LOGIC-DRIVEN CONCEPT BOTTLENECK FOR INTERPRETABLE RADIOLOGY REPORT GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Automated radiology report generation helps alleviate radiologists' workload and enhance diagnostic accuracy. However, current deep learning-based radiology report generation models often lack interpretability and logical consistency. The semantically agnostic nature of their reasoning processes often fails to align visual findings with clinical diagnoses, leading to factual inconsistencies. To address these challenges, we propose LoCoGen, a novel model for radiology report generation, which leverages a logic-driven concept bottleneck to perform diagnostic inference without relying on ground-truth annotations for the intermediate clinical concepts. Specifically, LoCoGen learns to predict intermediate concepts by optimizing high-level disease classification through logic-driven formulas linking concepts and diseases. The resulting inference paths are then integrated with clinical context and visual features as prompts to guide a pre-trained language model in generating accurate and interpretable reports. Extensive experiments on the MIMIC-CXR dataset demonstrate that our model achieves superior performance compared to state-of-the-art models. Qualitative analysis further validates the interpretability and diagnostic accuracy of LoCoGen. Code is available at `https://anonymous.4open.science/r/LoCoGen-zj8snwin2j0d/`.

1 INTRODUCTION

Radiological imaging is essential for disease assessment, yet manual reporting is a time-consuming task that demands significant clinical expertise (Tian & Fu, 2020; Brady, 2017). Consequently, automated radiology report generation (RRG) has garnered significant attention in recent years. With the rapid advancement of deep learning, many AI-driven models have been proposed to enhance the efficiency and accuracy of reporting (Chen et al., 2020; 2021; Qin & Song, 2022; Tanida et al., 2023; Gu et al., 2024; Wang et al., 2025a). However, as Figure 1 (a), prevalent RRG models typically adopt an 'image-in, report-out' end-to-end paradigm, which suffer from the 'black-box' nature of deep neural networks, lacking transparency in their decision-making processes. Furthermore, these models typically learn an implicit mapping from visual features to report text, leading to factual inconsistencies or hallucinations. When such errors occur within opaque processes, they lack a traceable rationale, which ultimately erodes clinical trust.

Recently, Concept Bottleneck Models (CBMs) (Koh et al., 2020) have emerged as a promising direction for interpretable medical image analysis. By decomposing predictions into two stages, mapping inputs to human-interpretable intermediate concepts and then inferring targets from these concepts, CBMs provide transparent reasoning process and allow radiologists to refine final outputs by correcting intermediate concepts. Despite their potential, CBMs are limited by the high cost of manual concept labeling for ground-truth annotations, as shown in Figure 1 (b), especially in medical domain (Zhao et al., 2025; Gao et al., 2025; Bunnell et al., 2024; Dong et al., 2025; Marcinkevičs et al., 2024). As illustrated in Figure 1 (c), current models (Xing et al., 2025) often use high-level disease labels (*e.g.*, "Pleural Effusion") as pseudo-concept, which fail to capture the granular pathological signs (*e.g.*, "Blunting of Costophrenic Angle"). It leads to circular or vacuous reasoning, where the model's explanation provides no substantive evidence beyond the diagnosis itself.

In this paper, we propose **LoCoGen**, a novel model designed for faithful and interpretable report generation via a Logic-driven Concept Bottleneck (LCB). LoCoGen is able to produce intermediate clinical concepts and explicit inference paths without requiring concept-level annotations. Instead,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

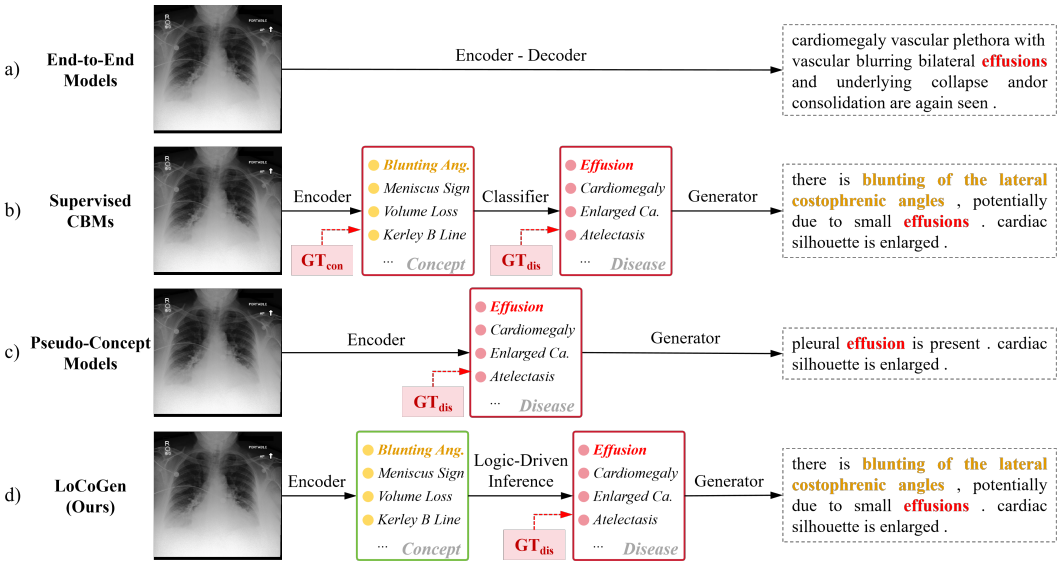

Figure 1: Comparison of RRG models. Our LoCoGen enables interpretable reasoning without extra concept-level annotations.

it relies solely on predefined concepts and logical formulas that map these concepts to high-level disease labels. Specifically, an image encoder first extracts region-level visual features, which are processed by the LCB to predict intermediate concepts and infer corresponding diseases. Unlike prior end-to-end models solely relying on opaque embeddings that may implicitly encode entangled non-semantic information, we utilize transparent logical formulas to derive diagnoses. This ensures the model learns explicit and clinically relevant reasoning logic. Subsequently, these inference paths are converted into natural language prompts. Combined with visual features and clinical context, these prompts guide a Pre-trained Language Model (PLM) in generating the final report.

The main contributions of our work are as follows:

- We propose LoCoGen, a logic-driven concept bottleneck framework that mimics human radiologists' diagnostic reasoning from clinical signs to diseases, constructing explicit inference paths to generate accurate and interpretable radiology reports.
- LoCoGen enables weakly supervised concept learning by converting logical formulas into differentiable soft logic that backpropagates disease-level supervision to concepts, eliminating the need for concept annotations.
- Extensive experiments on the MIMIC-CXR dataset demonstrate that LoCoGen achieves state-of-the-art (SOTA) performance while providing interpretability.

2 Related Work

2.1 Radiology Report Generation

RRG originates from image captioning in computer vision, and is primarily built upon encoder-decoder architectures. A line of methods improve cross-modal representation learning through memory-augmented architectures (Chen et al., 2020; 2021), reinforcement learning-based optimization (Qin & Song, 2022), and cross-modal alignment pretraining (Yu et al., 2025; Liu et al., 2024; Yin et al., 2025), with the goal of enhancing report generation. Subsequently, some methods explore region-aware modeling by explicitly incorporating lesion localization and anatomical structures. Methods such as RGRG (Tanida et al., 2023), COMG (Gu et al., 2024), and HKRG (Wang et al., 2025a) improve report generation by aligning visual features with organs, regions, and pathological areas, thereby strengthening fine-grained clinical descriptions.

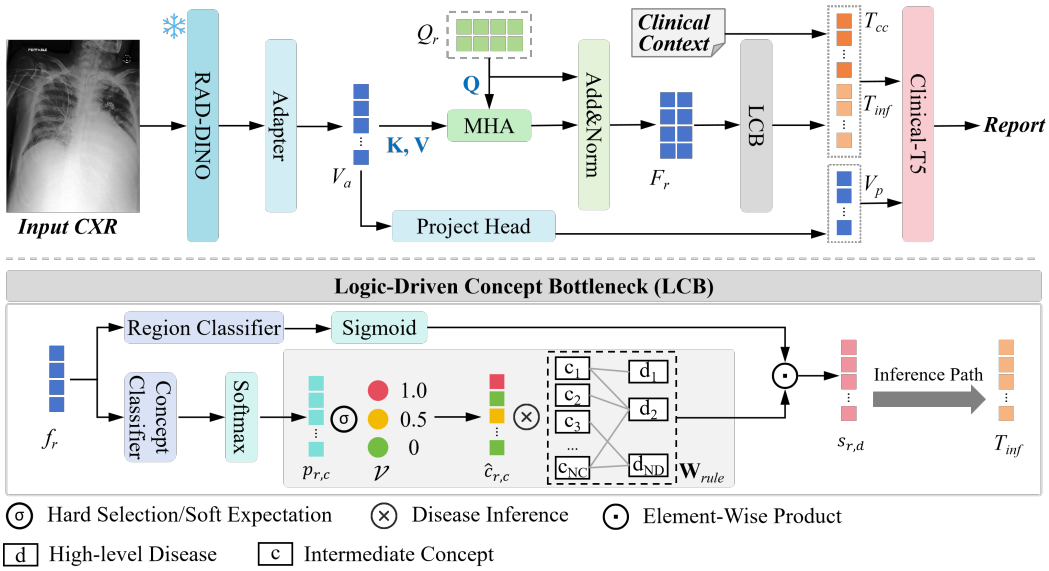

Figure 2: The framework of LoCoGen.

Despite these advances, interpretability remains limited. Although methods such as MCA-RG (Xing et al., 2025) introduce concept-based reasoning, their label-correlated concepts often result in "predicting results from results" and fail to reveal genuine diagnostic processes.

2.2 Concept-Based Diagnosis

The black-box nature of deep learning models limits their interpretability and clinical trustworthiness, motivating the adoption of concept-based interpretable models such as CBMs in medical domains. CGP (Zhao et al., 2025) and CRL (Gao et al., 2025) incorporate concept bottlenecks into dermoscopic, histopathological, and breast ultrasound image classification, achieving competitive predictive performance while providing concept-based explanations. RadAlign (Gu et al., 2025) further leverages medical concepts to align lesion localization with diagnostic decisions. However, these methods rely on manually annotated intermediate concepts, which are costly and difficult to obtain in clinical settings. To reduce this dependency, methods such as LaBo (Yang et al., 2023), Explicd (Gao et al., 2024), and MVP-CBM (Wang et al., 2025b) exploit Large Language Models (LLMs) or expert knowledge to construct diagnostic concepts. Nevertheless, they still require explicit concept-level supervision or additional concept alignment objectives, limiting their applicability in scenarios without concept annotations. In contrast, we explore a logic-driven concept bottleneck that transforms clinical rules into differentiable logical paths, allowing disease-level supervision to constrain and optimize intermediate concept representations.

3 Method

3.1 Visual Feature Extraction

As illustrated in Figure 2, given a radiology image, we first utilize a pre-trained visual encoder backbone to extract visual features. To adapt the visual features to RRG task without losing the pre-trained knowledge, we freeze the original parameters of the backbone and introduce a lightweight trainable adapter. Let $\mathbf{H} \in \mathbb{R}^{L \times D}$ represent the feature sequence from the frozen backbone, with L and D denoting the sequence length and hidden dimension respectively. The adapter is instantiated as a multi-layer perceptron (MLP) consisting of two linear projection layers, with a GELU activation and dropout regularization applied in between. The adapted visual features $\mathbf{V_a}$ are computed via a residual connection:

$$\mathbf{V_a} = \mathrm{LN}(\mathbf{H} + \mathrm{MLP}(\mathrm{LN}(\mathbf{H}))), \tag{1}$$

where $\mathrm{LN}(\cdot)$ is layer normalization. In contrast to models that entangle visual information from the entire image, which usually obscure the spatial specificity required for precise localization, we utilize a set of learnable region queries $\mathbf{Q_r} \in \mathbb{R}^{N_R \times D}$ to explicitly capture features from distinct anatomical regions and align with region-specific clinical concepts. Here, N_R denotes the number of anatomical regions. We employ a cross-attention mechanism to aggregate these region-specific visual features from the global context:

$$\mathbf{F}_r = \mathrm{LN}(\mathbf{Q_r} + \mathrm{MHA}(\mathbf{Q_r}, \mathbf{V_a}, \mathbf{V_a})). \tag{2}$$

Let $\mathbf{f}_r \in \mathbb{R}^D$ denote the extracted feature for the r-th region.

3.2 Logic-Driven Concept Bottleneck

To bridge the gap between visual features and high-level disease diagnosis via explicit clinical concepts without relying on extra annotations, we propose the LCB module. This module utilizes logic rules to infer disease presence from regional concepts. Crucially, we implement the clinical rules as differentiable logic gates, establishing a symbolic pathway from intermediate clincial concepts to high-level disease. This formulation allows gradients from high-level disease supervision to back-propagate through the logic rules, thereby implicitly refining the parameters for concept prediction.

We first assess the pathological status of each anatomical region. To compute the region-level abnormality probability p_r for the r-th region, we utilize a binary classifier composed of a linear layer and a sigmoid activation:

$$p_r = \mathrm{Sigmoid}(\mathbf{f}_r \mathbf{W}_{region} + \mathbf{b}_{region}), \tag{3}$$

where $\mathbf{W}_{region}$ and $\mathbf{b}_{region}$ are learnable parameters of the region classifier.

For each region, we predict a set of intermediate clinical concepts. To capture the granularity of clinical findings, each concept is modeled as a discrete variable with three states: *absent*, *uncertain/mild*, and *present*, mapped to a scalar set $\mathcal{V} = \{0, 0.5, 1.0\}$. The concept classifier predicts a probability distribution for the c-th concept in the r-th region:

$$p_{r,c} = [p_{r,c}^{(0)}, p_{r,c}^{(0.5)}, p_{r,c}^{(1.0)}] = \mathrm{Softmax}(\mathbf{f}_r \mathbf{W}_{concept} + \mathbf{b}_{concept}), \tag{4}$$

where $\mathbf{W}_{concept}$ and $\mathbf{b}_{concept}$ denote the learnable weights and bias of the concept classifier, respectively. To make the discrete states differentiable during training, we explicitly compute the expected activation value $\hat{c}_{r,c}$ via a weighted sum:

$$\hat{c}_{r,c} = \sum_{k \in \mathcal{V}} k \cdot p_{r,c}^{(k)}. \tag{5}$$

During inference, we discretize $\hat{c}_{r,c}$ by selecting the state with the maximum probability to obtain a deterministic clinical status.

To establish a transparent and clinically grounded reference path, we introduce a logic-driven inference mechanism that maps intermediate clinical concepts to final disease diagnoses. Instead of relying on black-box projections, we employ a fixed rule-mapping matrix $\mathbf{W}_{rule} \in \mathbb{R}^{N_C \times N_D}$ to encode the deterministic logical relationships between N_C clinical concepts and N_D diseases. To explicitly illustrate how concepts contribute to a specific disease, let $w_{c,d}$ denote the element at the c-th row and d-th column of $\mathbf{W}_{rule}$. This value strictly represents the predefined clinical correlation indicating how strongly the c-th concept supports the d-th disease. The raw diagnostic logit for disease d derived from region r is calculated as the weighted sum of all activated concepts in that region. To ensure spatial-clinical consistency and mitigate hallucinations, this diagnostic logit is strictly gated by the regional abnormality probability p_r. Formally, the final diagnostic score $s_{r,d}$ for disease d inferred from region r is formulated as:

$$s_{r,d} = (p_r + \mathbb{I}_{train} \cdot \epsilon) \left(\sum_{c=1}^{N_C} \hat{c}_{r,c} w_{c,d} \right) + b_d, \tag{6}$$

where b_d is a learnable bias for the d-th disease. During training ($\mathbb{I}_{train} = 1$), a small smoothing factor $\epsilon = 0.05$ is added to p_r to ensure a non-zero gradient flow, preventing the disease inference branch from vanishing when regions are initially undetected. During inference ($\mathbb{I}_{train} = 0$), ϵ is removed. Since $\mathbf{W}_{rule}$ remains frozen, minimizing the disease-level loss effectively forces the model

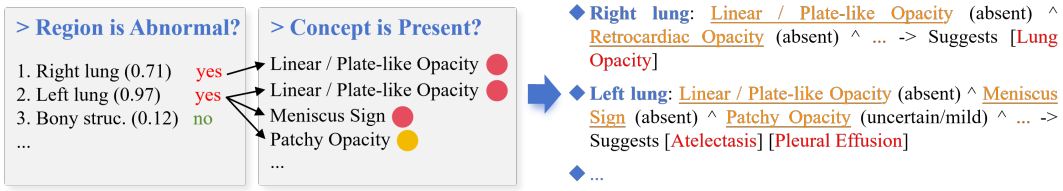

Figure 3: Illustration of the logic-driven inference path from regional clinical concepts to disease. Red circles indicate concepts with an *absent* state, while orange circles indicate concepts with an *uncertain/mild* state.

to learn accurate intermediate concept representations through back-propagation. This mechanism enables LoCoGen to acquire concept-level supervision from image-level disease labels, significantly reducing the annotation burden while ensuring the inference path is mathematically traceable and interpretable.

Overall, we train the Visual Feature Extractor and our LCB module using a composite loss function $\mathcal{L}_{lcb}$, which combines two binary cross-entropy losses $\mathcal{L}_{reg}$ and $\mathcal{L}_{dis}$:

$$\begin{aligned}
\mathcal{L}_{lcb} &= \mathcal{L}_{reg} + \mathcal{L}_{dis}, \\
\mathcal{L}_{reg} &= -\frac{1}{N_R}\sum_{r=1}^{N_R}\left[y_r \log p_r + (1-y_r)\log(1-p_r)\right], \\
\mathcal{L}_{dis} &= -\frac{1}{N_R N_D}\sum_{r=1}^{N_R}\sum_{d=1}^{N_D}\Big[y_{r,d}\log\sigma(s_{r,d}) + (1-y_{r,d})\log(1-\sigma(s_{r,d}))\Big],
\end{aligned} \tag{7}$$

where $y_r, y_{r,d} \in \{0,1\}$ denote the binary ground-truth labels for region abnormality and region-level disease d, respectively. $\sigma(\cdot)$ is the sigmoid function.

3.3 Prompt-Guided Report Generation

To explicitly guide the generation process with the established diagnostic logic, we translate the numerical inference results into a natural language sequence. Specifically, as shown in Figure 3, for a targeted region r, we identify the key diagnostic evidence by retrieving the top-k contributing concepts. The contribution of each concept is determined by its activation magnitude in $\hat{\mathbf{c}}$ weighted by the corresponding logical rules in $\mathbf{W}_{rule}$. This derivation is then formatted into a structured text template for each region as inference path $\mathbf{T}_{inf}$:

$$\begin{aligned}
\mathcal{C}_r &= [\texttt{Concept}_i : \hat{c}_{r,i}]_{i=1}^{k}, \\
\mathcal{S}_r &= [\texttt{Disease}_j : \mathbf{s}_{r,j}]_{j\in\mathcal{D}_r}, \\
\mathbf{T}_{inf} &= \texttt{[Region]}(p_r), \texttt{Clues:}\mathcal{C}_r, \texttt{Suggests:}\mathcal{S}_r,
\end{aligned} \tag{8}$$

where p_r denotes the region abnormality probability, $\hat{c}_{r,i}$ represents the activation strength of the i-th concept, and $s_{r,j}$ is the diagnostic score for the j-th disease within the set of detected diseases $\mathcal{D}_r$ for the region r. Finally, the inference path $\mathbf{T}_{inf}$ is concatenated with the clinical context $\mathbf{T}_{cc}$ and the visual prompts $\mathbf{V}_p$ to form the comprehensive prompt that guides the PLM in generating the final report. $\mathbf{V}_p$ is projected from region-specific visual features $\mathbf{V}_a$ via a learnable head to align with the PLM's embedding space, thereby injecting fine-grained visual evidence and global context directly into the generation process. Simultaneously, $\mathbf{T}_{cc}$ guides the model to prioritize diagnostic information relevant to the patient's clinical background, mimicking the workflow of human radiologists to establish the linguistic logical framework required to organize the final report.

During the training of the report generator, we freeze the parameters of the visual feature extractor and LCB module. Given the target report $Y = \{y_1, y_2, \ldots, y_T\}$, the generation loss $\mathcal{L}_{gen}$ is defined via the cross-entropy criterion:

$$\mathcal{L}_{gen} = -\sum_{t=1}^{T}\log P(y_t | y_{<t}, \mathbf{V}_p, \mathbf{T}_{cc}, \mathbf{T}_{inf}). \tag{9}$$

4 EXPERIMENTS

4.1 SETUP

Datasets. To validate the performance of our proposed framework, we conduct extensive experiments on the MIMIC-CXR dataset (Johnson et al., 2019b;a), which serves as a large-scale benchmark containing 377,110 chest X-ray images and 227,835 associated reports. We utilize the official data splits for training, validation, and testing. Furthermore, We incorporate the corresponding regional annotations from Chest ImaGenome (Wu et al., 2021), including the attributes *anatomicalfinding*, *disease*, *tubesandlines*, and *devices*. We map the information contained in these attributes to the 14 observation categories defined by CheXpert Labeler (Irvin et al., 2019) to obtain region-level ground-truth labels, which are subsequently used to supervise LCB training.

Metrics. To provide a comprehensive evaluation, we adopt both Natural Language Generation (NLG) and Clinical Efficacy (CE) metrics. In NLG metrics, we report BLEU-n (BL-n) (Papineni et al., 2002), METEOR (MTR) (Banerjee & Lavie, 2005), and ROUGE-L (RG-L) (Chin-Yew, 2004) to measure the textual similarity and fluency relative to ground-truth reports. In CE metrics, we leverage CheXbert (Smit et al., 2020) to classify each report into 14 pre-defined clinical observations, subsequently calculating the precision (P), recall (R), and F1-score based on these labels to quantify diagnostic consistency. Furthermore, we use RadGraph (Jain et al., 2021) to evaluate the structural consistency of the reports by matching medical entities and their inter-relational dependencies in a graph-based format.

Implementation Details. LoCoGen is configured to predict $N_R = 8$ anatomical regions, $N_C = 37$ intermediate clinical concepts, and $N_D = 13$ high-level diseases. The rule-mapping matrix $\mathbf{W}_{rule}$ is manually established according to domain-specific clinical priors. During the construction of the inference path, we set $k = 3$ to select the contributing concepts. To evaluate LoCoGen across different architectures, we instantiate it using both generic and domain-specific backbones. For visual encoding, we employ ViT-B/16 and RAD-DINO (Perez-Garcia et al., 2025). For language generation, we adopt a standard 6-layer Transformer decoder and Clinical T5 (Lu et al., 2022). Our model is trained in two stages, which involves 5 epochs with a learning rate of 1e-4 and a batch size of 32 in stage 1, and 5 epochs with the same learning rate but a batch size of 2 in stage 2. During inference, the beam search size is set to 3. All the decision threshold for classification is set to 0.5. We conduct our experiments using a single NVIDIA GeForce RTX 3090 GPU.

4.2 COMPARISON WITH SOTA MODELS

We compare LoCoGen with recent report generation methods, including end-to-end models (*e.g.*, AHP), concept-based baselines (*e.g.*, MCA-RG), region-aware baselines (*e.g.*, RGRG) and other SOTA baselines. Note that traditional CBMs are excluded as they strictly require dense human-annotated concept labels, which are unavailable in the MIMIC-CXR dataset.

NLG Results. Table 1 presents the NLG performance of LoCoGen compared with SOTA baselines on the MIMIC-CXR dataset. Our complete model (using RAD-DINO + Clinical T5) achieves new SOTA results across almost all metrics, notably outperforming previous SOTA baselines and advancing BLEU-4 score from 0.143 to 0.183, METEOR from 0.168 to 0.241, and ROUGE-L score from 0.310 to 0.377. Although D2Gen obtains a higher BLEU-1 score, LoCoGen's substantial lead in higher-order n-gram and semantic metrics (*e.g.*, BLEU-4 and ROUGE-L) demonstrates its superiority in generating coherent long-span clinical descriptions. Furthermore, while pseudo-concept-based methods like MCA-RG attempt to improve the transparency during generation, their disease-to-disease shortcuts lead to an opaque process of "predicting results from results", offering limited guidance for report generation. In contrast, LoCoGen mimics human radiologists by constructing explicit sign-to-disease inference paths, providing structured guidance that outperforms MCA-RG across all NLG metrics.

CE Results. Table 1 summarizes the factual correctness of generated clinical findings. Compared with our RAD-DINO + Clinical T5 backbone baseline, LoCoGen consistently improves all CE metrics, increasing F1 from 0.451 to 0.501, precision from 0.554 to 0.573, and recall from 0.381 to 0.446. These gains demonstrate the effectiveness of the proposed clinical reasoning framework beyond the underlying visual and language backbones. We further compare LoCoGen with existing methods that incorporate regional supervision or clinical concept prediction. RGRG also

Table 1: Comparison of NLG and CE performance on the MIMIC-CXR dataset. Best results are in bold, while the second-best are underlined.

Methods	NLG Metrics						CE Metrics			
	BL-1	BL-2	BL-3	BL-4	MTR	RG-L	F1	P	R	RG
R2Gen (Chen et al., 2020)	0.353	0.218	0.145	0.103	0.142	0.277	0.276	0.333	0.273	0.211
CMN (Chen et al., 2021)	0.353	0.218	0.148	0.106	0.142	0.278	0.278	0.334	0.275	0.223
CMM+RL (Qin & Song, 2022)	0.381	0.232	0.155	0.109	0.151	0.287	0.292	0.342	0.294	-
RGRG* (Tanida et al., 2023)	0.334	0.204	0.133	0.089	0.163	0.252	0.461	0.522	0.413	0.210
COMG (Gu et al., 2024)	0.363	0.235	0.167	0.124	0.128	0.290	0.345	0.424	0.291	-
SEI (Liu et al., 2024)	-	0.238	-	0.128	0.154	0.296	-	0.460	-	0.249
MCA-RG (Xing et al., 2025)	0.411	0.256	0.175	0.128	0.163	0.300	0.408	0.472	0.406	-
HKRG (Wang et al., 2025a)	<u>0.417</u>	0.253	0.172	<u>0.143</u>	0.167	<u>0.310</u>	0.339	0.384	0.324	-
KIA (Yin et al., 2025)	0.415	<u>0.274</u>	<u>0.187</u>	0.138	0.167	0.307	0.461	0.504	0.425	-
AHP (Yu et al., 2025)	0.400	0.250	0.172	0.126	0.154	0.285	-	-	-	0.180
D2Gen (Wu et al., 2026)	**0.428**	0.272	<u>0.187</u>	0.131	-	0.232	-	-	-	-
ViT-B/16 + Transformers	0.263	0.163	0.116	0.089	0.143	0.274	0.309	0.412	0.248	0.239
+ LoCoGen (Ours)	0.313	0.195	0.138	0.104	0.158	0.279	0.401	0.500	0.334	0.250
Δ *Improvement*	+19.0%	+19.6%	+19.0%	+16.9%	+10.5%	+1.8%	+29.8%	+21.4%	+34.7%	+4.6%
ViT-B/16 + Clinical T5	0.257	0.166	0.122	0.095	0.148	0.285	0.324	0.432	0.259	0.258
+ LoCoGen (Ours)	0.281	0.186	0.138	0.108	0.161	0.300	0.424	0.516	0.360	0.279
Δ *Improvement*	+9.3%	+12.0%	+13.1%	+13.7%	+8.8%	+5.3%	+30.9%	+19.4%	+39.0%	+8.1%
RAD-DINO + Transformers	0.331	0.205	0.143	0.107	0.165	0.282	0.456	0.537	0.396	0.258
+ LoCoGen (Ours)	0.332	0.212	0.153	0.116	<u>0.174</u>	0.302	<u>0.494</u>	<u>0.564</u>	<u>0.439</u>	0.289
Δ *Improvement*	+0.3%	+3.4%	+7.0%	+8.4%	+5.5%	+7.1%	+8.3%	+5.0%	+10.9%	+12.0%
RAD-DINO + Clinical T5	0.304	0.203	0.151	0.118	0.171	0.308	0.451	0.554	0.381	<u>0.292</u>
+ LoCoGen (Ours)	0.399	**0.290**	**0.226**	**0.183**	**0.241**	**0.377**	**0.501**	**0.573**	**0.446**	**0.308**
Δ *Improvement*	+31.3%	+42.9%	+49.7%	+55.1%	+40.9%	+22.4%	+11.1%	+3.4%	+17.1%	+5.5%

exploits region-level supervision, using region bounding boxes together with corresponding region sentences. In comparison, LoCoGen achieves higher F1 (0.501 vs. 0.461), precision (0.573 vs. 0.522), and recall (0.446 vs. 0.413), suggesting that supervising fine-grained regional clinical findings provides effective intermediate guidance for clinically faithful report generation. MCA-RG similarly introduces clinical labels as pseudo-intermediate concepts, but these labels are already close to the final diagnostic outcomes. In contrast, LoCoGen decomposes the generation process into a hierarchical clinical reasoning chain, first identifying fine-grained clinical signs, then inferring higher-level diagnoses, and finally generating the report. This may contribute to the stronger CE performance, particularly in F1 and precision. For fine-grained clinical entity and relation evaluation, LoCoGen achieves a RadGraph score of 0.308, compared with 0.292 for the RAD-DINO + Clinical T5 backbone and 0.249 for SOTA baseline SEI, corresponding to relative improvements of 5.5% and 23.7%, respectively. This suggests that LoCoGen better preserves clinically meaningful relationships in the generated reports.

Effectiveness Across Diverse Backbones. To validate its robustness, we evaluate LoCoGen across four distinct backbone configurations. As shown in Tables 1, integrating our framework yields substantial improvements in both NLG and CE metrics across all settings. When equipped with domain-specific backbones (RAD-DINO + Clinical T5), the relative NLG gains are remarkably high (*e.g.*, +55.1% on BLEU-4 score, +40.9% on METEOR score). We attribute this synergy to two factors: (1) Since our intermediate concepts are reversely supervised by disease labels, RAD-DINO exhibits a much stronger perception of concept-level lesions than generic models; (2) Clinical T5 possesses richer prior knowledge of concept-disease relationships. Conversely, when applied to generic architectures, LoCoGen brings massive CE improvements ($\sim$30% relative gain in F1 score), effectively injecting crucial diagnostic capabilities into general-purpose models. In summary, these results indicate the versatility of our proposed framework. They suggest that integrating logic-driven diagnostic inference can effectively enhance both the linguistic fluency and clinical accuracy of generated reports across varying baseline architectures.

4.3 Ablation Study

To investigate the distinct contribution of each component in LoCoGen, we conduct a comprehensive ablation study. The results are detailed in Table 2.

Effect of the Inference Path (w/o T_{inf}). The exclusion of the explicit inference path causes a drastic drop across all metrics. Most notably, the RadGraph score plummets from 0.308 to 0.203,

Table 2: Ablation study on different modules.

Modules	NLG Metrics						CE Metrics			
	BL-1	BL-2	BL-3	BL-4	MTR	RG-L	F1	P	R	RG
w/o T_{inf}	0.337	0.209	0.141	0.102	0.179	0.288	0.451	0.528	0.393	0.203
w/o V_p	0.347	0.249	0.193	0.156	0.228	0.357	0.410	0.519	0.340	0.275
w/o T_{cc}	0.340	0.210	0.142	0.103	0.178	0.289	0.475	0.544	0.422	0.231
w/o Concept	0.322	0.215	0.160	0.124	0.178	0.315	0.475	0.562	0.411	0.303
w/o Region	0.328	0.220	0.163	0.128	0.180	0.314	0.481	0.566	0.419	0.302
Full	**0.399**	**0.290**	**0.226**	**0.183**	**0.241**	**0.377**	**0.501**	**0.573**	**0.446**	**0.308**



INDICATION: Gastrointestinal bleed, assessment for endotracheal tube position and interval change.

Groud-Truth: As compared to the previous radiograph, there is no relevant change. Monitoring and support devices are constant. Constant cardiomegaly with relatively extensive retrocardiac atelectasis and the potential presence of a small left pleural effusion. Mild pulmonary edema. Areas of atelectasis at the right lung base. No newly occurred parenchymal opacities. No pneumothorax.

Baseline: There is no pneumothorax. ~~There is no pulmonary edema.~~ Left lower lobe atelectasis is unchanged. Bibasilar opacities are consistent with atelectasis. Endotracheal tube tip is 3.5 cm above the carina. The nasogastric tube tip is in the stomach. Right internal jugular line tip is in the mid SVC. Moderate cardiomegaly. NG tube tip is in the stomach.

LoCoGen (Ours): As compared to the previous radiograph, the patient has been intubated. The tip of the endotracheal tube projects 4 cm above the carina. The course of the nasogastric tube is unremarkable, the tip of the tube projects over the middle parts of the stomach. There is no evidence of complications, notably no pneumothorax. Unchanged moderate cardiomegaly with bilateral areas of atelectasis and a small left pleural effusion. No newly appeared focal parenchymal opacities.

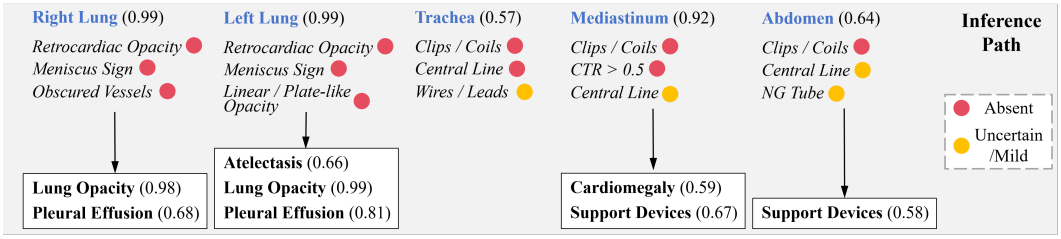

Figure 4: Qualitative comparison of reports generated by LoCoGen and the baseline. Distinct colors in reports indicate the matched descriptions. Strikethrough highlights incorrect descriptions of clinical findings.

and BLEU-4 score drops from 0.183 to 0.102. This decline confirms that inference paths not only provide critical diagnostic logic for factual correctness but also guide the report generator to generate coherent reports.

Effect of the Visual Features (w/o V_p). Without visual features, the model retains strong performance especially in METEOR score and ROUGE-L score, demonstrating that combining inference paths and clinical context is already capable of generating highly competitive reports. Nevertheless, integrating V_p further elevates the performance, pushing the F1 score from 0.410 to 0.501. This indicates that visual features serve as supplementary information, capturing fine-grained visual nuances not explicitly articulated in text prompts.

Effect of the Clinical Context (w/o T_{cc}). Interestingly, removing the clinical context leads to a massive drop in higher-order NLG metrics (BLEU-4 score drops to 0.103, METEOR score to 0.178) but only a slight decline in CE metrics. This contrast suggests that while T_{inf} and V_p ensure clinical keyword recall, T_{cc} governs the linguistic framework.

Effect of the Intermediate Concepts (w/o Concept**).** We ablate the intermediate clinical concepts from inference paths, forcing the model to predict diseases without fine-grained clinical signs. This removal leads to a notable decrease in both semantic fluency (BLEU-4 score drops to 0.124) and clinical accuracy (F1 score drops to 0.475). This indicates that intermediate concepts provide a referenceable logic foundation, which grounds the model's reasoning during report generation.

Impact of Region-Level Abnormality Detection (w/o Region**).** We analyze the role of region-level abnormality classification, which determines whether specific anatomical regions contain lesions.

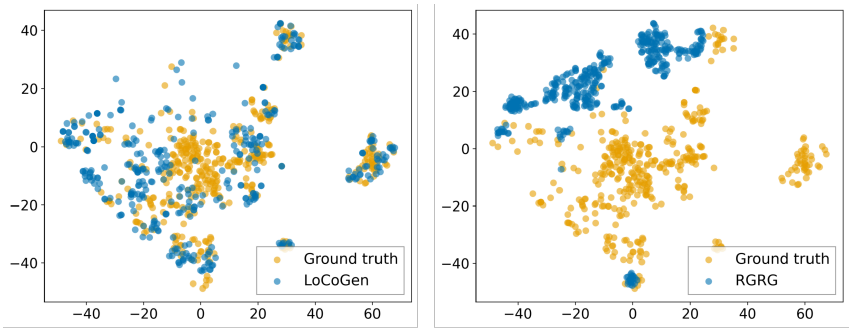

Figure 5: Visualization of feature distributions using t-SNE for the RGRG (baseline) and LoCoGen (Ours) models. Points are colored by source (Orange: ground truth; Blue: model output).

Removing this module leads to a substantial decline in NLG performance (BLEU-4 score drops from 0.183 to 0.128) and a decrease in CE F1 score (from 0.501 to 0.481). This indicates that region-level classification provides spatial anatomical priors. By isolating abnormal regions beforehand, it effectively directs subsequent concept predictions toward clinically relevant anatomical structures, preventing the model from being distracted by normal regions.

4.4 Qualitative Analysis

Case Study. As illustrated in Figure 4, LoCoGen demonstrates superior logical consistency and interpretability compared to the baseline (RGRG). By integrating clinical context, LoCoGen demonstrated superior context awareness by prioritizing the explicitly requested "assessment for endotracheal tube position." in the clinical context. While both models failed to report "Edema", LoCoGen detected a strong signal for the intermediate concept "Obscured Vessels", providing a relevant diagnostic clue. Additionally, LoCoGen captured a small "Effusion" overlooked by baseline. These results confirm that LoCoGen generates more accurate and interpretable reports by utilizing LCB to construct an explicit inference path.

Semantic Representation Visualization. Figure 5 visualizes the semantic representations of ground-truth and generated reports. Specifically, we first encode all reports with the same frozen CXR-BERT (Boecking et al., 2022) encoder and then jointly project the resulting representations into a two-dimensional space using t-SNE. The representations of LoCoGen-generated reports exhibit substantial overlap with the ground-truth distribution and cover a broad range of the reference semantic space. In contrast, RGRG predictions tend to concentrate in several relatively compact regions with less correspondence to the ground-truth distribution. These results suggest that LoCoGen better preserves the semantic distribution and diversity of clinical reports, indicating that its logic-driven intermediate concept reasoning provides more informative and case-specific guidance for report generation.

5 Conclusion

In this paper, we propose LoCoGen, a novel framework designed to generate clinically accurate and interpretable radiology reports by mimicking the diagnostic reasoning process of radiologists. LoCoGen constructs traceable inference paths by predicting intermediate clinical concepts and calculating disease probabilities via logic-driven formulas in the LCB module. Crucially, the logic-driven inference allows concept-level representations to be optimized implicitly via disease-level supervision, eliminating the need for dense concept annotations. These inference paths effectively condition the report generator to produce more accurate reports, while simultaneously providing explicit diagnostic interpretability. In future work, we plan to incorporate longitudinal prior studies to capture temporal disease progression, as well as leverage complementary spatial information from multi-view chest X-rays to enhance lesion localization and report generation.

AI USE STATEMENT

In this work, we used generative AI tools for design or provide feedback on research methodology or experiments, and assist with translation. We have not used generative AI tools to generate synthetic datasets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, implement methods, clean and reformat datasets, support qualitative and thematic data analysis, or interpret results, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for sourcing/searching for information, identifying relevant literature, drafting parts of the research paper, and editing the research paper to improve readability. We have reviewed all AI-assisted work. AI-assisted suggestions for research methodology and experiments were critically evaluated by the authors and verified through the corresponding literature, implementations, and experiments where applicable. AI-assisted manuscript text was reviewed and revised by the authors to ensure its accuracy and consistency with the actual work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide sufficient details to facilitate the reproducibility of our work. The proposed method and its major components are described in Section 3, together with the corresponding formulations and training objectives. Implementation and training details are provided in Section 4.1. The source code is available at `https://anonymous.4open.science/r/LoCoGen-zj8snwin2j0d/`.

REFERENCES

Satanjeev Banerjee and Alon Lavie. 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.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pp. 1–21. Springer, 2022.

Adrian P Brady. Error and discrepancy in radiology: inevitable or avoidable? *Insights into Imaging*, 8:171–182, 2017.

Arianna Bunnell, Yannik Glaser, Dustin Valdez, Thomas Wolfgruber, Aleen Altamirano, Carol Zamora González, Brenda Y. Hernandez, Peter Sadowski, and John A Shepherd. Learning a clinically-relevant concept bottleneck for lesion detection in breast ultrasound. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pp. 650–659, 2024.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1439–1449, November 2020.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. Cross-modal memory networks for radiology report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5904–5914, August 2021.

Lin Chin-Yew. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Yiheng Dong, Yi Lin, and Xin Yang. CoPA: Hierarchical concept prompting and aggregating network for explainable diagnosis. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, pp. 67–76, 2025.

Yibo Gao, Hangqi Zhou, Zheyao Gao, Bomin Wang, Shangqi Gao, Sihan Wang, and Xiahai Zhuang. Learning concept-driven logical rules for interpretable and generalizable medical image classification. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, pp. 291–300, 2025.

Yunhe Gao, Difei Gu, Mu Zhou, and Dimitris Metaxas. Aligning human knowledge with visual concepts towards explainable medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 46–56, 2024.

Difei Gu, Yunhe Gao, Yang Zhou, Mu Zhou, and Dimitris Metaxas. Radalign: Advancing radiology report generation with vision-language concept alignment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 484–494. Springer, 2025.

Tiancheng Gu, Dongnan Liu, Zhiyuan Li, and Weidong Cai. Complex organ mask guided radiology report generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7995–8004, January 2024.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Thirty-Third AAAI Conference on Artificial Intelligence*, 2019.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Q. Truong, Minh Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, and Pranav Rajpurkar. Radgraph: Extracting clinical entities and relations from radiology reports. In *In Advances in Neural Information Processing Systems (NeurIPS 2021)*, 2021.

Alistair Johnson, Tom Pollard, Seth Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, 2019a.

Alistair Johnson, Tom Pollard, Roger Mark, Seth J. Berkowitz, and Steven Horng. MIMIC-CXR database, 2019b. (version 2.0.0) *PhysioNet*. RRID:SCR_007345.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 5338–5348, July 2020.

Kang Liu, Zhuoqi Ma, Xiaolu Kang, Zhusi Zhong, Zhicheng Jiao, Grayson Baird, Harrison Bai, and Qiguang Miao. Structural entities extraction and patient indications incorporation for chest X-ray report generation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pp. 433–443, 2024.

Qiuhao Lu, Dejing Dou, and Thien Nguyen. ClinicalT5: A generative language model for clinical text. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 5436–5443, December 2022.

Ričards Marcinkevičs, Patricia Reis Wolfertstetter, Ugne Klimiene, Kieran Chin-Cheong, Alyssia Paschke, Julia Zerres, Markus Denzinger, David Niederberger, Sven Wellmann, Ece Ozkan, Christian Knorr, and Julia E. Vogt. Interpretable and intervenable ultrasonography-based machine learning models for pediatric appendicitis. *Medical Image Analysis*, 91:103042, 2024.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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.

Fernando Perez-Garcia, Harshita Sharma, Sam Bond-Taylor, Kenza Bouzid, Valentina Salvatelli, Maximilian Ilse, Shruthi Bannur, Daniel C. Castro, Matthew Schuurmans, Stephanie L. Hyland, and Ayya Faisal. Exploring scalable medical image encoders beyond text supervision. *Nature Machine Intelligence*, 7(1):119–130, 2025.

Han Qin and Yan Song. Reinforced cross-modal alignment for radiology report generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 448–458, May 2022.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y. Ng, and Matthew P. Lungren. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1500–1519, 2020.

Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7433–7442, June 2023.

Yingjie Tian and Saiji Fu. A descriptive framework for the field of deep learning applications in medical images. *Knowledge-Based Systems*, 210:106445, 2020.

Bo Wang, Peihong Teng, Hongda Zhang, Feiyang Yang, Zeyu Wang, Xingcheng Yi, Tianyang Zhang, Chunhui Wang, Adriano Jose Tavares, and Hao Xu. HKRG: Hierarchical knowledge integration for radiology report generation. *Expert Systems with Applications*, 271:126622, 2025a.

Chunjiang Wang, Kun Zhang, Yandong Liu, Zhiyang He, Xiaodong Tao, and S Kevin Zhou. Mvp-cbm: multi-layer visual preference-enhanced concept bottleneck model for explainable medical image classification. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 529–537, 2025b.

Joy Wu, Nkechinyere Agu, Ismini Lourentzou, Arjun Sharma, Joseph A. Paguio, Jasper S. Yao, Edward C. Dee, William G. Mitchell, Satyananda Kashyap, Andrea Giovannini, Leo A. Celi, and Mehdi Moradi. Chest imagenome dataset, 2021. (version 2.0.0). *PhysioNet*. RRID:SCR_007345.

Puzhen Wu, Hexin Dong, Yi Lin, Yihao Ding, and Yifan Peng. A disease-aware dual-stage framework for chest x-ray report generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 40, pp. 33953–33961, 2026.

Qilong Xing, Zikai Song, Youjia Zhang, Na Feng, Junqing Yu, and Wei Yang. MCA-RG: Enhancing LLMs with medical concept alignment for radiology report generation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, pp. 380–390, 2025.

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19187–19197, 2023.

Heng Yin, Shanlin Zhou, Pandong Wang, Zirui Wu, and Yongtao Hao. KIA: Knowledge-guided implicit vision-language alignment for chest X-ray report generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 4096–4108, January 2025.

Ting Yu, Wangwen Lu, Yan Yang, Weidong Han, Qingming Huang, Jun Yu, and Ke Zhang. Adapter-enhanced hierarchical cross-modal pre-training for lightweight medical report generation. *IEEE Journal of Biomedical and Health Informatics*, 29(7):5303–5316, 2025.

Lei Zhao, Changjian Chen, Bin Pu, Xiaoming Qi, Fengfeng Peng, Chunlian Wang, Kenli Li, and Guanghua Tan. Concept-induced graph perception model for interpretable diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 215–225. Springer, 2025.