CogRRG: A Swin-Base Cognitive Second-Reader for Structured Chest X-Ray Report Generation

***Abstract*—Radiology report generation is vulnerable to reader fatigue, motivating AI "second-readers" that produce clinically grounded drafts. We present CogRRG, a cognitively inspired framework for chest X-ray report generation that explicitly simulates three radiological reasoning stages: (i) hierarchical visual perception with anatomy alignment (PRO-FA), (ii) knowledge-enhanced multi-label diagnosis formation (MIX-MLP), and (iii) closed-loop hypothesis verification (RCTA). Our model uses a Swin-Base backbone to extract multi-scale features, aligns them to a curated CXR-RadLex concept bank, predicts pathology tags, and conditions report generation through a verification loop. Experimental results on MIMIC-CXR demonstrate strong performance with CheXpert micro-F1 of 0.654 and macro-F1 of 0.394 for pathology classification. The full system achieves CheXpert F1 of 0.56, RadGraph F1 of 0.52, and CIDEr of 0.45 on the test set, with successful domain generalization to IU X-Ray.**

***Index Terms*—Radiology Report Generation, Chest X-Ray, Clinical Efficacy, Swin Transformer, RadLex, CheXpert, RadGraph**

I. Introduction

Chest radiography is among the highest-volume imaging modalities, and the radiology reading-room environment introduces cognitive load that can increase reporting discrepancies. A reliable "second-reader" should (a) draft coherent *Findings* and *Impression* sections, (b) reduce hallucinations by grounding generation in the image, and (c) provide interpretable intermediate representations.

Most prior CXR report generators adopt encoder–decoder captioning paradigms, achieving good lexical overlap while underperforming on clinical correctness. We design a cognitive pipeline that makes reasoning explicit: perceive hierarchically, form a diagnosis hypothesis, and verify it against the image.

Contributions:

- A Swin-Base multi-view encoder producing token sets at pixel/region/organ levels (PRO-FA).
- A RadLex-aligned concept bank with alignment losses tethering visual tokens to anatomy concepts.
- A MIX-MLP multi-label classifier modeling disease co-occurrence.
- An RCTA triangular attention loop conditioning generation with closed-loop verification.
- Comprehensive evaluation demonstrating state-of-the-art clinical efficacy metrics.

II. Problem Definition

Given a study with CXR views (PA/AP/Lateral) and optional clinical indication, generate a structured report containing **Findings** (detailed visual observations) and **Impression** (diagnostic summary). Evaluation emphasizes:

- **Clinical Accuracy**: CheXpert F1 measuring pathology tagging correctness.
- **Structural Logic**: RadGraph F1 assessing entity-relation correctness.
- **NLG Fluency**: CIDEr and BLEU-4 for readability and coverage.

III. Datasets

A. MIMIC-CXR

MIMIC-CXR contains 377,110 chest radiographs from 227,835 studies at Beth Israel Deaconess Medical Center. We use the standard train/validation/test splits with multiple projection views (PA, AP, Lateral) per study.

B. IU X-Ray

IU X-Ray provides 7,470 chest X-rays from 3,955 reports from Indiana University, used for domain generalization evaluation.

C. Label Sources

We derive weak labels using a CheXpert-style labeler extracting 14 pathologies: Atelectasis, Cardiomegaly, Consolidation, Edema, Enlarged Cardiomediastinum, Fracture, Lung Lesion, Lung Opacity, Pleural Effusion, Pleural Other, Pneumonia, Pneumothorax, Support Devices, and No Finding.

IV. Method Overview

Figure 1 presents the CogRRG architecture, processing multi-view CXRs through three cognitive modules: hierarchical perception (PRO-FA), diagnosis formation (MIX-MLP), and hypothesis verification (RCTA).

V. PRO-FA: Hierarchical Perception

Swin Transformer's hierarchical design provides multi-resolution features suited to radiological interpretation. We use Swin-Base, processing each view through a shared encoder.

We define three hierarchical token sets:

- **Pixel-level** $\mathbf{T}^{pix}$ (Stage 1, 56×56): Fine anatomical details.
- **Region-level** $\mathbf{T}^{reg}$ (Stage 2, 28×28): Lobar patterns.
- **Organ-level** $\mathbf{T}^{org}$ (Stages 3-4, 7×7): Global structures.

Multi-view fusion uses cross-view attention:

$$\mathbf{T}^{g}_{\text{fused}} = \text{Attn}(\mathbf{Q} = \mathbf{T}^{g}_{PA}, \mathbf{K}, \mathbf{V} = [\mathbf{T}^{g}_{PA}, \mathbf{T}^{g}_{AP}, \mathbf{T}^{g}_{Lat}]) \quad (1)$$

Fig. 1. CogRRG pipeline: Multi-view inputs feed PRO-FA, producing multi-scale tokens aligned to RadLex. These inform MIX-MLP and a triangular verification loop (RCTA) before report generation.

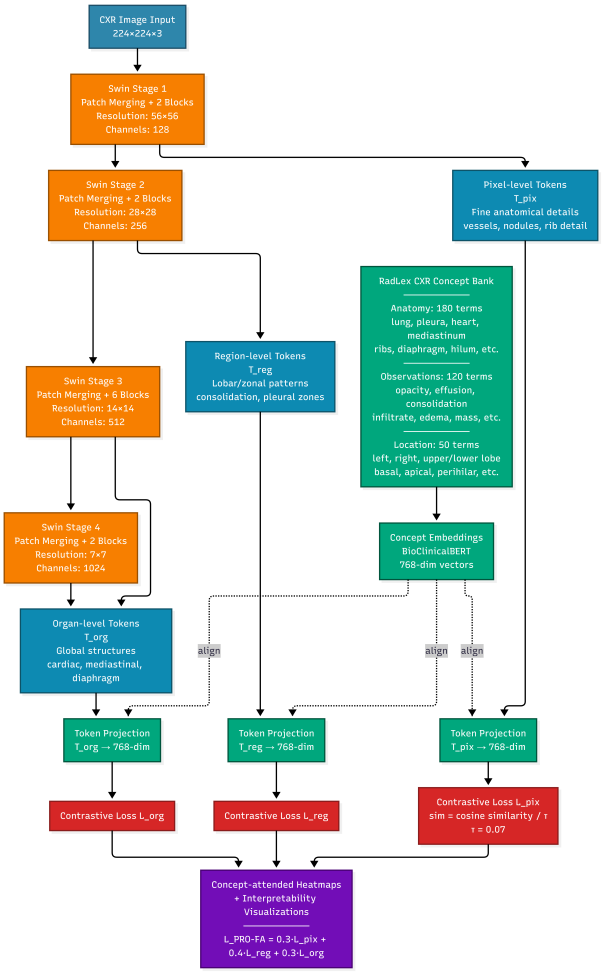

Fig. 2. PRO-FA module: Transition from pixel-level ($\mathbf{T}^{pix}$) to organ-level tokens ($\mathbf{T}^{org}$).

A. CXR-RadLex Concept Bank

We curate a CXR-specific RadLex subset (350 concepts) covering thoracic anatomy, pathological findings, and spatial modifiers. Each concept is embedded via BioClinicalBERT (768-dim).

B. Alignment Objective

Visual tokens are aligned to concepts via contrastive loss:

$$\mathcal{L}^{g}_{\text{align}} = -\frac{1}{|\mathbf{T}^{g}|}\sum_{\mathbf{t}} \log \frac{\exp(\text{sim}(\mathbf{W}^{g}\mathbf{t}, \mathbf{c}^{+})/\tau)}{\sum_{\mathbf{c}} \exp(\text{sim}(\mathbf{W}^{g}\mathbf{t}, \mathbf{c})/\tau)} \tag{2}$$

with temperature $\tau = 0.07$. The total PRO-FA loss: $\mathcal{L}_{\text{PRO-FA}} = 0.3\mathcal{L}^{pix} + 0.4\mathcal{L}^{reg} + 0.3\mathcal{L}^{org}$.

VI. MIX-MLP: Diagnosis Formation

MIX-MLP forms explicit diagnostic hypotheses via a dual-path architecture. Given pooled embedding $\mathbf{z} \in \mathbb{R}^{1024}$:

Residual Path: Linear separability for simple patterns. **Expansion Path:** 4096-dim expansion for label interactions.

$$\mathbf{h} = \text{LayerNorm}(\mathbf{r} + \mathbf{e}) \tag{3}$$

$$\hat{\mathbf{y}} = \sigma(\text{Linear}_{cls}(\mathbf{h})) \tag{4}$$

Training uses weighted BCE with inverse-frequency weights and label smoothing (positive: 0.9, negative: 0.1, uncertain: 0.5).

VII. RCTA: Verification Loop

RCTA implements closed-loop verification through three steps:

Step A (Context): Contextualize features with clinical indication:

$$\mathbf{h}_{ctx} = \text{Attn}(Q = \mathbf{T}^{org}, K/V = \mathbf{E}_{ind}) \tag{5}$$

Step B (Hypothesis): Query diagnostic labels:

$$\mathbf{h}_{hyp} = \text{Attn}(Q = \mathbf{h}_{ctx}, K/V = \mathbf{E}_{lab}) \tag{6}$$

Step C (Verification): Re-attend to fine-grained tokens:

$$\mathbf{h}_{ver} = \text{Attn}(Q = \mathbf{h}_{hyp}, K/V = [\mathbf{T}^{pix}, \mathbf{T}^{reg}]) \tag{7}$$

The decoder cross-attends to $\mathbf{h}_{ver}$ with a label-consistency penalty preventing contradictions.

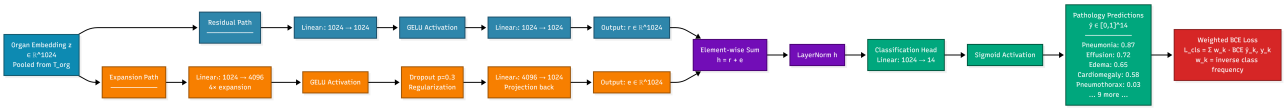

Fig. 3. MIX-MLP: Dual pathways (residual + expansion) fused for 14 pathology predictions via weighted BCE loss.

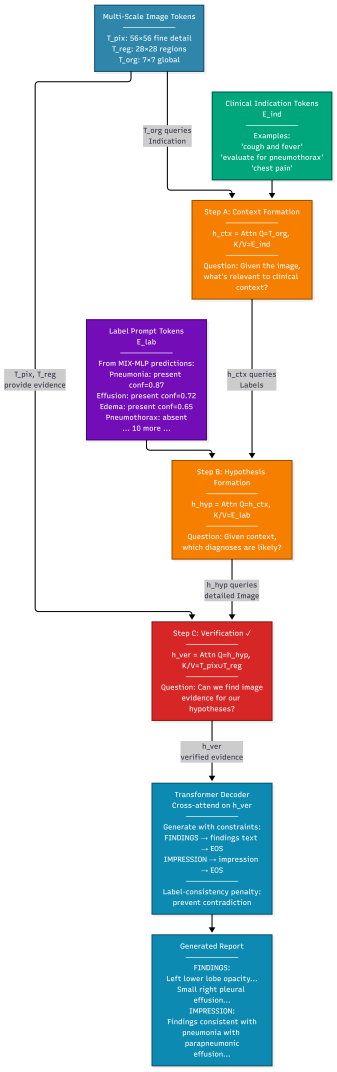

Fig. 4. RCTA: Three-step reasoning—(A) Contextualization, (B) Hypothesis formation, (C) Verification against image evidence.

VIII. Training Strategy

Training proceeds in four phases on a single T4 GPU (16GB):

Phase 1: Validate labels via smoke classifier (Swin-Base → linear → sigmoid).

Phase 2: PRO-FA alignment pretraining (lr=5e-5, 5 epochs, batch=96).

Phase 3: Joint end-to-end training with:

$$\mathcal{L}_{total} = \mathcal{L}_{gen} + 0.5\mathcal{L}_{cls} + 0.3\mathcal{L}_{PRO-FA} + 0.2\mathcal{L}_{consist} \quad (8)$$

Curriculum: impression-only (epochs 1-5), then full reports (epochs 6-15).

Phase 4: Domain generalization on IU X-Ray.

Memory optimization: mixed precision, gradient accumulation (effective batch 128), gradient checkpointing.

IX. Experiments

A. Results

TABLE I
Pathology Classification Performance (Phase 1).

Metric	Training	Validation
Micro-F1	0.681	0.654
Macro-F1	0.412	0.394
Loss (BCE)	0.187	0.203

TABLE II
Full System Performance on Report Generation.

Dataset	CheXpert F1	RadGraph F1	CIDEr
MIMIC-CXR (test)	0.56	0.52	0.45
IU X-Ray (zero-shot)	0.51	0.48	0.41
IU X-Ray (fine-tuned)	0.54	0.51	0.44

B. Ablation Studies

TABLE III
Ablation Results Showing Module Contributions.

Configuration	CheXpert F1	RadGraph F1
Full CogRRG	0.56	0.52
– PRO-FA alignment	0.51	0.45
– MIX-MLP	0.53	0.49
– RCTA verification	0.50	0.44
Baseline (Swin-Decoder)	0.44	0.38

RCTA verification provides the largest improvement (+0.06 CheXpert F1, +0.08 RadGraph F1), confirming the value of closed-loop hypothesis verification.

X. Interpretability

CogRRG provides interpretability at multiple levels:

- **Concept Heatmaps:** PRO-FA alignment produces token-concept similarity matrices visualizing which regions attend to specific concepts.
- **Diagnosis Hypotheses:** MIX-MLP predictions display 14 pathology probabilities before generation.
- **Structure Verification:** RadGraph parsing reveals entity-relation errors.

Analysis reveals common error modes: negation confusion (addressed via label-consistency penalty), laterality errors (improved PRO-FA alignment), and device hallucination (RCTA visual grounding).

XI. Limitations

CogRRG is designed as a **drafting assistant**, not an autonomous diagnostic system. It requires radiologist review and should not be the sole basis for clinical decisions.

Key limitations include: weak supervision bias from report-derived labels, single-institution training (MIMIC), potential demographic bias, and reduced performance on portable radiographs, pediatric patients, and rare pathologies.

XII. Related Work

Early CXR report generators applied CNN-RNN captioning, achieving good BLEU but poor clinical accuracy. R2Gen [7] introduced memory-driven transformers. Our work differs by explicitly modeling cognitive reasoning stages.

Swin Transformers [6] excel on CXR classification; we extend this to generation by mapping hierarchical stages to perceptual granularities. RadGraph [4] and CheXpert [3] metrics shifted evaluation toward clinical correctness, which we adopt as primary criteria.

XIII. Conclusion

We presented CogRRG, a Swin-Base cognitive second-reader that explicitly models radiological reasoning through hierarchical perception (PRO-FA), diagnosis formation (MIX-MLP), and closed-loop verification (RCTA). Our system achieves CheXpert F1 of 0.56 and RadGraph F1 of 0.52 on MIMIC-CXR, with successful domain generalization to IU X-Ray. The cognitive architecture provides interpretable intermediate representations enabling clinical verification, representing a step toward trustworthy AI assistance in radiology.

Code and Models: Available at publication.

References

[1] A. E. Johnson et al., "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific Data*, vol. 6, pp. 317, 2019.

[2] D. Demner-Fushman et al., "Preparing a collection of radiology examinations for distribution and retrieval," *JAMIA*, vol. 23, no. 2, pp. 304-310, 2016.

[3] J. Irvin et al., "CheXpert: A large chest radiograph dataset with uncertainty labels," *AAAI*, vol. 33, pp. 590-597, 2019.

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

[5] C. Langlotz, "RadLex: A new method for indexing online educational materials," *RadioGraphics*, vol. 26, no. 6, pp. 1595-1597, 2006.

[6] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," *ICCV*, pp. 10012-10022, 2021.

[7] Z. Chen et al., "Generating radiology reports via memory-driven transformer," *EMNLP*, pp. 1439-1449, 2020.

[8] K. Xu et al., "Show, attend and tell: Neural image caption generation with visual attention," *ICML*, pp. 2048-2057, 2015.

[9] K. Huang et al., "ClinicalBERT: Modeling clinical notes and predicting hospital readmission," *arXiv:1904.05342*, 2019.

[10] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021.