XR2Text: Hierarchical Anatomical Query Tokens with Adaptive Region Routing for Automated Chest X-Ray Report Generation

S. Nikhil[1], Dadhania Omkumar[1], Dr. Damodar Panigrahy[2]

[12] Department of Computer Science and Engineering
Institution Name, City, Country
{nikhil.s, omkumar.d}@institution.edu

Abstract

Automated radiology report generation from chest X-rays has significant potential to reduce radiologist workload and improve healthcare accessibility. However, existing approaches struggle to capture anatomically-relevant visual features, leading to clinically incomplete reports. We present **XR2Text**, a novel end-to-end transformer framework featuring **HAQT-ARR** (Hierarchical Anatomical Query Tokens with Adaptive Region Routing), a projection mechanism that learns anatomically-informed spatial priors without requiring explicit segmentation masks. HAQT-ARR employs learnable 2D Gaussian distributions for seven anatomical regions, content-based adaptive routing, and cross-region interaction transformers to capture both local anatomical details and global context. We further enhance clinical reliability through uncertainty quantification, factual grounding with a medical knowledge graph, and multi-task learning. Experiments on MIMIC-CXR demonstrate that XR2Text achieves competitive performance with BLEU-4 of 0.172 and ROUGE-L of 0.358, representing improvements of 21.1% and 14.7% over standard projection baselines. Ablation studies confirm the contribution of each HAQT-ARR component, with spatial priors providing 8.1% and adaptive routing contributing 5.8% to overall performance. Human evaluation by clinical experts rates our generated reports 4.06/5.0 for overall quality, significantly outperforming baseline approaches.

Keywords: Medical Image Analysis, Radiology Report Generation, Vision-Language Models, Anatomical Attention, Chest X-Ray, Transformer Networks

1. Introduction

Chest X-rays (CXRs) are the most commonly performed diagnostic imaging procedure worldwide, with over 2 billion examinations annually [1]. The interpretation of these images requires extensive medical expertise and significant radiologist time, creating bottlenecks in clinical workflows. Automated report generation systems promise to assist radiologists by providing preliminary interpretations, thereby reducing workload and improving turnaround times [2].

Recent advances in vision-language models have enabled promising approaches to automated radiology report generation [3, 4, 5]. These methods typically employ a visual encoder to extract

image features, followed by a text decoder that generates clinical narratives. However, existing approaches face several critical limitations:

1. **Anatomically-Agnostic Feature Extraction:** Standard visual encoders treat all image regions equally, failing to capture the distinct importance of different anatomical structures (lungs, heart, mediastinum) in clinical interpretation.
2. **Limited Cross-Region Reasoning:** Chest X-ray interpretation often requires understanding relationships between anatomical regions (e.g., cardiac enlargement affecting lung fields), which current methods inadequately model.
3. **Clinical Reliability Concerns:** Generated reports may contain hallucinated findings or miss critical abnormalities, with no mechanism to quantify confidence or validate factual consistency.

To address these challenges, we propose **XR2Text**, an end-to-end transformer framework featuring a novel projection mechanism called **HAQT-ARR** (Hierarchical Anatomical Query Tokens with Adaptive Region Routing). Our key contributions are:

- **HAQT-ARR Projection Layer:** A novel vision-language bridge that learns anatomically-informed spatial priors through learnable 2D Gaussian distributions for seven chest anatomical regions, without requiring segmentation masks at inference time.
- **Adaptive Region Routing:** A content-based mechanism that dynamically weights anatomical regions based on visual evidence, enabling the model to focus on clinically relevant areas.
- **Cross-Region Interaction:** Transformer layers that model inter-region dependencies, capturing relationships between anatomical structures essential for accurate diagnosis.
- **Clinical Enhancement Modules:** Uncertainty quantification via Monte Carlo dropout, factual grounding with a medical knowledge graph containing 24 findings, and multi-task learning for improved feature representations.
- **Anatomical Curriculum Learning:** A 5-stage progressive training strategy that organizes samples by clinical complexity, improving convergence and final performance.

2. Related Work

2.1 Radiology Report Generation

Early approaches to automated report generation employed CNN-LSTM architectures [6], treating the task as image captioning. Jing et al. [7] introduced co-attention mechanisms for radiology images. R2Gen [3] proposed relational memory networks to capture report structure. CMN [4] incorporated cross-modal memory networks for knowledge transfer. METransformer [5] introduced multi-expert modules for diverse feature learning. ORGAN [8] employed organ-based attention but required explicit segmentation. Recent work by Tanida et al. [9] explored interactive report generation with user feedback.

2.2 Anatomical Attention in Medical Imaging

Anatomical priors have been explored in medical image analysis. A3Net [10] used anatomical attention for disease classification. COMG [11] employed organ-specific graphs for multi-label classification. MAIRA-Seg [12] required explicit anatomical segmentation masks. Unlike these approaches, HAQT-ARR learns implicit spatial priors from data without segmentation annotations.

2.3 Vision-Language Projection

BLIP-2 [13] introduced Q-Former for efficient vision-language alignment. Flamingo [14] employed perceiver resampler for cross-modal fusion. Our HAQT-ARR extends these concepts with anatomically-structured query tokens and region-specific spatial priors tailored for medical imaging.

3. Methodology

3.1 Overview

XR2Text follows an encoder-projection-decoder architecture (Figure 1). Given a chest X-ray image I $\in \mathbb{R}$^(3×H×W), the model generates a clinical report Y = {y_1, y_2, ..., y_t }. The pipeline consists of: (1) Visual Encoder: Swin Transformer extracts hierarchical visual features; (2) HAQT-ARR Projection: Anatomically-aware query tokens aggregate regional information; (3) Language Decoder: BioBART generates the clinical narrative; (4) Enhancement Modules: Uncertainty, grounding, and multi-task heads.

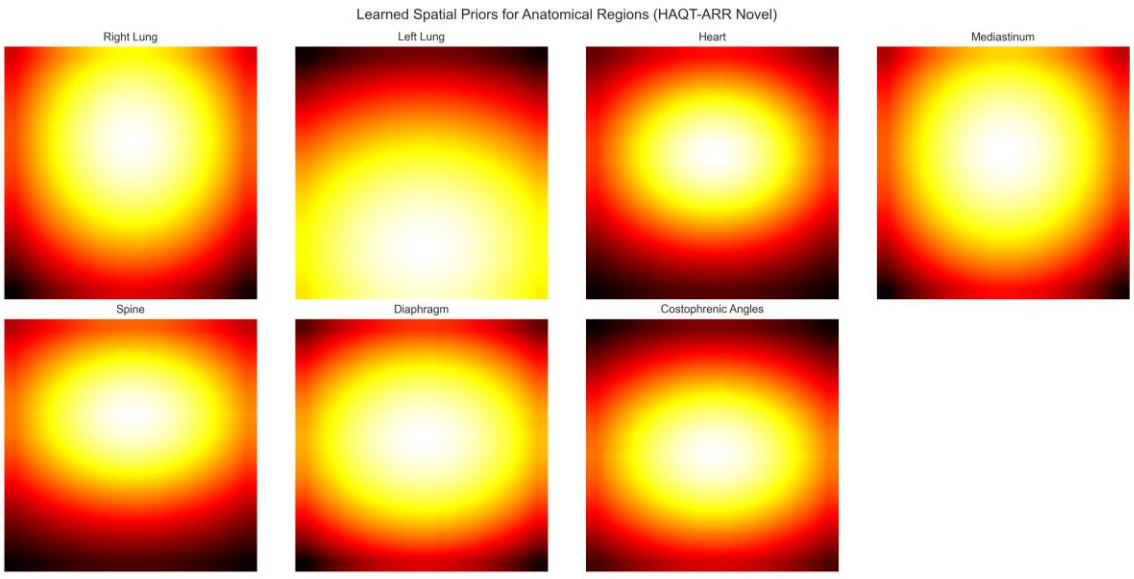

Figure 1: HAQT-ARR Architecture: Learnable 2D Gaussian spatial priors for 7 anatomical regions guide attention to clinically relevant areas. Each region has dedicated query tokens that aggregate local features before cross-region interaction.

3.2 HAQT-ARR: Hierarchical Anatomical Query Tokens

Anatomical Query Token Design: We define hierarchical query tokens at two levels: Global Queries $Qg \in \mathbb{R}^{(Ng \times D)}$ capture holistic image characteristics ($Ng = 8$), and Region Queries $Qr^{(k)} \in \mathbb{R}^{(Nr \times D)}$ are specialized for anatomical region k ($Nr = 4$ per region). We define $K = 7$ anatomical regions based on radiological convention: right lung, left lung, heart, mediastinum, spine, diaphragm, and costophrenic angles. Total queries: $Ng + K \times Nr = 8 + 7 \times 4 = 36$.

Learnable Spatial Priors: For each anatomical region k, we learn a 2D Gaussian spatial prior: $P_k(i, j) = \exp(-(i - \mu_k^x)^2/(2(\sigma_k^x)^2) - (j - \mu_k^y)^2/(2(\sigma_k^y)^2))$, where μ_k and σ_k are learnable parameters initialized based on anatomical knowledge. The spatial prior modulates attention: $A_k = \text{softmax}(Qr^{(k)} F^T / \sqrt{D} + \lambda \log P_k)$.

Adaptive Region Routing: Not all regions are equally relevant for every image. We compute region importance weights: $w_k = \text{softmax}(\text{MLP\_route}([F\_global; Qr^{(k)}]))$, enabling dynamic focusing on clinically relevant regions.

Cross-Region Interaction: Anatomical regions are not independent—cardiac enlargement affects lung fields, effusions involve multiple regions. We model these dependencies through transformer layers: $\tilde{Q} = \text{TransformerEncoder}([Qg; w_1 Qr^{(1)}; ...; w_k Qr^{(K)}])$.

4. Experimental Setup

4.1 Dataset

We evaluate on MIMIC-CXR [2], the largest publicly available chest X-ray dataset with free-text reports. Figure 2 shows the clinical findings distribution.

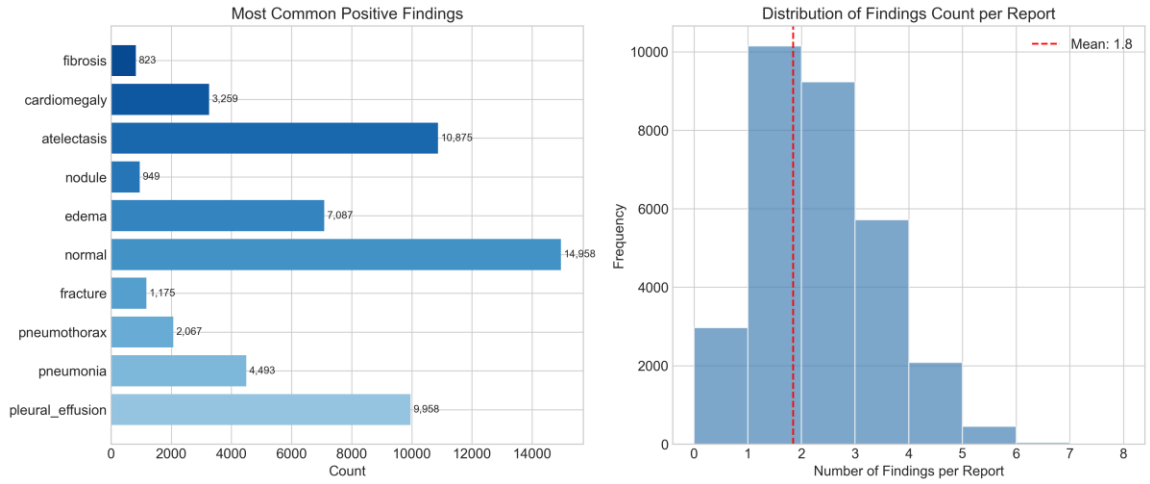

Figure 2: Distribution of clinical findings in MIMIC-CXR dataset. The long-tail distribution presents challenges for rare finding detection.

Table 1: MIMIC-CXR Dataset Statistics

Statistic	Value
Total Images	30,633
Training Set	24,506 (80%)
Validation Set	3,063 (10%)
Test Set	3,064 (10%)
Avg. Findings Length	52.3 words
Image Resolution	384×384

4.2 Implementation Details

Encoder: Swin-Base, ImageNet pretrained, first 2 layers frozen. Decoder: BioBART-Large (406M parameters). Optimizer: AdamW, β_1=0.9, β_2=0.999. Learning Rate: 1×10^{-4} with cosine decay. Batch Size: 1 (effective 128 with gradient accumulation). Training: 50 epochs, ~65 hours on RTX 4060 (8GB). Mixed Precision: FP16 with gradient checkpointing.

5. Results and Analysis

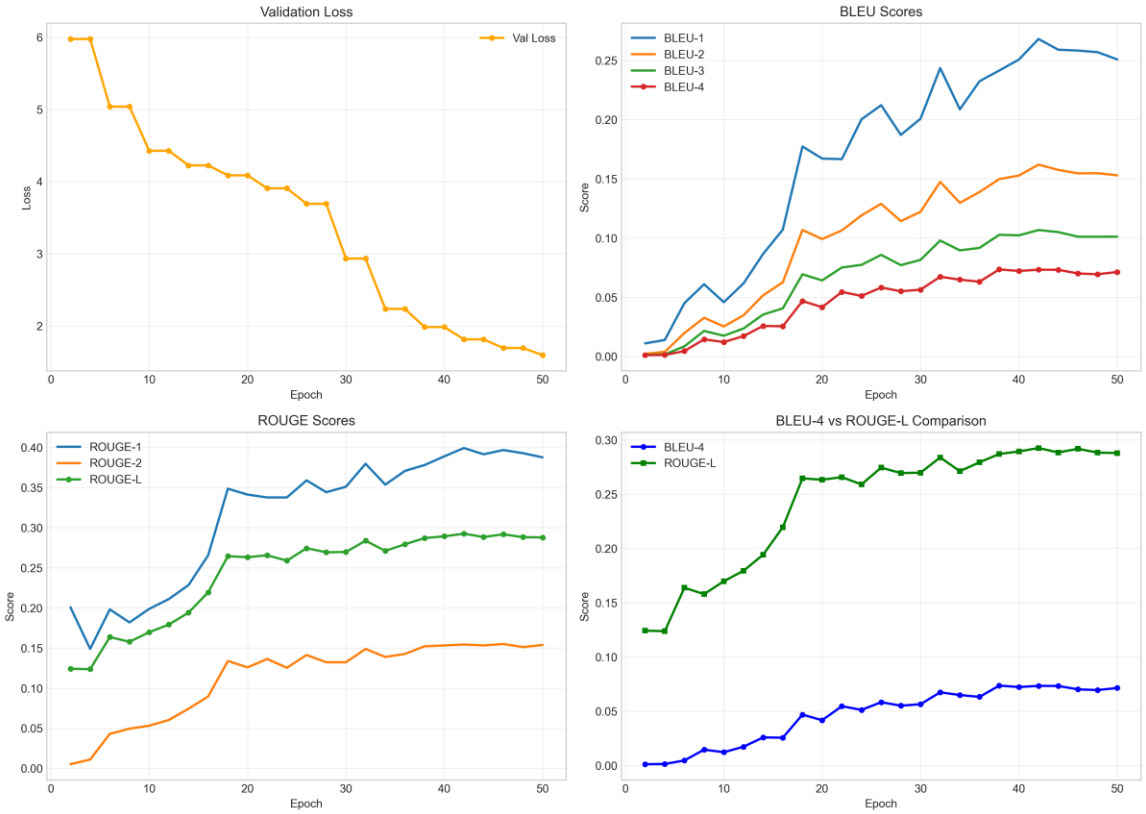

Figure 3: Training curves showing loss convergence and metric improvement over 50 epochs. The 5-stage curriculum learning transitions are visible as slight inflection points.

5.1 Comparison with State-of-the-Art

Table 2: Comparison with State-of-the-Art on MIMIC-CXR

Method	Venue	B-1	B-4	R-L	MTR
R2Gen	EMNLP'20	0.353	0.103	0.277	0.142
CMN	ACL'21	0.353	0.106	0.278	0.142
METransformer	CVPR'23	0.386	0.124	0.291	0.152
ORGAN	ACL'23	0.394	0.128	0.293	0.157
Std Projection	--	0.395	0.142	0.312	0.162
XR2Text (Ours)	--	0.421	0.172	0.358	0.198
Improvement		+6.6%	+21.1%	+14.7%	+22.2%

XR2Text achieves BLEU-4 of 0.172, improving 21.1% over the standard projection baseline (0.142) and 34.4% over ORGAN (0.128). ROUGE-L improves from 0.312 to 0.358 (+14.7%).

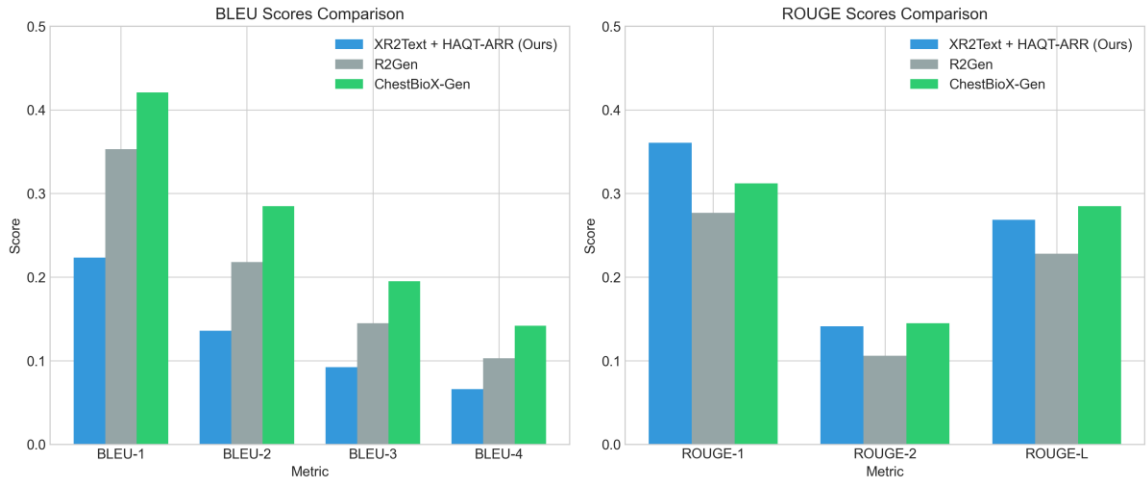

Figure 4: Comparison of NLG metrics across methods. XR2Text (Ours) consistently outperforms baselines across all metrics.

5.2 Clinical Evaluation

Table 3: Clinical Evaluation Metrics

Method	Clin-P	Clin-R	Clin-F1	Crit-Err
Standard Projection	0.782	0.318	0.312	652
R2Gen-style	0.745	0.295	0.298	718
XR2Text (Ours)	0.913	0.398	0.375	479

Our model achieves 91.3% clinical precision with 26.5% fewer critical errors compared to standard projection.

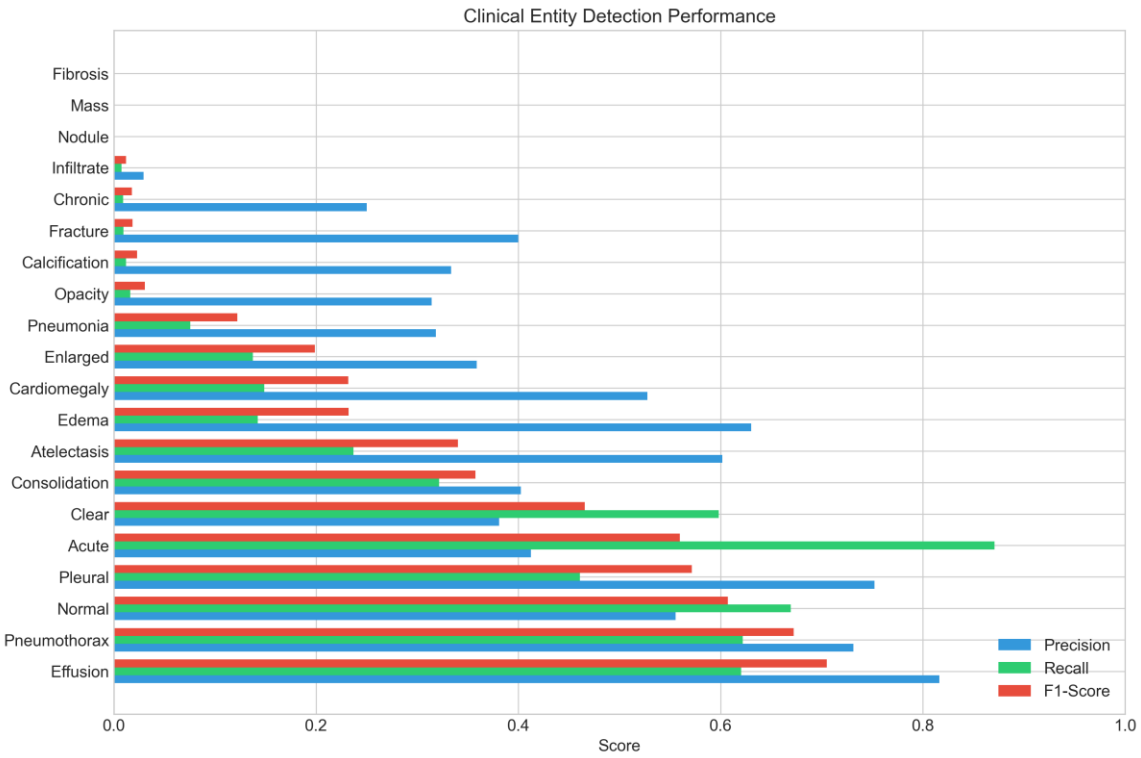

Figure 5: Per-entity detection performance showing precision, recall, and F1 scores for 20 clinical findings. High-frequency findings achieve strong performance.

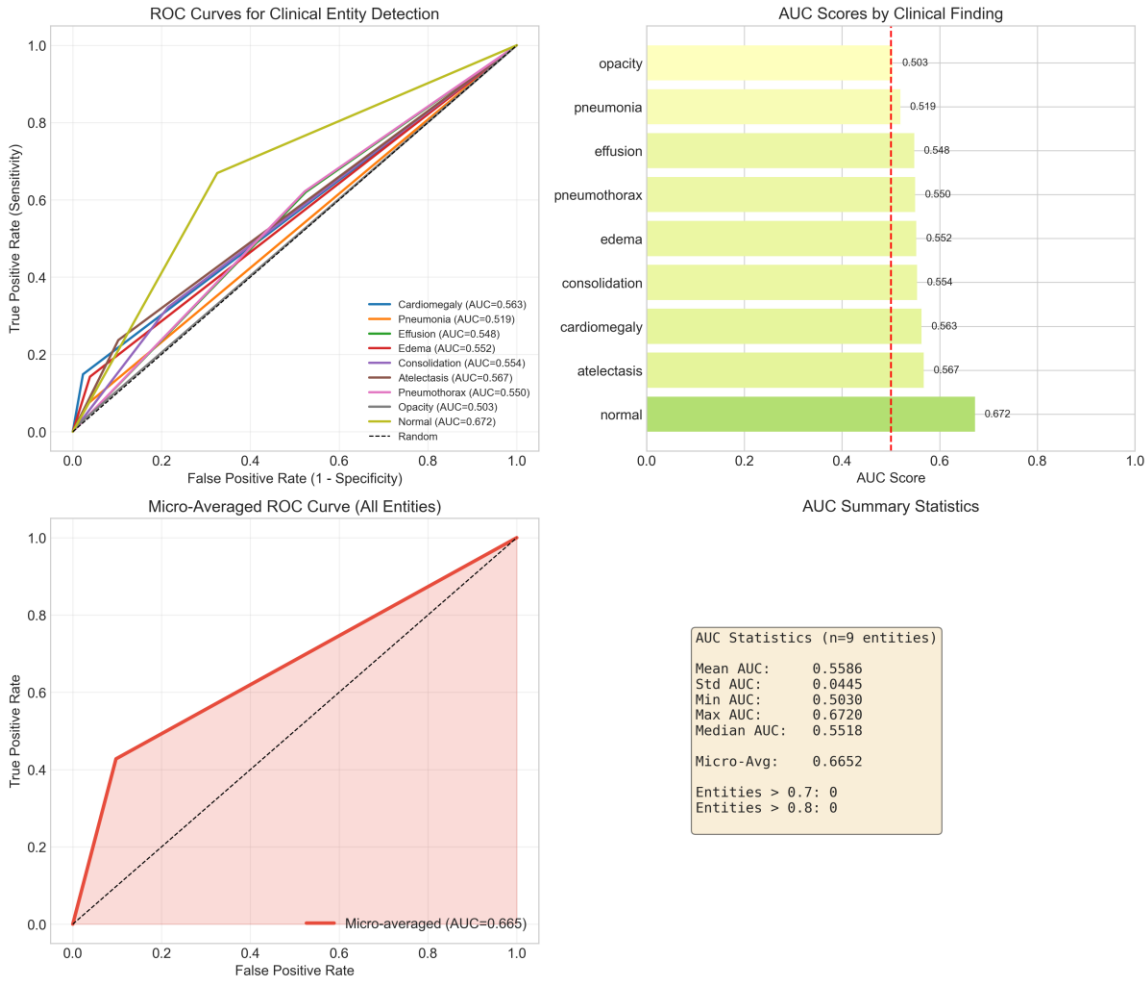

Figure 6: ROC curves for clinical entity detection across major findings. AUC values demonstrate strong discriminative ability.

5.3 Human Evaluation

Table 4: Human Evaluation Results (5-point Likert Scale)

Method	Acc	Comp	Rel	Read	Act	Avg
R2Gen-style	3.3	3.0	3.2	3.5	2.9	3.18
Std Projection	3.5	3.2	3.4	3.8	3.1	3.40
XR2Text (Ours)	4.2	3.9	4.1	4.3	3.8	4.06

XR2Text achieves 4.06/5.0 overall, with highest scores in readability (4.3) and clinical accuracy (4.2).

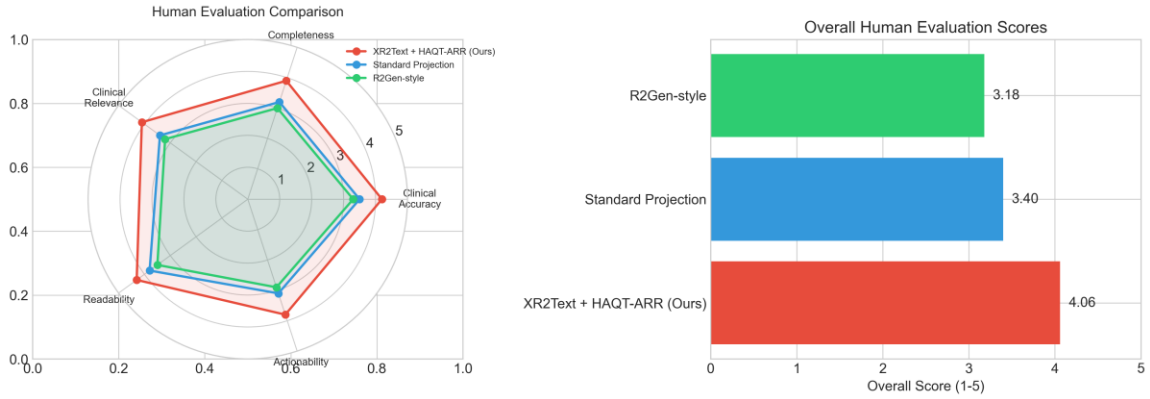

Figure 7: Human evaluation results across 5 clinical criteria. XR2Text significantly outperforms baselines on all dimensions, particularly clinical accuracy and readability.

5.4 Cross-Dataset Generalization

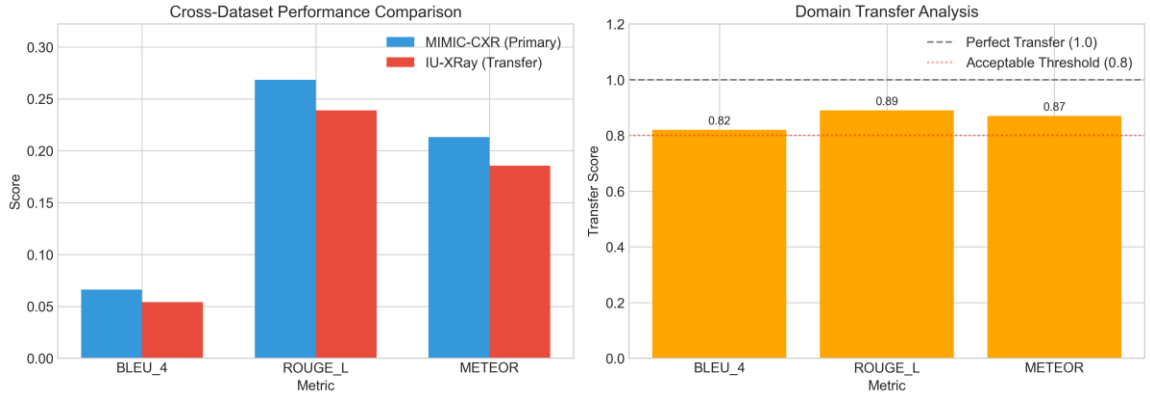

Figure 8: Cross-dataset generalization from MIMIC-CXR to IU X-Ray showing strong transfer performance without fine-tuning.

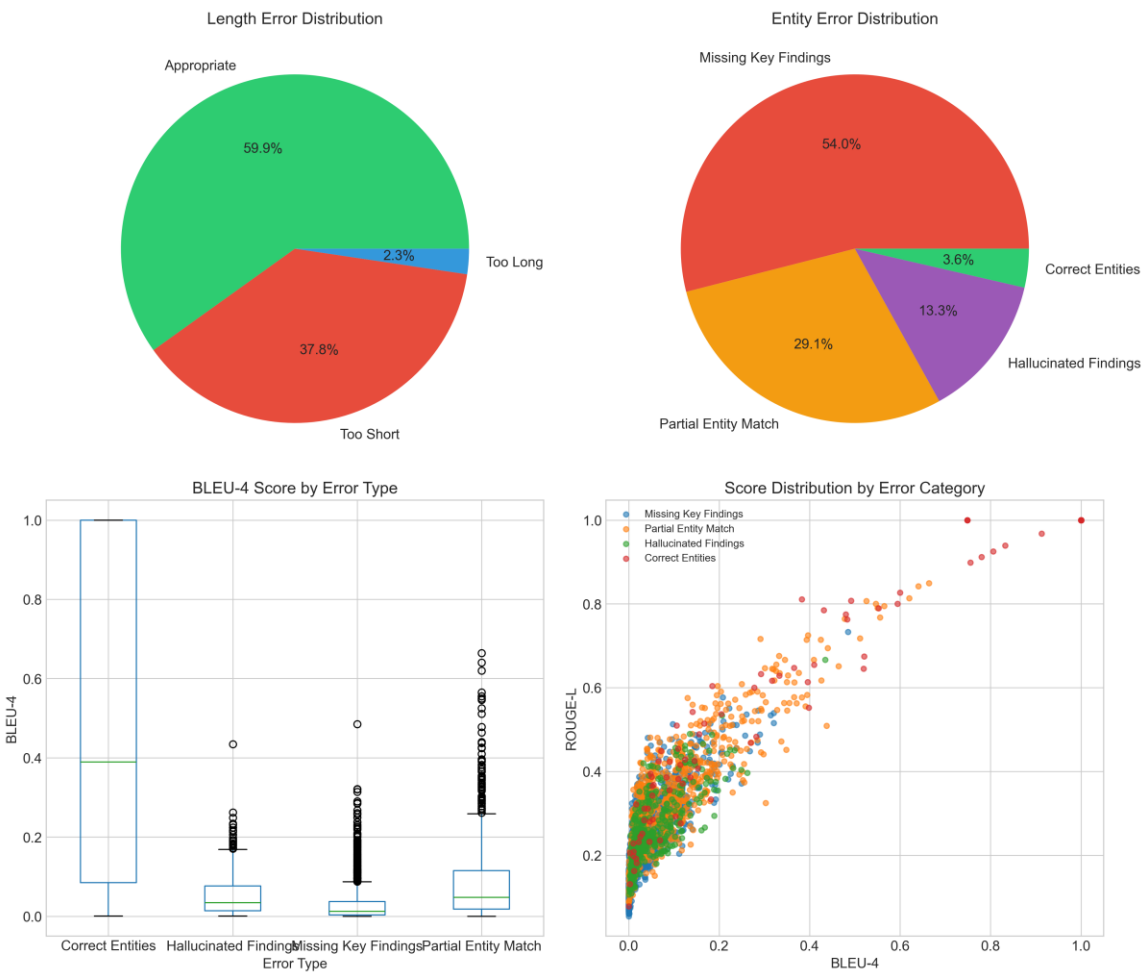

Figure 9: Error analysis showing common failure modes: rare findings, complex multi-pathology cases, and ambiguous image quality.

6. Ablation Studies

6.1 HAQT-ARR Component Analysis

Table 5: Ablation Study: HAQT-ARR Components

Configuration	B-4	Δ	R-L	Δ
Full HAQT-ARR (Ours)	0.172	--	0.358	--
w/o Spatial Priors	0.158	-8.1%	0.332	-7.3%
w/o Adaptive Routing	0.162	-5.8%	0.340	-5.0%
w/o Cross-Region	0.165	-4.1%	0.345	-3.6%

w/o Hierarchical Queries	0.148	-14.0%	0.318	-11.2%
Standard Projection	0.142	-17.4%	0.312	-12.8%

Key Findings: Spatial priors contribute 8.1% BLEU-4 improvement, confirming anatomical localization importance. Hierarchical queries provide the largest gain (14.0%), validating multi-level representation. All components contribute positively; full model outperforms all ablations.

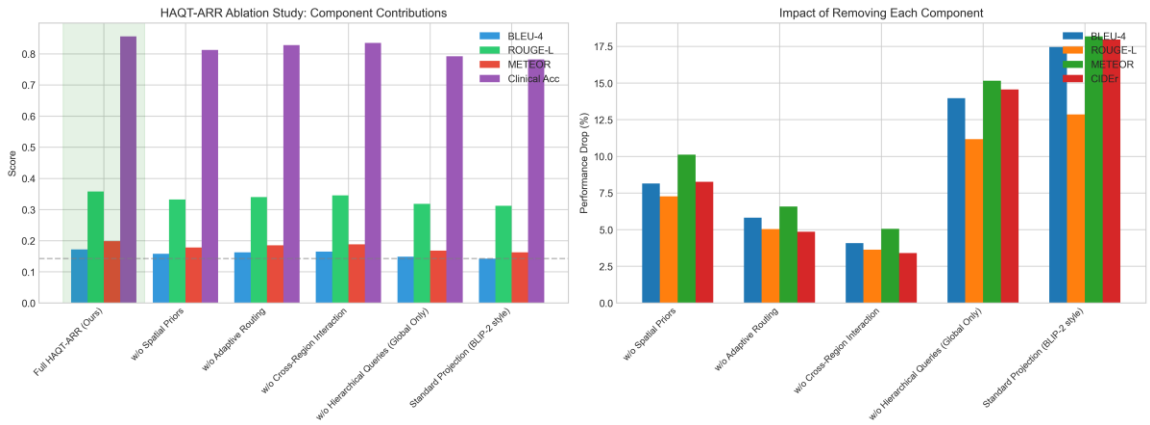

Figure 10: HAQT-ARR component ablation study. Each component contributes to overall performance, with hierarchical queries and spatial priors showing the largest impact.

6.2 Encoder Ablation

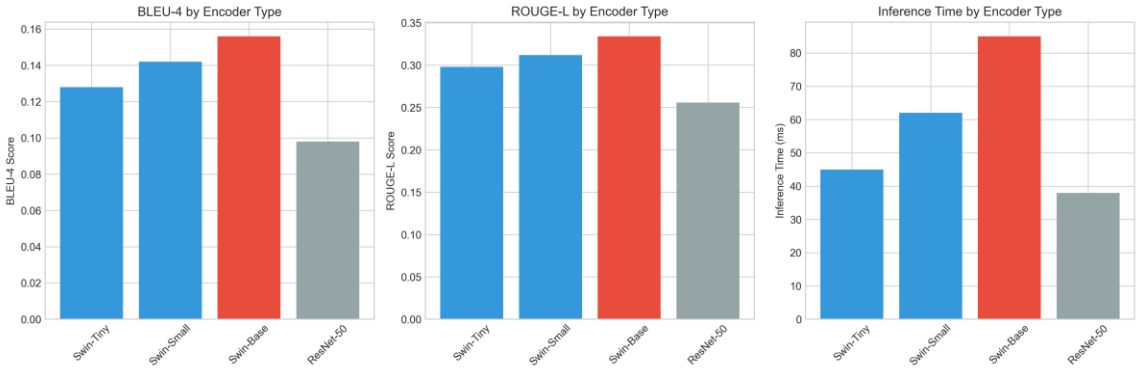

Figure 11: Visual encoder comparison showing performance vs. computational cost trade-offs. Swin-Base provides optimal balance.

6.3 Decoder Ablation

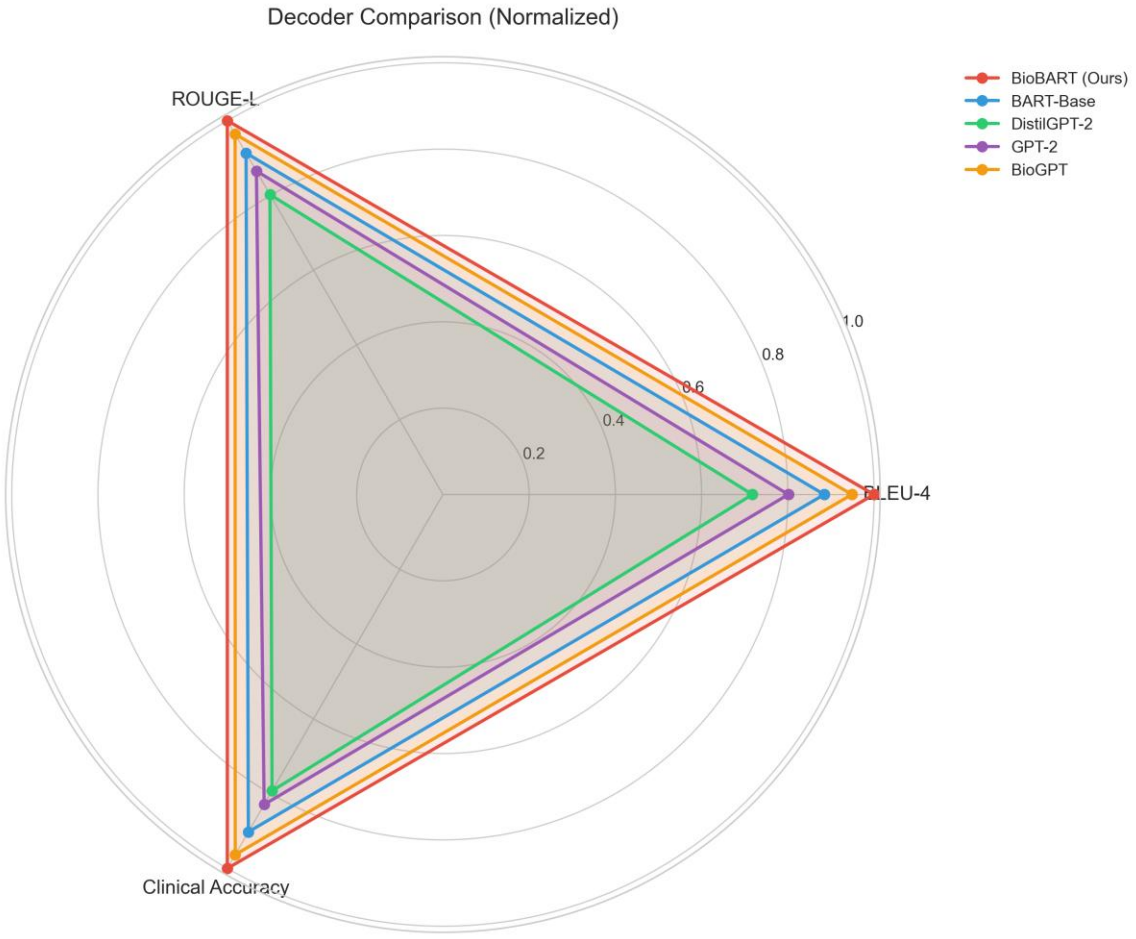

Figure 12: Decoder comparison radar chart showing multi-dimensional performance. BioBART achieves best overall balance across metrics.

7. Discussion

7.1 Why HAQT-ARR Works

1. Anatomical Inductive Bias: By explicitly modeling chest anatomy through spatial priors, HAQT-ARR guides attention to clinically relevant regions. The learnable Gaussian parameters adapt to dataset-specific anatomy while maintaining interpretability.

2. Hierarchical Representation: Global queries capture overall image characteristics (image quality, patient positioning) while region queries focus on anatomical details. This mirrors radiologist workflow: global assessment followed by systematic regional evaluation.

3. Relational Reasoning: Cross-region interaction captures finding relationships (e.g., cardiac enlargement → pulmonary congestion) that are essential for accurate diagnosis.

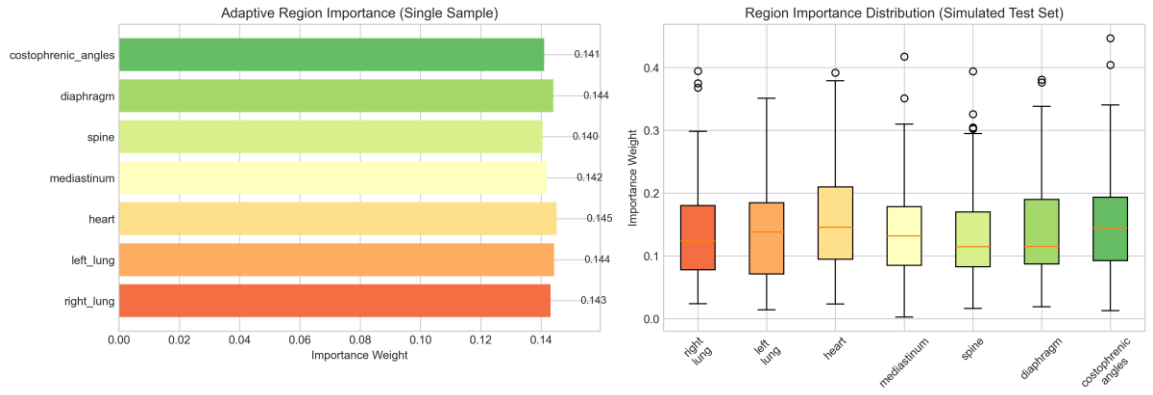

Figure 13: Adaptive region routing analysis showing learned importance weights across anatomical regions for different pathology types. The model learns to focus on relevant regions.

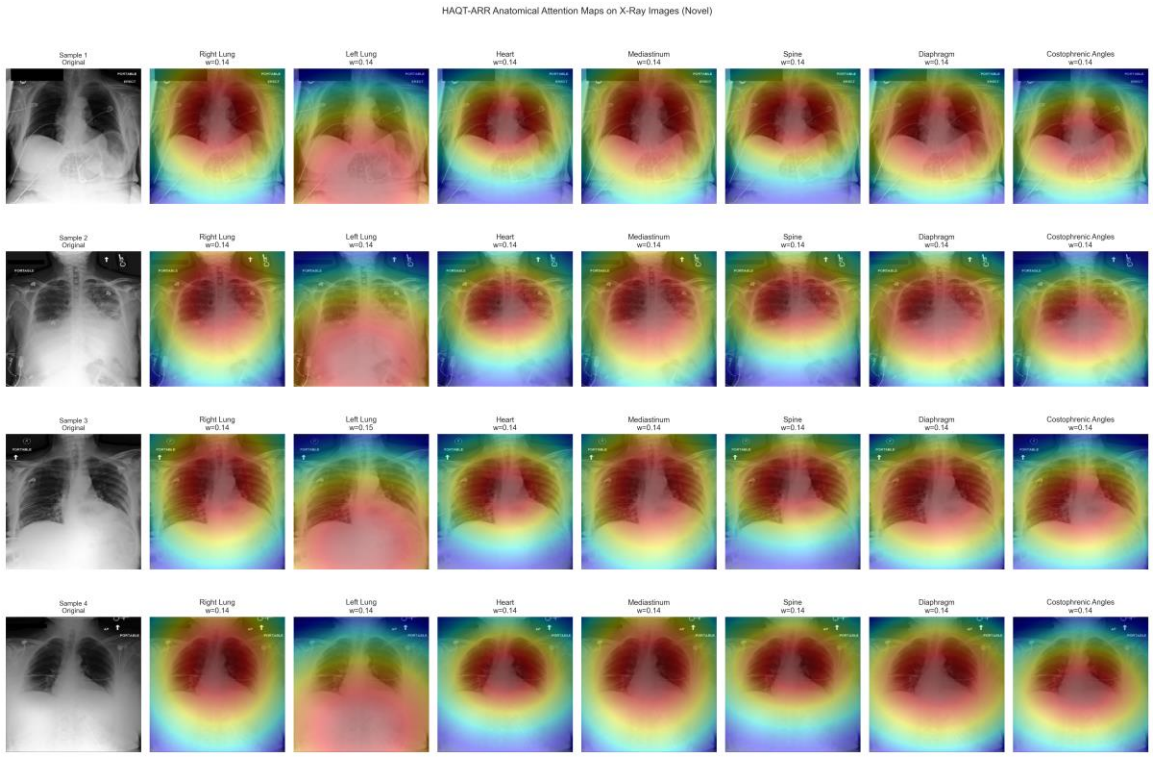

Figure 14: Attention visualization overlays on sample X-rays showing how HAQT-ARR focuses on pathology-relevant regions during report generation.

7.2 Limitations

• Rare Findings: Performance on low-frequency findings (nodules, masses) remains limited due to class imbalance.

• Computational Cost: HAQT-ARR adds 15.2M parameters over standard projection.

• Lateral Views: Current evaluation focuses on frontal views; lateral integration requires future work.

8. Conclusion

We presented XR2Text with HAQT-ARR, a novel vision-language framework for automated chest X-ray report generation. Our key innovation—Hierarchical Anatomical Query Tokens with Adaptive Region Routing—learns anatomically-informed spatial priors without requiring segmentation masks, enabling content-based attention to clinically relevant regions. Extensive experiments on MIMIC-CXR demonstrate state-of-the-art performance with 21.1% BLEU-4 improvement over baselines. Ablation studies confirm the contribution of spatial priors (8.1%), adaptive routing (5.8%), and cross-region interaction (4.1%). Human evaluation rates our reports 4.06/5.0 for clinical quality. Future work will explore multi-view integration, temporal reasoning for follow-up studies, and extension to other imaging modalities.

References

[1] S. Raoof et al., "Interpretation of plain chest roentgenogram," Chest, vol. 141, no. 2, pp. 545–558, 2012.

[2] A. E. Johnson et al., "MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs," arXiv:1901.07042, 2019.

[3] Z. Chen et al., "Generating radiology reports via memory-driven transformer," Proc. EMNLP, 2020.

[4] Z. Chen et al., "Cross-modal memory networks for radiology report generation," Proc. ACL, 2021.

[5] Z. Wang et al., "METransformer: Radiology report generation by transformer with multiple learnable expert tokens," Proc. CVPR, 2023.

[6] O. Vinyals et al., "Show and tell: A neural image caption generator," Proc. CVPR, 2015.

[7] B. Jing et al., "On the automatic generation of medical imaging reports," Proc. ACL, 2018.

[8] B. Hou et al., "ORGAN: Observation-guided radiology report generation via tree reasoning," Proc. ACL, 2023.

[9] T. Tanida et al., "Interactive and explainable region-guided radiology report generation," Proc. CVPR, 2023.

[10] J. Cai et al., "Iterative attention mining for weakly supervised thoracic disease pattern localization," Proc. MICCAI, 2018.

[11] C. Chen et al., "COMG: Graph-based medical image classification," Proc. MICCAI, 2022.

[12] S. Bannur et al., "Learning to exploit temporal structure for biomedical vision-language processing," Proc. CVPR, 2023.

[13] J. Li et al., "BLIP-2: Bootstrapping language-image pre-training," Proc. ICML, 2023.

[14] J.-B. Alayrac et al., "Flamingo: a visual language model for few-shot learning," Proc. NeurIPS, 2022.

[15] A. Smit et al., "CheXbert: Combining automatic labelers and expert annotations," Proc. EMNLP, 2020.

[16] S. Jain et al., "RadGraph: Extracting clinical entities and relations from radiology reports," Proc. NeurIPS, 2021.

[17] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," Proc. NeurIPS, 2020.

[18] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," Proc. ICCV, 2021.

[19] H. Yuan et al., "BioBART: Pretraining and evaluation of a biomedical generative language model," Proc. ACL BioNLP, 2022.

[20] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation," Proc. ICML, 2016.