-CXR dataset (Johnson et al. 2019) (220,000+ CXR-report pairs), using standard preprocessing protocols, and evaluate cross-dataset performance on IU-Xray (Demner-Fushman et al. 2016) with its standard split (Chen et al. 2020). Linguistic quality is assessed using BLEU-4 (Papineni et al. 2002) and ROUGE-L (Lin 2004), while clinical accuracy is evaluated with F1-RadGraph (Delbrouck et al. 2022), RadCliQ$_0$ (Yu et al. 2023), and CheXbert-based metrics (Smit et al. 2020) (vector similarity, macro/micro F1 for 5 and 14 observations).

Experiments were conducted using PyTorch on an NVIDIA Tesla A800 GPU. Stage I involved MGVE pre-training with a frozen Faster R-CNN model (Hu et al. 2023) to segment 26 anatomical regions, and Rad-DINO embeddings (Pérez-García et al. 2025) were extracted. Training used a learning rate of 1e-4, batch size 128, and the model with the lowest validation loss was selected. In Stage II, the multimodal large language model (MLLM) was fine-tuned for 3 epochs at 3e-5 learning rate and batch size 8. Stage III froze LLM parameters and continued training for 3 more epochs with the same settings.

Method	RL	Vec	F1-Rad	RadCliQ$_0$↓
Qwen2.5-VL	0.065	0.051	0.045	5.09
Qwen2.5-VL*	0.226	0.553	0.292	2.82
LLaVA-Rad	0.184	0.583	0.294	2.85
ReflexNet$_{w/oSR}$	0.236	0.590	0.312	2.68
ReflexNet	0.244	0.585	0.325	2.64
ReflexNet$^\#$	**0.259**	**0.612**	**0.384**	**2.46**

Table 2: Evaluation results on the IU-Xray dataset. Vec represents CheXbert vector similarity. RefleXNet$^\#$ refers to the variant trained with an additional two epochs of self-reflection optimization on IU-Xray in the stage III.

Main Results

Comparison with Baselines Table 1 presents a comprehensive comparison of RefleXNet and recent baselines on the MIMIC-CXR dataset. RefleXNet achieves the highest scores across all key clinical factual correctness metrics, including an F1-RadGraph of 0.311 and a RadCliQ-v0 score of 2.96 (lower is better). It also establishes new benchmarks on CheXbert-based abnormality classification, with a micro-5 F1 of 0.607 and a macro-5 F1 of 0.536. Compared with LLaVA-Rad, a leading multimodal large language model, RefleXNet consistently improves F1-RadGraph, RadCliQ-v0, and both micro and macro F1 scores, highlighting the effectiveness of our targeted self-reflective learning and multi-scale visual fusion strategy. Importantly, RefleXNet achieves these results with only 3 billion parameters, significantly fewer than models like Med-PaLM and LLaVA-Rad, which use 7 billion or more. While RefleXNet also performs strongly on language generation metrics such as BLEU-4 and ROUGE-L, its main advantage is observed in clinically relevant and factual evaluations. Additionally, results for Qwen2.5-VL before and after continued training on MIMIC-CXR confirm that RefleXNet's improvements stem from the proposed learning strategy rather than backbone changes. These findings demonstrate the clinical accuracy, efficiency, and scalability of RefleXNet.

Model Generalization We further assess the generalization ability of RefleXNet and baseline MLLMs on the IU-Xray dataset in a zero-shot setting, given the limited sample size of IU-Xray. As shown in Table 2, RefleXNet outperforms leading MLLMs such as LLaVA-Rad and Qwen2.5-VL across all clinical factual accuracy metrics, achieving an F1-RadGraph of 0.325 and a RadCliQ-v0 of 2.64, along with competitive scores on ROUGE-L and CheXbert vector similarity. When targeted self-reflection optimization is applied for two additional epochs (with the LLM parameters frozen), RefleXNet$^\#$ shows further improvements, with F1-RadGraph increasing to 0.384 and RadCliQ-v0 decreasing to 2.46. These results indicate that RefleXNet generalizes robustly under strict evaluation settings and benefits from lightweight, focused adaptation, consistently improving performance even when only a small number of additional samples are available.

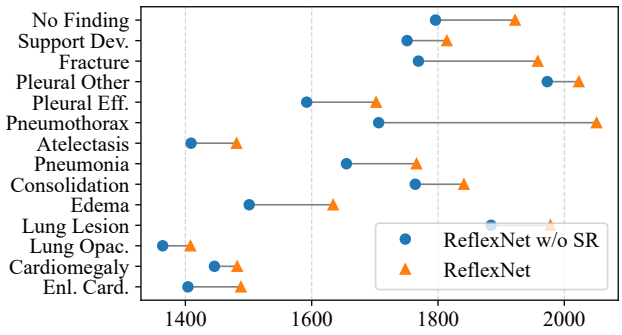


Figure 2: Comparison of the number of correctly predicted cases for 14 abnormalities between RefleXNet with and without self-reflection optimization.

Analysis of Targeted Self-Reflection Optimization We systematically evaluate the effectiveness of targeted self-reflection by comparing RefleXNet models trained with and without this strategy on the MIMIC-CXR test set. As shown in Table 1, introducing self-reflection consistently improves all major evaluation metrics, with the most notable gains observed in clinical correctness. Specifically, F1-RadGraph increases by 2% and Micro-F1 for the 14 abnormality categories improves by 6.5%.

To further examine its impact, we analyze the number of correctly predicted cases for each of the 14 CheXpert abnormality observations, using CheXbert to extract labels from generated reports and comparing them to the ground truth. Figure 2 shows that self-reflection results in more correct predictions across all abnormality observations. The improvement is particularly significant for pneumothorax, where correct predictions increase by 345. Before self-reflection, the model had high recall (0.6027) but very low precision (0.0987) for pneumothorax, resulting in frequent false positives. After applying self-reflection, precision rises to 0.3860 and recall becomes 0.3014, yielding a much more balanced and clinically reliable prediction, and the F1-score increases from 0.1696 to 0.3385. For other findings such as cardiomegaly, improvements are more modest, which may be due to already strong baseline performance (pre-optimization F1-score 0.6433; post-optimization F1-score 0.6772). Overall, these results indicate that targeted self-reflection effectively corrects both over- and under-prediction, resulting in more accurate and clinically meaningful report generation.

Ablation Study

To evaluate the contribution of each major component in our MGVE framework, we conduct a set of ablation experiments on the abnormality prediction task. Table 3 presents the results for three key variants: removing the balanced multi-label (BaML) loss, omitting multi-scale visual feature extraction and region-patch fusion, and disabling the region-abnormality mapping matrix. For completeness, we include the ablation of graph-based relational refinement (ReGAT) and related analysis in the supplementary materials.

Method	Micro			Macro		
	P	R	F1	P	R	F1
w/oBaML	**0.592**	0.386	0.467	**0.376**	0.234	0.264
w/oPatch	0.430	0.534	0.476	0.257	0.324	0.276
w/oMap	0.501	0.556	0.527	0.372	0.396	0.357
Ours	0.463	**0.626**	**0.533**	0.354	**0.461**	**0.388**

Table 3: Ablation study on the abnormality prediction task, showing the performance impact of removing key components of our MGVE model.

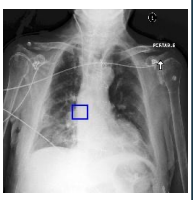
Impact of Balanced Multi-Label (BaML) Loss The first variant ($_{w/o\text{BaML}}$) replaces our adaptive, difficulty- and imbalance-aware loss with a standard binary cross-entropy loss. As BaML is designed to address severe label imbalance by assigning greater weight to rare or challenging samples, its removal causes the most notable drop in performance, with micro-F1 decreasing from 0.533 to 0.467 and macro-F1 from 0.388 to 0.264. This result underscores the importance of class-balanced optimization, as the model without BaML tends to overfit common findings and overlook rare but clinically significant abnormalities.

Impact of Multi-Scale Visual Feature Extraction The second variant ($_{w/o\text{Patch}}$) excludes patch-level feature extraction and fusion, relying solely on region-level features. Our full model employs region-aware cross-attention to integrate patch-level details, providing essential spatial and contextual cues for each anatomical region. Removing patch features results in decreased micro-precision (from 0.463 to 0.430), F1 score (from 0.533 to 0.476), and macro-F1, demonstrating that fine-grained patch representations enhance the model's ability to detect subtle or overlapping abnormalities that region-level features alone may miss.

Impact of Region-Abnormality Mapping The third variant ($_{w/o\text{Map}}$) removes the region-abnormality mapping matrix, allowing abnormality-specific aggregation to attend freely over all regions. This leads to a drop in macro-F1 (from 0.388 to 0.357), indicating reduced consistency across abnormality classes. Although micro-precision slightly increases, detailed analysis reveals instability in certain predictions. For example, precision for "Enlarged Cardiomediastinum" rises to 0.5, but recall falls sharply to 0.015, indicating that most true cases are missed. This underscores the importance of our mapping matrix in guiding the model to attend to clinically relevant regions, leading to more balanced and reliable abnormality detection.

Case Study

Figure 3 presents a representative example comparing reports generated by RefleXNet before and after self-reflection optimization, alongside the expert reference. In the initial output, the model incorrectly reports a small right pneumothorax and right basilar opacity suggestive of atelectasis or aspiration, and imprecisely localizes the tip of the right internal jugular central venous catheter to the mid superior vena cava (SVC).



Clinical Indication: right ij central venous catheter placement, assess position.
Ground Truth: portable ap upright chest radiograph obtained. midline sternotomy wires and mediastinal clips are again noted. there has been interval placement of a right ij central venous catheter with its tip located in the distal svc or cavoatrial junction. no pneumothorax. otherwise, no change. **(Support Devices)**

ReflexNet w/o SR: portable ap upright chest radiograph was provided. there has been interval placement of a right ij central venous catheter with its tip in the region of the mid svc. there is a small right pneumothorax with right basilar opacity which could represent atelectasis or aspiration. otherwise, no change. **(Lung Opacity, Atelectasis, Pneumothorax, Support Devices)**

RefleXNet: portable ap upright chest radiograph was provided. midline sternotomy wires and mediastinal clips are again noted. there has been interval placement of a right ij central venous catheter with its tip in the region of the low svc. there is no pneumothorax. otherwise, no change. (Support Devices)

Figure 3: Case study demonstrating improvements from self-reflection optimization.

After applying self-reflection optimization, the generated report shows clear improvements in both factual accuracy and anatomical localization. The model correctly omits any mention of pneumothorax or basilar opacity, aligning with the reference report. The catheter tip description is refined to the low SVC region, which is closer to the expert annotation specifying the distal SVC or cavoatrial junction. Additionally, the model begins to include details such as midline sternotomy wires and mediastinal clips, further matching the expert report.

This case demonstrates that targeted self-reflection not only corrects abnormality predictions but also enhances the anatomical precision and relevance of the generated text. By incorporating report-level feedback, the model can refine region-level representations and better capture clinical details. Overall, self-reflection optimization enables RefleXNet to generate radiology reports that are more accurate and consistent with real-world clinical documentation.

Conclusion

In this work, we presented RefleXNet, a unified framework that integrates multi-scale visual feature fusion, anatomical relational modeling, and a targeted self-reflection strategy for chest X-ray diagnosis and report generation. The model leverages report-level feedback to guide the selective refinement of abnormality predictions and the associated regional features, ensuring that corrections are clinically grounded rather than purely data-driven. This combination leads to improved clinical accuracy, stronger alignment between visual cues and textual findings, and more consistent diagnostic reasoning. Extensive experiments demonstrate that RefleXNet not only outperforms state-of-the-art baselines across both classification and report-generation tasks, while remaining computationally efficient, reinforcing its practicality for scalable CXR interpretation.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62372380, U22B2036, Key Research and Development Plan of Shaanxi Province under Grant 2024GX-ZDCYL-01-05, National Key Research and Development Project under Grant 2022YFB3104005, Natural Science Basic Research Program of Shaanxi under Grants 2024JC-YBMS-513, the CCF-NSFOCUS 'Kunpeng' Research Fund under Grant CCF-NSFOCUS2025003, and Huawei Research Project under Grant TC20250730001.

References

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.

Chen, Z.; Song, Y.; Chang, T.; and Wan, X. 2020. Generating Radiology Reports via Memory-driven Transformer. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP*, 1439–1449. Association for Computational Linguistics.

Chen, Z.; Varma, M.; Delbrouck, J.; Paschali, M.; Blankemeier, L.; Veen, D. V.; Valanarasu, J. M. J.; Youssef, A.; Cohen, J. P.; Reis, E. P.; Tsai, E. B.; Johnston, A.; Olsen, C.; Abraham, T. M.; Gatidis, S.; Chaudhari, A. S.; and Langlotz, C. P. 2024. CheXagent: Towards a Foundation Model for Chest X-Ray Interpretation. *CoRR*, abs/2401.12208.

Delbrouck, J.; Chambon, P. J.; Bluethgen, C.; Tsai, E. B.; Almusa, O.; and Langlotz, C. P. 2022. Improving the Factual Correctness of Radiology Report Generation with Semantic Rewards. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, 4348–4360. Association for Computational Linguistics.

Demner-Fushman, D.; Kohli, M. D.; Rosenman, M. B.; Shooshan, S. E.; Rodriguez, L.; Antani, S.; Thoma, G. R.; and McDonald, C. J. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2): 304–310.

Driess, D.; Xia, F.; Sajjadi, M. S. M.; Lynch, C.; Chowdhery, A.; Ichter, B.; Wahid, A.; Tompson, J.; Vuong, Q.; Yu, T.; Huang, W.; Chebotar, Y.; Sermanet, P.; Duckworth, D.; Levine, S.; Vanhoucke, V.; Hausman, K.; Toussaint, M.; Greff, K.; Zeng, A.; Mordatch, I.; and Florence, P. 2023. PaLM-E: An Embodied Multimodal Language Model. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, 8469–8488. PMLR.

Hou, W.; Xu, K.; Cheng, Y.; Li, W.; and Liu, J. 2023. ORGAN: Observation-Guided Radiology Report Generation via Tree Reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8108–8122. Toronto, Canada: Association for Computational Linguistics.

Hu, X.; Gu, L.; An, Q.; Zhang, M.; Liu, L.; Kobayashi, K.; Harada, T.; Summers, R. M.; and Zhu, Y. 2023. Expert Knowledge-Aware Image Difference Graph Representation Learning for Difference-Aware Medical Visual Question Answering. In Singh, A. K.; Sun, Y.; Akoglu, L.; Gunopulos, D.; Yan, X.; Kumar, R.; Ozcan, F.; and Ye, J., eds., *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023*, 4156–4165. ACM.

Huang, Z.; Zhang, X.; and Zhang, S. 2023. KiUT: Knowledge-injected U-Transformer for Radiology Report Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, 19809–19818. IEEE.

Hyland, S. L.; Bannur, S.; Bouzid, K.; Castro, D. C.; Ranjit, M.; Schwaighofer, A.; Pérez-García, F.; Salvatelli, V.; Srivastav, S.; Thieme, A.; Codella, N.; Lungren, M. P.; Wetscherek, M. T.; Oktay, O.; and Alvarez-Valle, J. 2023. MAIRA-1: A specialised large multimodal model for radiology report generation. *CoRR*, abs/2311.13668.

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

Johnson, A. E. W.; Pollard, T. J.; Berkowitz, S. J.; Greenbaum, N. R.; Lungren, M. P.; Deng, C.; Mark, R. G.; and Horng, S. 2019. MIMIC-CXR: A large publicly available database of labeled chest radiographs. *CoRR*, abs/1901.07042.

Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; and Gao, J. 2023a. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Li, L.; Gan, Z.; Cheng, Y.; and Liu, J. 2019. Relation-Aware Graph Attention Network for Visual Question Answering. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 10312–10321. IEEE.

Li, M.; Lin, B.; Chen, Z.; Lin, H.; Liang, X.; and Chang, X. 2023b. Dynamic Graph Enhanced Contrastive Learning for Chest X-Ray Report Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, 3334–3343. IEEE.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*,

74–81. Barcelona, Spain: Association for Computational Linguistics.

Lin, Q.; Zhu, Y.; Mei, X.; Huang, L.; Ma, J.; He, K.; Peng, Z.; Cambria, E.; and Feng, M. 2025. Has multimodal learning delivered universal intelligence in healthcare? A comprehensive survey. *Inf. Fusion*, 116: 102795.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

McBee, M. P.; Awan, O. A.; Colucci, A. T.; Ghobadi, C. W.; Kadom, N.; Kansagra, A. P.; Tridandapani, S.; and Auffermann, W. F. 2018. Deep Learning in Radiology. *Academic Radiology*, 25(11): 1472–1480.

Mei, X.; Mao, R.; Cai, X.; Yang, L.; and Cambria, E. 2024. Medical Report Generation via Multimodal Spatio-Temporal Fusion. In Cai, J.; Kankanhalli, M. S.; Prabhakaran, B.; Boll, S.; Subramanian, R.; Zheng, L.; Singh, V. K.; César, P.; Xie, L.; and Xu, D., eds., *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, 4699–4708. ACM.

OpenAI. 2023. GPT-4V(ision) System Card.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. ACL.

Perera, D.; Habib, G.; Xu, Q.; Tan, D.; He, K.; Cambria, E.; and Feng, M. 2025. Beyond Prediction: Reinforcement Learning as the Defining Leap in Healthcare AI. *arXiv preprint arXiv:2508.21101*.

Pérez-García, F.; Sharma, H.; Bond-Taylor, S.; Bouzid, K.; Salvatelli, V.; Ilse, M.; Bannur, S.; Castro, D. C.; Schwaighofer, A.; Lungren, M. P.; et al. 2025. Exploring scalable medical image encoders beyond text supervision. *Nature Machine Intelligence*, 1–12.

Rahmat, T.; Ismail, A.; and Aliman, S. 2018. Chest x-rays image classification in medical image analysis. *Applied Medical Informatics*, 40(3-4): 63–73.

Sloan, P.; Clatworthy, P.; Simpson, E.; and Mirmehdi, M. 2024. Automated Radiology Report Generation: A Review of Recent Advances. *CoRR*, abs/2405.10842.

Smit, A.; Jain, S.; Rajpurkar, P.; Pareek, A.; Ng, A. Y.; and Lungren, M. P. 2020. Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, 1500–1519. Association for Computational Linguistics.

Tanida, T.; Müller, P.; Kaissis, G.; and Rueckert, D. 2023. Interactive and Explainable Region-guided Radiology Report Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, 7433–7442. IEEE.

Tu, T.; Azizi, S.; Driess, D.; Schaekermann, M.; Amin, M.; Chang, P.; Carroll, A.; Lau, C.; Tanno, R.; Ktena, I.; Mustafa, B.; Chowdhery, A.; Liu, Y.; Kornblith, S.; Fleet, D. J.; Mansfield, P. A.; Prakash, S.; Wong, R.; Virmani, S.; Semturs, C.; Mahdavi, S. S.; Green, B.; Dominowska, E.; y Arcas, B. A.; Barral, J. K.; Webster, D. R.; Corrado, G. S.; Matias, Y.; Singhal, K.; Florence, P.; Karthikesalingam, A.; and Natarajan, V. 2023. Towards Generalist Biomedical AI. *CoRR*, abs/2307.14334.

Wu, J.; He, K.; Mao, R.; Li, C.; and Cambria, E. 2023. MEGACare: Knowledge-guided multi-view hypergraph predictive framework for healthcare. *Inf. Fusion*, 100: 101939.

Wu, J.; Mei, X.; Mao, R.; He, K.; and Cambria, E. 2026. TAKECare: A temporal-hierarchical framework with knowledge fusion for personalized clinical predictive modeling. *Inf. Fusion*, 126: 103620.

Wu, J. T.; Agu, N.; Lourentzou, I.; Sharma, A.; Paguio, J. A.; Yao, J. S.; Dee, E. C.; Mitchell, W.; Kashyap, S.; Giovannini, A.; Celi, L. A.; and Moradi, M. 2021. Chest ImaGenome Dataset for Clinical Reasoning. In Vanschoren, J.; and Yeung, S., eds., *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Yao, T.; Pan, Y.; Li, Y.; and Mei, T. 2018. Exploring Visual Relationship for Image Captioning. In Ferrari, V.; Hebert, M.; Sminchisescu, C.; and Weiss, Y., eds., *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIV*, volume 11218 of *Lecture Notes in Computer Science*, 711–727. Springer.

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. 2023. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9).

Zambrano Chaves, J. M.; Huang, S.-C.; Xu, Y.; Xu, H.; Usuyama, N.; Zhang, S.; Wang, F.; Xie, Y.; Khademi, M.; Yang, Z.; et al. 2025. A clinically accessible small multimodal radiology model and evaluation metric for chest X-ray findings. *Nature Communications*, 16(1): 3108.

Zhang, H.; Li, Z.; Li, H.; Zhou, X.; Zhang, J.; and Li, Y. 2025. TransFR: Transferable Federated Recommendation with Adapter Tuning on Pre-trained Language Models. arXiv:2402.01124.

Zhang, Y.; Wang, X.; Xu, Z.; Yu, Q.; Yuille, A. L.; and Xu, D. 2020. When Radiology Report Generation Meets Knowledge Graph. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, 12910–12917. AAAI Press.

Zhao, Y.; Gong, M.; Zhang, M.; Qin, A. K.; Jiang, F.; and Li, J. 2025. SPCNet: Deep Self-Paced Curriculum Network Incorporated With Inductive Bias. *IEEE Transactions on Neural Networks and Learning Systems*, 36(8): 15029–15042.