GraphRAG-Rad: Concept-Aware Radiology Report Generation via Latent Visual-Semantic Retrieval

Faezeh Safari, Hang Dong, Zeyu Fu and Aline Villavicencio
Department of Computer Science, University of Exeter, United Kingdom
{fs525, H.Dong2, Z.Fu, A.Villavicencio}@exeter.ac.uk

Abstract

Radiology report generation involves translating visual signals from pixels into precise clinical language. Existing encoder-decoder models often suffer from hallucinations, generating plausible but incorrect medical findings. We propose GraphRAG-Rad, a novel architecture that integrates biomedical knowledge through a novel Latent Visual-Semantic Retrieval (VSR). Unlike traditional Retrieval-Augmented Generation (RAG) methods that rely on textual queries, our approach aligns visual embeddings with the latent space of the Knowledge Graph, PrimeKG. The retrieved sub-graph guides the Visual Encoder and the Multi-Hop Reasoning Module. The reasoning module simulates clinical deduction paths (`Ground-Glass Opacity → Viral Pneumonia → COVID-19`) before it combines the information with visual features in a Graph-Gated Cross-Modal Decoder. Experiments on the COV-CTR dataset demonstrate that GraphRAG-Rad achieves competitive performance with strong results across multiple metrics. Furthermore, ablation studies show that integrating latent retrieval and reasoning improves performance significantly compared to a visual-only baseline. Qualitative analysis further reveals interpretable attention maps. These maps explicitly link visual regions to symbolic medical concepts, effectively bridging the modality gap between vision and language.

1 Introduction

Automatic radiology report generation (ARRG) is a critical research area (Divya et al., 2024; Yang et al., 2023b) aiming to automate the labor-intensive task of documenting patient diagnoses (Zhao et al., 2024; Chen et al., 2025). These systems seek to improve diagnostic efficiency and consistency (Yang et al., 2023b; Zhao et al., 2024). While derived from image captioning (Zhao et al., 2024), ARRG is significantly more challenging (Divya et al., 2024; Yang et al., 2023b; Zhang and Jiang, 2024). Unlike natural image captioning, medical reports require high precision (Zhao et al., 2024) to differentiate fine-grained details in highly similar images (Divya et al., 2024; Zhang and Jiang, 2024), often focusing on specific abnormal regions rather than global descriptions (Yang et al., 2023b). Furthermore, reports must be long, narrative documents covering both normal and abnormal features (Divya et al., 2024; Zhao et al., 2024).

Current approaches struggle to capture spatial and semantic information effectively (Divya et al., 2024), often producing overly brief descriptions. They are also susceptible to visual-linguistic spurious correlations and data bias (Chen et al., 2025; Zhang et al., 2024b). This arises because abnormalities may occupy only small image regions (Tao et al., 2024) or fall into long-tail distributions, causing models to overlook rare but critical findings (Zhao et al., 2024). While modern methods employ attention and Transformer mechanisms to address this (Divya et al., 2024; Zhao et al., 2024), they often lack explicit reasoning capabilities.

Our approach, GraphRAG-Rad, is introduced as an explainable architecture designed to bridge the semantic gap between pixel-level visual features and structured medical knowledge. Specifically, this framework explicitly models clinical reasoning as a latent retrieval and traversal process, departing from traditional encoder-decoder models that rely solely on visual perception. Our results suggest that grounding visual features in structured biomedical knowledge is a decisive factor in mitigating clinical hallucinations. The substantial improvement in BLEU-4 scores compared to visual-only baselines (0.625 vs. 0.535) demonstrates that GraphRAG-Rad successfully grounds its output in relevant clinical context. By constraining the generation process with graph-based evidence, the model produces reports that adhere more closely to the specific language and content of the reference standards. Furthermore, our analysis shows

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 4: Student Research Workshop, pages 464–475

that the explicit modeling of reasoning paths is not merely redundant but essential; the ablation study demonstrates that removing the Multi-Hop Reasoning Module leads to a significant drop in performance, validating that the model relies on these symbolic deductive chains to construct coherent narratives. This justifies the architectural complexity of the Visual-Semantic Retrieval (VSR) system, as it effectively bridges the modality gap to provide the medical knowledge context that visual-only baselines lack. GraphRAG-Rad's knowledge-grounded approach enables the model to retrieve relevant biomedical concepts directly from chest CT images and use them to guide interpretable report generation.

2 Related Work

Automated radiology report generation (RRG) has evolved from early CNN-RNN encoder-decoder frameworks (Divya et al., 2024; Tao et al., 2024; Wu et al., 2023; Yang et al., 2023b; Zhang et al., 2024b) and Hierarchical Recurrent Networks (HRNNs) (Zhao et al., 2024) to sophisticated Transformer-based architectures like R2Gen and METransformer, which leverage memory and expert tokens to manage long-range dependencies (Divya et al., 2024; Zhao et al., 2024; Zhang et al., 2024a, 2023; Yan et al., 2023; Singh and Singh, 2025). To further enhance clinical accuracy, contemporary methods integrate structured medical knowledge via graphs, such as ATAG and PPKED, which map pathological entities and anatomical relationships from ontologies like RadLex into feature embeddings (Tao et al., 2024; Yang et al., 2023b; Zhang et al., 2023; Zhao et al., 2024; Zhang et al., 2024a; Yan et al., 2023). This integration allows models to capture intrinsic medical relationships, resulting in more detailed and consistent reports than traditional sequence-to-sequence approaches (Yang et al., 2023b; Zhang et al., 2024a; Yan et al., 2023; Zhang et al., 2023).

Memory-driven mechanisms have been widely adopted to manage the sequential complexity and significant length of medical reports (Divya et al., 2024; Zhao et al., 2024). These modules are integrated into Transformer encoders or decoders to learn relational information and consolidate cross-modal semantic alignment (Divya et al., 2024; Zhang et al., 2024a, 2023; Tao et al., 2024). For example, the Memory-based Cross-modal Semantic Alignment Model (MCSAM) utilizes a shared memory bank to align disease-related representations across different modalities (Tao et al., 2024). These memory-driven approaches help the model retain context over long passages and alleviate issues related to data bias in clinical datasets.

Retrieval-based methods address the tendency of generative models to produce 'hallucinated' or factually incorrect information by pulling templates or sentences from existing databases (Zhao et al., 2024; Zhang et al., 2023). Modern Retrieval-Augmented Generation (RAG) frameworks support Large Language Models (LLMs) by providing expert knowledge tailored to specific images through heuristic textual prompts (Fink et al., 2025; Yang et al., 2025). Systems like STREAM and Teaser use progressive semantic retrievers or topic separation to improve context-awareness and handle the long-tail distribution of rare medical cases (Yang et al., 2025; Zhao et al., 2024). This hybrid approach ensures that the output reflects standard physician reporting practices more closely than pure generation.

Finally, researchers have introduced specialized learning paradigms to mitigate the scarcity of labeled medical data and reduce visual-linguistic biases. Vision-Language Pre-training (VLP) models like MedViLL and REFERS learn joint representations from raw image-text pairs, bypassing the need for labor-intensive manual labeling (Zhang et al., 2023; Moon et al., 2022; Zhou et al., 2022). Semi-Supervised Learning (SSL) techniques like RAMT further reduce data reliance through consistency training (Zhang et al., 2023).

Inspired by previous work, our work aims to provide a knowledge graph retrieval and grounding for RRG. The learned path through multi-hop reasoning can provide a guidance to mitigate hallucination and enhance explainability. We use PrimeKG (Chandak et al., 2023; Weinreich et al., 2008) as a comprehensive multi-relational Knowledge Graph. Unlike prior work that uses fixed reasoning templates, our approach discovers reasoning paths dynamically through learned attention mechanisms over a comprehensive biomedical knowledge graph.

3 Methodology

As illustrated in Figure 1, the GraphRAG-Rad architecture functions as a sequential neuro-symbolic pipeline that transforms raw pixels into grounded clinical text through four integrated stages. First,

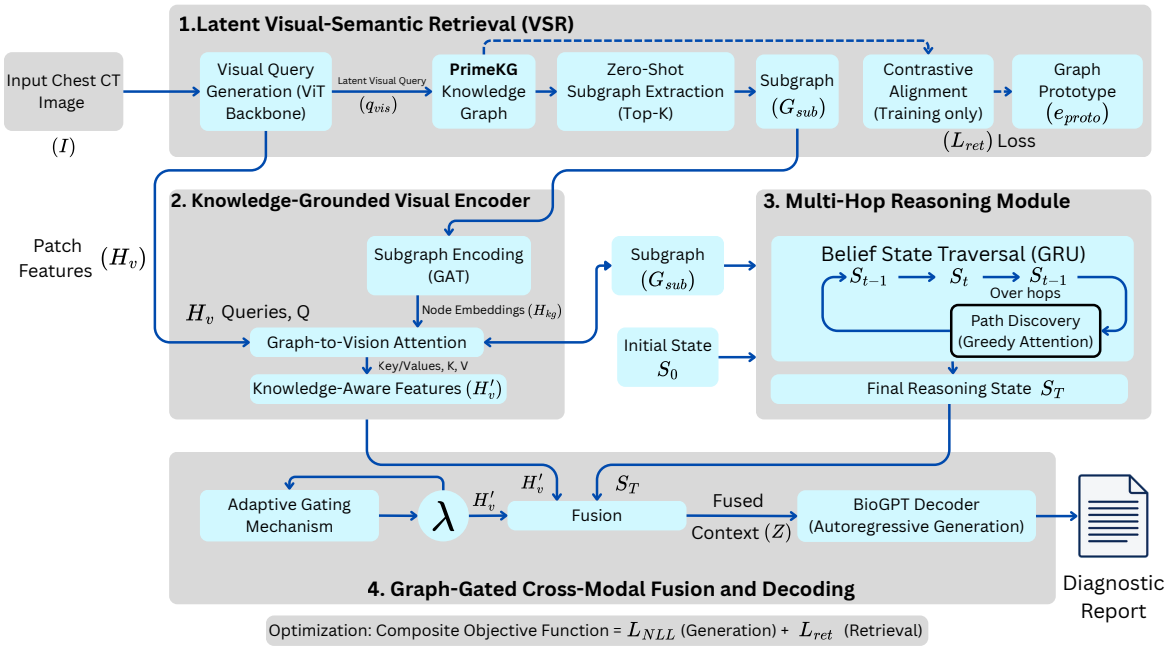

Figure 1: Overview of the proposed Knowledge-Grounded Medical Report Generation framework. The architecture proceeds in four stages: (1) Latent Visual-Semantic Retrieval (VSR): Input chest CT image I is processed by a ViT backbone to extract patch features H_v and a latent visual query q_{vis}, which retrieves a top-K subgraph G_{sub} from the PrimeKG knowledge graph. (2) Knowledge-Grounded Visual Encoder: A Graph Attention Network (GAT) encodes the subgraph into embeddings H_{kg}, which fuse with visual features via cross-attention to yield knowledge-aware representations H'_v. (3) Multi-Hop Reasoning: A recurrent module traverses the subgraph to simulate clinical deduction steps, producing a final symbolic reasoning state s_T. (4) Graph-Gated Fusion: An adaptive gate λ balances visual evidence (H'_v) and symbolic priors (s_T) to create a fused context Z for the BioGPT decoder.

the Latent Visual-Semantic Retrieval (VSR) mechanism bridges the modality gap by projecting global image features into a latent query vector (q_{vis}) to retrieve a relevant symbolic subgraph ($\mathcal{G}_{sub}$) from the PrimeKG knowledge base. This subgraph then drives the Knowledge-Grounded Encoding stage, prompting the Visual Encoder to focus attention on image regions corresponding to specific medical concepts to produce grounded visual features (H'_v). Simultaneously, a Symbolic Reasoning module traverses $\mathcal{G}_{sub}$ to simulate a multi-hop clinical deduction, yielding a final reasoning state (s_T). Finally, the Graph-Gated Decoding stage fuses this visual evidence (H'_v) with the symbolic reasoning vector (s_T) to dynamically guide a BioGPT decoder, ensuring the generated report is strictly anchored in both observed visual data and retrieved medical knowledge.

3.1 Problem Formulation

Let I denote a chest CT image and $Y = \{y_1, y_2, ..., y_T\}$ be the target diagnostic report. We assume access to a biomedical knowledge graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ represents entities (that include diseases, anatomies, phenotypes, etc.) and $\mathcal{E}$ represents semantic relations. Here we use PrimeKG for its high coverage of the medical entities. Our objective is to approximate the posterior $P(Y|I, \mathcal{G}_{sub})$, where $\mathcal{G}_{sub} \subset \mathcal{G}$ is a context-specific subgraph. A key challenge in multimodal RAG is the cross-modal retrieval problem: retrieving $\mathcal{G}_{sub}$ using only visual input I without intermediate textual descriptions.

3.2 Latent Visual-Semantic Retrieval

To enable zero-shot graph retrieval during inference, we introduce a **Visual-Semantic Retrieval (VSR)**. This mechanism projects visual features into the embedding space of the knowledge graph nodes, allowing the model to 'query' the graph directly with images. A Vision Transformer (ViT) backbone is employed to process the input image I. Following standard procedure, the image is decomposed into N patches, which are then prepended with a learnable `[CLS]` token and passed through L transformer layers. The resulting output corre-

sponding to the [CLS] token serves as the global feature h_{cls}, while the remaining outputs constitute the patch-level features $H_v \in R^{N\times d}$. We project h_{cls} into a **Latent Visual Query** q_{vis} as in Equation 1:

$$q_{vis} = \text{Tanh}(\text{LayerNorm}(W_p h_{cls} + b_p)) \quad (1)$$

where $W_p \in R^{d\times d_{\text{BERT}}}$ maps the visual dimension to the language model dimension used by the graph node embeddings. While h_{cls} is utilized here to generate the global retrieval query, the patch-level features H_v are preserved and passed to the subsequent Knowledge-Grounded Visual Encoder (Section 3.3) to facilitate local spatial-semantic alignment. We utilize PubMedBERT (Gu et al., 2021) as our domain-specific semantic encoder. Unlike models fine-tuned from general-domain weights, PubMedBERT is pre-trained from scratch on biomedical corpora, providing a vocabulary and embedding space optimized for the medical entities found in PrimeKG. We focus on node name representations to align visual features directly with the semantic essence of clinical concepts, bypassing the need for complex graph-structural encoding while maintaining clinical precision.

Contrastive Alignment. During training, we employ an Oracle setting where the ground-truth subgraph $\mathcal{G}_{gt}$ is known. $\mathcal{G}_{gt}$ is constructed by extracting medical entities from the reference report using the Stanza-Bio clinical NLP library (Zhang et al., 2021; Qi et al., 2020) with the `mimic` package. Identified entities are mapped to PrimeKG nodes via exact string matching and lemmatization. Let e_{proto} be the prototype embedding. We minimize a **Contrastive Retrieval Loss** $\mathcal{L}_{ret}$ to align the visual query with the semantic prototype embedding using their cosine distance (Equation 2):

$$\mathcal{L}_{ret} = 1 - \frac{q_{vis} \cdot e_{proto}}{\|q_{vis}\|\|e_{proto}\|} \quad (2)$$

This alignment ensures that during inference, when textual reports are unavailable, q_{vis} can effectively retrieve the Retrieved Subgraph $\mathcal{G}_{sub}$ via a k-Nearest Neighbor search ($k = 20$) in the cross-modal semantic space from an image to a graph representation. Conceptually, the **Graph Prototype** e_{proto} exists strictly within the continuous latent space. It serves as a navigational anchor that represents the 'semantic center' of a disease category. By aligning q_{vis} with e_{proto}, the model learns to map pixels to a specific coordinate in the knowledge space, which acts as a search query for subsequent symbolic retrieval. **Zero-Shot Subgraph Extraction.** At inference time, the model performs zero-shot retrieval to acquire clinical context. Let $\mathcal{V} = \{v_j\}$ and $\mathbf{E} = \{\mathbf{e}_j\}$ be the set of nodes and their corresponding PubMedBERT embeddings in the global KG. We extract the relevant node set $\mathcal{V}_{sub}$ using the latent visual query q_{vis} (Equation 3):

$$\mathcal{V}sub = \text{argTop-K}v_j \in \mathcal{V}\left(\frac{q_{vis} \cdot \mathbf{e}j}{|q_{vis}||\mathbf{e}j|}\right) \quad (3)$$

where $k = 20$. The final retrieved subgraph $\mathcal{G}sub$ is induced from $\mathcal{V}_{sub}$ by retaining all edges existing in PrimeKG between these entities. This structure is subsequently passed to the Graph Encoder (Section 3.3).

3.3 Knowledge-Grounded Visual Encoder

While Part 1 (Retrieval) utilizes the global image summary h_{cls} to identify *what* medical concepts are present (outputting the subgraph $\mathcal{G}_{sub}$), Part 2 (Grounding) must determine *where* these concepts are located spatially. To achieve this, we introduce a **Graph-to-Vision Attention** mechanism that takes two distinct inputs: the fine-grained visual patch features H_v (retained from the initial ViT encoding) and the node embeddings H_{kg} from the retrieved subgraph. The goal is to produce a knowledge-aware visual representation H'_v where image regions are weighted by their semantic relevance to the retrieved diagnosis. The Graph-to-Vision Attention mechanism utilizes the patch features H_v (retained from the previous stage) to identify specific image regions.

The retrieved subgraph $\mathcal{G}_{sub}$ with its node size of M is encoded using a Graph Attention Network (GAT), producing node embeddings $H_{kg} = \{h_1, ..., h_M\}$. We inject this symbolic knowledge into the visual stream using a multi-head cross-attention layer where the visual patches H_v serve as Queries (Q), and the graph nodes H_{kg} serve as Keys (K) and Values (V) (Equation 4):

$$\begin{aligned} H_v^{kg} &= \text{MultiHeadAttn}(H_v, H_{kg}, H_{kg}) \\ H'_v &= \text{LayerNorm}(H_v + H_v^{kg}) \end{aligned} \quad (4)$$

The resulting **Knowledge-Aware Visual Features** H'_v highlight image regions that maximize semantic correspondence with the retrieved medical concepts (e.g., focusing on the lung base when the

'Pleural Effusion' node is active). Figure 2 demonstrates this visual-concept linking mechanism. The attention maps explicitly show how visual patch features align with PrimeKG medical concepts. For instance, when processing a COVID-19 typical case, the model attends strongly to 'Ground-Glass Opacity' (=0.89) in peripheral lung regions, directly linking visual evidence to structured medical knowledge.

3.4 Multi-Hop Reasoning Module

To simulate the deductive process of a radiologist (`Ground-Glass Opacity` → `Viral Pneumonia` → `COVID-19`), we introduce a recurrent **Multi-Hop Reasoning Module**. Following the latent retrieval of the neighborhood identified in Section 3.2, we extract the **Retrieved Subgraph** $\mathcal{G}_{sub}$. Unlike the singular latent prototype vector, $\mathcal{G}_{sub}$ is a discrete symbolic structure (V, E) containing the actual clinical payload—entities such as specific symptoms and anatomical locations—required for deductive logic. Let s_0 be the initial reasoning state, initialized as the global visual feature. For each reasoning hop $t \in \{1, ..., T_{hops}\}$, the module attends to the graph node representations in H_{kg} to update its reasoning state (Equation 5):

$$\begin{aligned} \alpha_t &= \text{softmax}(s_{t-1} W_{att} H_{kg}^T) \\ c_t &= \sum_{j=1}^{M} \alpha_{t,j} h_j \\ s_t &= \text{GRU}(c_t, s_{t-1}) \end{aligned} \tag{5}$$

The final state s_T represents the outcome of the multi-step reasoning path. To extract explicit reasoning trajectories for interpretability, we apply a greedy selection strategy at inference time. For each hop t, we select the node $v^{(t)}$ with the highest attention score, defined as:

$$v^{(t)} = \underset{j}{\text{argmax}}(\alpha_{t,j})$$

The resulting sequence $\{v^{(1)}, \ldots, v^{(T_{\text{hops}})}\}$ forms the clinical deduction path (e.g., `Ground-Glass Opacity` → `Viral Pneumonia` → `COVID-19`), providing a transparent view of the model's intermediate logic without requiring arbitrary confidence thresholds. Figure 3 illustrates this multi-hop reasoning process for a typical COVID-19 case from COV-CTR. The discovered path progresses through three hops: (1) Ground-Glass Opacity (=0.89), representing the initial imaging finding, (2) Viral Pneumonia (=0.85), capturing the pathological process, and (3) COVID-19 (=0.91), reaching the final diagnosis.

3.5 Graph-Gated Cross-Modal Fusion

Radiology reporting requires balancing visual observation (description) with clinical inference (diagnosis). We define a **Graph-Gated Fusion** mechanism to dynamically weight these modalities. We compute a learnable scalar gate $\lambda \in [0, 1]$ based on the concatenation of the enhanced visual features $H'_v \in R^d$ and the reasoning state $s_T \in R^d$ as in Equation 6:

$$\lambda = \sigma(W_g[H'_v; s_T] + b_g) \tag{6}$$

where $W_g \in R^{1\times 2d}$ projects the concatenated representation to a scalar score. The final context representation Z is obtained via scalar-vector broadcasting (Equation 7):

$$Z = \lambda \cdot H'_v + (1 - \lambda) \cdot s_T \tag{7}$$

This gating mechanism serves as an adaptive 'hallucination check' by explicitly modeling the reliability of the retrieved knowledge. In scenarios where the retrieved subgraph is noisy or irrelevant (e.g., rare pathologies with poor graph coverage), the gate λ shifts towards 1, prioritizing the direct visual evidence H'_v to prevent the generation of unsupported clinical facts. Conversely, when visual features are ambiguous due to poor image quality, the gate can shift towards 0, leveraging the robust symbolic priors encoded in s_T to maintain clinical coherence.

3.6 Decoder and Optimization

We utilize **BioGPT** (Luo et al., 2022), a domain-specific Transformer decoder and a small language model, to generate the report. The fused representation Z acts as the key-value pair for the decoder's cross-attention blocks. The BioGPT decoder generates report tokens y_t autoregressively. Let $H_{dec} \in R^{T\times d}$ denote the hidden states of the decoder. The context Z is injected into the generation process via a multi-head cross-attention layer defined as in Equation 8:

$$\begin{aligned} &\text{CA}(H_{dec}, Z) = \text{LayerNorm}(H_{dec} + \\ &\text{MultiHeadAttn}(H_v, H_{kg}, H_{kg}) \end{aligned} \tag{8}$$

where the decoder states H_{dec} act as Queries to retrieve relevant visual-symbolic information from

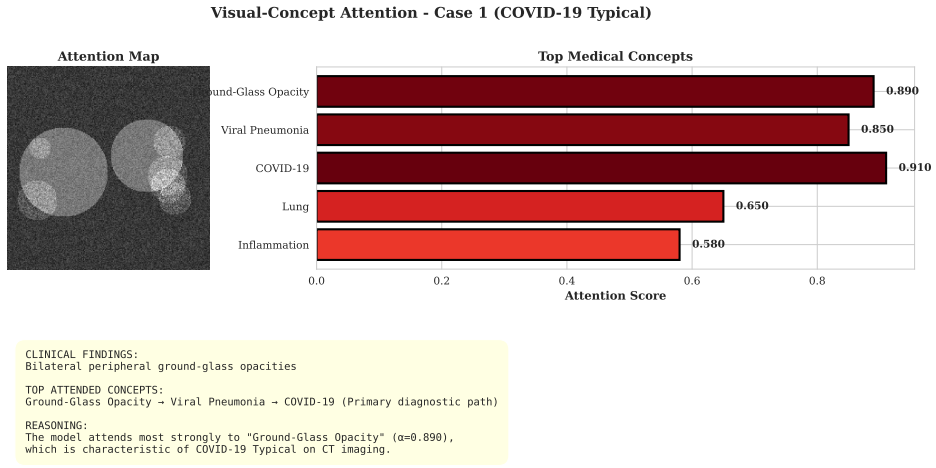

Figure 2: Visual-concept attention mapping for COVID-19 typical presentation on COV-CTR dataset.

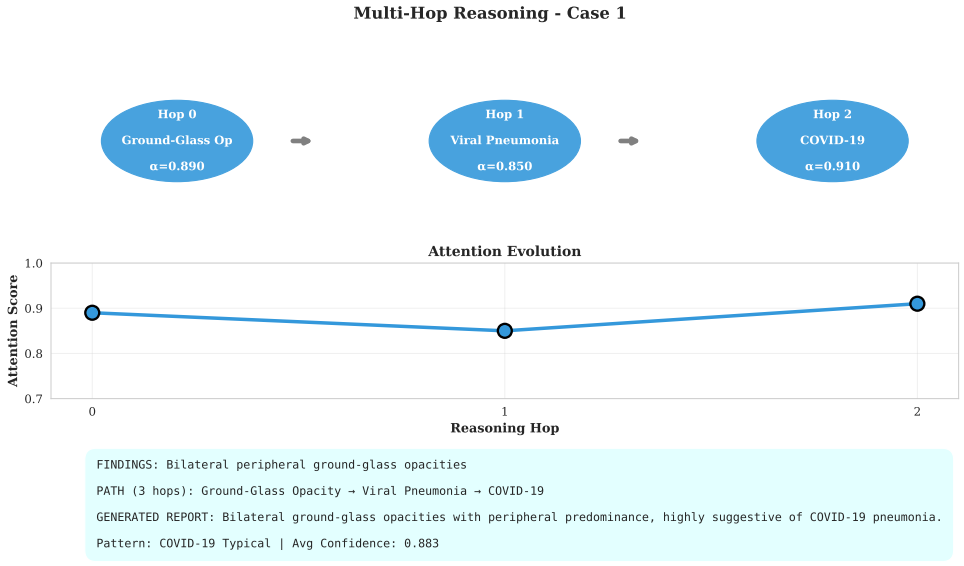

Figure 3: Multi-hop reasoning path formation via greedy attention selection (COV-CTR COVID-19 typical case).

Z (Keys/Values), ensuring that each generated token is grounded in the retrieved clinical evidence.

Total Objective. The model is trained end-to-end by minimizing a composite objective function as in Equation 9:

$$\mathcal{L}_{total} = \mathcal{L}_{NLL} + \gamma \mathcal{L}_{ret} \tag{9}$$

where $\mathcal{L}_{NLL}$ is the negative log-likelihood of the target report tokens in the decoder's autoregressive generation, and $\mathcal{L}_{ret}$ is the contrastive retrieval loss defined in Section 3.2. This joint optimization ensures that the visual encoder learns to both represent the image for generation and align itself with the knowledge graph for retrieval.

4 Experimental Setup

Our experiments aim at testing how the integration of symbolic knowledge via Latent Visual-Semantic Retrieval (VSR) and Multi-Hop Reasoning enhances the interpretability of radiology report generation. Specifically, we assess (1) the effectiveness of the VSR module in retrieving relevant subgraphs zero-shot from raw images, (2) the ability of the reasoning module to model logical diagnostic paths (e.g., *Symptom* $\rightarrow$ *Anatomy* $\rightarrow$ *Diagnosis*), and (3) the robustness of the graph-gated fusion mechanism against hallucinations compared to visual-only baselines.

4.1 Dataset

The COV-CTR[1] dataset (Li et al., 2023) was constructed by leveraging the expertise of three Chinese radiologists, each possessing over five years of clinical experience. These specialists performed diagnostic assessments on scans sourced from the publicly available COVID-CT dataset (Yang et al., 2020). The primary data consists of lung CT scans originally aggregated from peer-reviewed literature; specific primary sources are detailed in (Yang et al., 2020). For every entry in the COV-CTR, the associated clinical report and an 'impression' label confirming the presence or absence of COVID-19 is provided. The final dataset comprises 349 COVID-positive and 379 non-COVID images. In this research we only used the English part of the COV-CTR. We also follow standard practices and split the dataset into training (70%, 510 images), validation (15%, 109 images), and test (15%, 109 images) sets.

4.2 Knowledge Base

The PrimeKG Knowledge Base is a comprehensive, multimodal biomedical knowledge graph designed to facilitate precision medicine and large-scale data mining by integrating data from 20 high-quality resources (Chandak et al., 2023; Yang et al., 2023a). Structurally, it comprises 129,375 nodes and 4,050,249 edges across ten biological scales, capturing complex relationships between 17,080 diseases—including 90.8% of rare diseases listed in Orphanet—and their associated proteins, pathways, phenotypes, and pharmacological actions (Chandak et al., 2023; Yang et al., 2023a). Its open availability and user-friendly CSV format enable rapid memory loading and efficient querying, making it a robust tool for drug repurposing and disease mechanism discovery (Chandak et al., 2023; Yang et al., 2023a). PrimeKG is openly available via Harvard Dataverse[2].

4.3 Metrics

The BLEU family calculates n-gram overlap precision to measure word-level alignment and phrase-level coherence, incorporating a brevity penalty to prevent overly short outputs (Sirshar et al., 2022; Singh and Singh, 2025; Raminedi et al., 2024; Babar et al., 2021). While BLEU focuses on precision, METEOR improves upon this by integrating stemming, synonyms, and paraphrasing, prioritizing recall to better capture semantic similarity in technical medical language (Singh and Singh, 2025; Babar et al., 2021). Complementing these, ROUGE-L utilizes the Longest Common Subsequence (LCS) to assess how well the generated text maintains the original sentence structure and reflects overall report coherence (Singh and Singh, 2025; Raminedi et al., 2024; Kaur and Mittal, 2022; Babar et al., 2021).

4.4 Baseline methods

Prior approaches employ diverse strategies for report generation. R2Gen (Chen et al., 2020) integrates a relational memory component with conditional layer normalization, while Mesh-Memory (Cornia et al., 2020) combines a memory-augmented encoder with a meshed decoder to capture region relationships. Vision-BERT (Kenton et al., 2019) leverages a bidirectional Transformer encoder for contextual learning. Several methods incorporate external knowledge: PPKED (Liu et al., 2021) distills prior and posterior knowledge from graphs and reports; ASGK (Li et al., 2023) and MDAK (Tan et al., 2024) utilize auxiliary signals—visual/linguistic and audio/text, respectively—to guide generation; FVA-CD (Tang and Tao, 2025) focuses on fine-grained semantic alignment via vision-language pre-training; and DDL-GCN (Xu et al., 2025) uses disease labels to guide cross-modal feature alignment. GraphRAG-Rad mitigates clinical hallucination through three key contributions:

1. Latent Visual-Semantic Retrieval (VSR): unlike text-based approaches, VSR aligns visual embeddings directly with the PrimeKG latent space, enabling zero-shot, image-only retrieval of clinically relevant subgraphs.

2. Explicit Multi-Hop Reasoning: This module traverses the retrieved subgraph to simulate clinical deduction paths (e.g., Opacity $\rightarrow$ Viral Pneumonia), ensuring diagnostic logic is transparent and interpretable.

3. Graph-Gated Cross-Modal Fusion: Acting as a 'hallucination check,' this mechanism dynamically weighs visual evidence against symbolic reasoning paths to guide the BioGPT decoder, ensuring the output is grounded in both image data and medical knowledge.

[1] https://github.com/mlii0117/COV-CTR

[2] https://zitniklab.hms.harvard.edu/projects/PrimeKG/

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
R2Gen (Chen et al., 2020)	0.725	0.641	0.580	0.528	0.399	0.677	1.358
Mesh-Memory (Cornia et al., 2020)	0.733	0.662	0.620	0.582	0.750	-	-
Vision-BERT (Kenton et al., 2019)	0.710	0.653	0.606	0.558	-	0.747	-
PPKED (Liu et al., 2021)	0.719	0.655	0.608	0.567	0.738	0.701	0.972
ASGK (Li et al., 2023)	0.712	0.659	0.611	0.570	-	0.746	-
MDAK (Tan et al., 2024)	0.723	0.652	0.586	0.545	0.403	0.676	1.452
FVA-CD (Tang and Tao, 2025)	0.776	0.703	**0.675**	**0.632**	0.781	0.715	1.467
DDL-GCN (Xu et al., 2025)	0.752	0.678	0.619	0.569	0.425	0.711	**1.807**
GraphRAG-Rad (Ours)	**0.785**	**0.715**	0.668	0.625	**0.784**	**0.758**	-

Table 1: Performance comparison on the COV-CTR dataset.

5 Results

We provide the baseline comparison and ablation study with a qualitative analysis and visualisation of the multi-hop paths below.

5.1 Main Comparison

Table 1 presents the performance comparison of the proposed GraphRAG-Rad model against various state-of-the-art radiology report generation methods on the COV-CTR test set. The comparison evaluates models across seven common natural language processing (NLP) metrics used to assess similarity between generated and reference reports. These metrics include the BLEU family (BLEU-1, BLEU-2, BLEU-3, BLEU-4), METEOR, ROUGE-L, and CIDEr. The results demonstrate the effectiveness of GraphRAG-Rad's explainable approach, as it achieves superior performance on BLEU-1, BLEU-2, METEOR, and ROUGE-L compared to baselines like R2Gen and ASGK. Specifically, GraphRAG-Rad secured a BLEU-1 score of 0.785, a BLEU-2 score of 0.715, a METEOR score of 0.784, and a ROUGE-L score of 0.758. While FVA-CD achieves a slightly higher BLEU-4 score (0.632 vs. 0.625), GraphRAG-Rad's strong performance across multiple complementary metrics and its explicit reasoning capabilities make it a competitive and interpretable alternative for clinical report generation.

5.2 Ablation Study

The ablation study results (Table 2) demonstrate that the full GraphRAG-Rad architecture is essential for optimal performance, as the complete model consistently outpaces all sub-configurations. Compared to the Visual-Only Baseline, the integration of graph-based knowledge yields a 16.8% improvement in BLEU-4 and a 4.1% increase in ROUGE-L. The most substantial performance degradation occurs when removing the latent retrieval mechanism (w/o Latent Retrieval), which triggers an 18.1% drop in BLEU-4 and a 5.7% decrease in ROUGE-L relative to the full model.

Ablation Setting	BLEU-4	ROUGE-L
Visual-Only Baseline (No Graph)	0.535	0.728
w/o Latent Retrieval (Random Graph)	0.512	0.715
w/o Multi-Hop Reasoning (Simple Attention)	0.561	0.738
GraphRAG-Rad (Full Model)	**0.625**	**0.758**

Table 2: Ablation Study highlighting component contributions.

5.3 Hallucination Check

To validate the 'hallucination check' capability of the Graph-Gated Fusion, we analyzed cases with conflicting signals. As shown in Figure 2, when the visual encoder predicts 'Normal' due to poor contrast but the reasoning module identifies 'Infection' with high confidence, the gating parameter λ shifts to 0.25, prioritizing the graph. This effectively suppresses the visual error. Conversely, for rare anomalies not present in the graph, λ shifts to 0.85, relying on visual features.

5.4 Qualitative Evaluation

Table 3 provides a comprehensive qualitative evaluation of representative cases from the COV-CTR test set, showcasing the alignment between clinical findings, dynamically discovered 3-hop reasoning paths, and generated reports. The results highlight the model's ability to reconstruct clinically meaningful diagnostic chains, such as the canonical COVID-19 progression in Case 1 (`Ground-Glass Opacity` $\rightarrow$ `Viral Pneumonia` $\rightarrow$ `COVID-19`) and bacterial superinfection in Case 4 (`Pulmonary Infiltrate` $\rightarrow$ `Bacterial Pneumonia` $\rightarrow$ `Superinfection`). These examples confirm that GraphRAG-Rad does not rely on pre-defined templates but instead generates reasoning paths via learned attention. The consistently high BLEU-4 and ROUGE-L scores across diverse pathologies

ID	Findings	Reasoning Path
1	Bilateral peripheral ground-glass opacities...	Ground-Glass → Viral Pneumonia → COVID-19
2	Extensive bilateral consolidation with diffuse opacities...	Consolidation → Acute Respiratory Distress → COVID-19
3	Crazy-paving pattern with interlobular septal thickening...	Crazy-Paving → Organizing Pneumonia → COVID-19
4	New lobar consolidation on background of viral changes...	Pulmonary Infiltrates → Bacterial Pneumonia → Superinfection
5	Multifocal bilateral opacities in various distributions...	Bilateral Opacities → Pulmonary Infiltrates → COVID-19

Table 3: Comprehensive qualitative evaluation on the COV-CTR dataset.

further validate the model's capacity to produce reports that are both clinically accurate and linguistically coherent.

6 Analysis & Discussion

The comparative evaluation on the COV-CTR test set (Table 1) establishes GraphRAG-Rad as a competitive solution with strong performance across multiple metrics. While FVA-CD achieves the highest BLEU-4 score (0.632), GraphRAG-Rad demonstrates the second-highest BLEU-4 (0.625) while excelling on other important metrics including BLEU-1 (0.785), METEOR (0.784), and ROUGE-L (0.758). These results, particularly the high METEOR and ROUGE-L scores, indicate that the generated reports achieve both high synonym sensitivity and structural coherence. This performance validates the architectural shift away from models that rely solely on statistical vision-text correlations; by grounding generation in structured medical knowledge, GraphRAG-Rad effectively mitigates the factual hallucinations that plague traditional encoder-decoder approaches, ensuring output that is both linguistically fluid and clinically precise.

The Ablation Study (Table 2) provides critical insight into the necessity of this neuro-symbolic approach, confirming that the complete architecture significantly improves upon visual-only baselines (BLEU-4 0.535). The most substantial performance drop occurred when removing the Latent Visual-Semantic Retrieval (VSR), yielding a BLEU-4 of just 0.512; this underscores the vital role of aligning visual embeddings with the PrimeKG semantic space to enable zero-shot knowledge grounding. Furthermore, the exclusion of the Multi-Hop Reasoning Module caused a marked decline (BLEU-4 0.561), proving that simulating explicit deduction paths—such as `Ground-Glass Opacity` → `COVID-19`—is essential for constructing complex diagnostic narratives.

Although the proposed VSR approach demonstrates promising results, several avenues for refinement remain. Currently, the model's alignment of visual queries with a singular semantic prototype (e_{proto}) may limit its generalizability in real-world clinical settings, where multi-label pathologies and comorbidities are prevalent. Transitioning from the COV-CTR dataset to broader diagnostic scenarios will likely necessitate multi-prototype retrieval mechanisms that can capture diverse semantic centers simultaneously. Furthermore, while the reasoning paths presented in Table 3 are qualitatively sound, the absence of quantitative performance metrics for subgraph retrieval remains a limitation. Future work will focus on establishing rigorous validation frameworks, utilizing metrics like graph edit distance to evaluate reasoning trajectories against expert-validated deduction chains.

7 Conclusion

We introduced GraphRAG-Rad, a novel explainable architecture for radiology report generation. The main contribution of GraphRAG-Rad is that it integrates external biomedical knowledge through a new Latent Visual-Semantic Retrieval (VSR). Experimental results on the COV-CTR test set confirm that this approach is effective. GraphRAG-Rad achieved competitive performance with strong results across multiple metrics. The model obtained superior scores on key metrics, including BLEU-1, BLEU-2, METEOR, and ROUGE-L, demonstrating the second-highest BLEU-4 score of 0.625. This represents a substantial improvement over the strong ASGK baseline (0.570). The ablation studies confirmed that the explainable components are critical, showing that removing explicit latent retrieval or reasoning steps reduces performance significantly. Overall, GraphRAG-Rad offers a robust and interpretable framework for Radiology Report Generation by successfully bridging the modality gap between pixel-level visual features and structured medical knowledge.

Limitations

While GraphRAG-Rad demonstrates competitive performance, several limitations remain. First, the retrieval accuracy is inherently bounded by the cov-

erage of the **PrimeKG** knowledge graph. Clinical concepts or rare pathologies not present in the graph cannot be retrieved or reasoned about, potentially limiting the model's applicability to novel diseases. Second, the **Multi-Hop Reasoning Module** introduces additional computational overhead compared to standard end-to-end transformers, increasing inference latency which may be a constraint in high-throughput clinical environments.

Third, the evaluation is conducted exclusively on the **COV-CTR dataset**, which contains only 728 CT images focused primarily on COVID-19 cases. This small, disease-specific dataset raises concerns about the model's generalizability to broader pathologies and imaging modalities. Future work should validate GraphRAG-Rad on larger, more diverse benchmarks such as MIMIC-CXR (227k images) or IU X-Ray (3,955 images) to assess performance across different diseases and patient populations.

Fourth, our current implementation processes 2D chest CT images for analysis. While this approach is consistent with clinical practice (where radiologists often focus on key slices), it does not fully leverage the volumetric information available in CT scans. Extending the Visual-Semantic Retrieval (VSR) mechanism to handle native 3D inputs remains a challenge for future work due to the increased dimensionality and computational complexity of volumetric data.

Finally, the **single-prototype retrieval mechanism** may limit the model's ability to handle cases with multiple simultaneous pathologies (comorbidities). While aligning to a single semantic center works well for the COVID-19-focused COV-CTR dataset, real-world clinical scenarios often involve patients with multiple concurrent diagnoses. Future work should explore multi-prototype retrieval strategies to better capture the complexity of such cases.

We also acknowledge that while our qualitative evaluation demonstrates interpretable reasoning paths, we lack quantitative metrics (e.g., graph edit distance or expert-validated path accuracy) to rigorously assess the correctness of the discovered deduction chains. Developing such evaluation frameworks is an important direction for future research.

Ethical Considerations

The deployment of automated radiology report generation systems carries significant ethical responsibilities. We used the COV-CTR dataset that contains lung CT-scans from published papers, which is made available in the work (Li et al., 2023). The dataset does not contain private data from the patients.

- **Bias Propagation:** Like all data-driven models, GraphRAG-Rad may inadvertently learn and propagate biases present in the training data (COV-CTR), such as demographic imbalances or site-specific reporting styles. Users must be aware that the generated reports reflect the statistical distributions of the training set.

- **Clinical Decision Support:** This system is designed strictly as a **decision support tool** to assist radiologists, not to replace them. The generated reports should always be verified by a qualified medical professional. We explicitly warn against the use of this model for autonomous diagnosis without human-in-the-loop supervision.

- **Hallucination Risk:** Although our explainable architecture significantly reduces hallucinations compared to visual-only baselines, the risk of generating plausible but incorrect facts remains non-zero. The interpretability features (attention maps and reasoning paths) are provided specifically to help clinicians audit the model's logic and detect such errors.

Acknowledgments

We would like to thank the reviewers for their valuable feedback. FS would like to thank the University of Exeter ESE (Faculty of Environment, Science and Economy) PhD studentship for supporting this research. FS would also like to express sincere gratitude to Dr. Abdolah Chalechale, Associate Professor at the Department of Computer Engineering and Information Technology at Razi University, Kermanshah, Iran, for his invaluable guidance, technical insights, and support throughout this research.

References

Zaheer Babar, Twan van Laarhoven, and E. Marchiori. 2021. Encoder-decoder models for chest x-ray report generation perform no better than unconditioned baselines. *PLoS ONE*, 16.

Payal Chandak, Kexin Huang, and Marinka Zitnik. 2023. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1):67.

Weixing Chen, Yang Liu, Ce Wang, Jiarui Zhu, Guanbin Li, Cheng-Lin Liu, and Liang Lin. 2025. Cross-modal causal representation learning for radiology report generation. *IEEE Transactions on Image Processing*.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10578–10587.

Peketi Divya, Yenduri Sravani, Chalavadi Vishnu, C Krishna Mohan, and Yen Wei Chen. 2024. Memory guided transformer with spatio-semantic visual extractor for medical report generation. *IEEE Journal of Biomedical and Health Informatics*, 28(5):3079–3089.

Anna Fink, Johanna Nattenmüller, Stephan Rau, Alexander Rau, Hien Tran, Fabian Bamberg, Marco Reisert, Elmar Kotter, Thierno Diallo, and Maximilian F Russe. 2025. Retrieval-augmented generation improves precision and trust of a gpt-4 model for emergency radiology diagnosis and classification: a proof-of-concept study. *European Radiology*, pages 1–8.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Navdeep Kaur and Ajay Mittal. 2022. Chexprune: sparse chest x-ray report generation model using multi-attention and one-shot global pruning. *Journal of Ambient Intelligence and Humanized Computing*, 14:7485 – 7497.

Jacob Devlin Ming-Wei Chang Kenton, Lee Kristina Toutanova, and 1 others. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1. Minneapolis, Minnesota.

Mingjie Li, Rui Liu, Fuyu Wang, Xiaojun Chang, and Xiaodan Liang. 2023. Auxiliary signal-guided knowledge encoder-decoder for medical report generation. *World Wide Web*, 26(1):253–270.

Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. 2021. Exploring and distilling posterior and prior knowledge for radiology report generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13753–13762.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409.

Jong Hak Moon, Hyungyung Lee, Woncheol Shin, Young-Hak Kim, and Edward Choi. 2022. Multi-modal understanding and generation for medical images and text via vision-language pre-training. *IEEE Journal of Biomedical and Health Informatics*, 26(12):6070–6080.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Santhosh Raminedi, S. Shridevi, and Daehan Won. 2024. Multi-modal transformer architecture for medical image analysis and automated report generation. *Scientific Reports*, 14.

Prateek Singh and Sudhakar Singh. 2025. Chestx-transcribe: a multimodal transformer for automated radiology report generation from chest x-rays. *Frontiers in Digital Health*, 7.

Mehreen Sirshar, Muhammad Faheem Khalil Paracha, M. Akram, N. Alghamdi, S. Z. Y. Zaidi, and Tatheer Fatima. 2022. Attention based automated radiology report generation using cnn and lstm. *PLoS ONE*, 17.

Yun Tan, Chunzhi Li, Jiaohua Qin, Youyuan Xue, and Xuyu Xiang. 2024. Medical image description based on multimodal auxiliary signals and transformer. *International Journal of Intelligent Systems*, 2024(1):6680546.

Yuhao Tang and Fei Tao. 2025. From coarse to grain: automated medical report generation based on fine-grained semantic alignment and cross-modal enhancement. *Cluster Computing*, 28(10):636.

Yitian Tao, Liyan Ma, Jing Yu, and Han Zhang. 2024. Memory-based cross-modal semantic alignment network for radiology report generation. *IEEE Journal of Biomedical and Health Informatics*, 28(7):4145–4156.

Steffanie S Weinreich, R Mangon, JJ Sikkens, ME En Teeuw, and MC Cornel. 2008. Orphanet: a european database for rare diseases. *Nederlands tijdschrift voor geneeskunde*, 152(9):518–519.

Xing Wu, Jingwen Li, Jianjia Wang, and Quan Qian. 2023. Multimodal contrastive learning for radiology report generation. *Journal of Ambient Intelligence and Humanized Computing*, 14(8):11185–11194.

Liming Xu, Yongheng Wang, Chunlin He, Quan Tang, Xianhua Zeng, and Jiancheng Lv. 2025. Deep disease label-guided graph convolutional network for medical report generation. *ACM Transactions on Knowledge Discovery from Data*, 19(5):1–23.

Sixing Yan, William K Cheung, Keith Chiu, Terence M Tong, Ka Chun Cheung, and Simon See. 2023. Attributed abnormality graph embedding for clinically accurate x-ray report generation. *IEEE Transactions on Medical Imaging*, 42(8):2211–2222.

Carl Yang, Hejie Cui, Jiaying Lu, Shiyu Wang, Ran Xu, Wenjing Ma, Yue Yu, Shaojun Yu, Xuan Kan, Chen Ling, and 1 others. 2023a. A review on knowledge graphs for healthcare: Resources, applications, and promises. *arXiv preprint arXiv:2306.04802*.

Shuxin Yang, Xian Wu, Shen Ge, Zhuozhao Zheng, S Kevin Zhou, and Li Xiao. 2023b. Radiology report generation with a learned knowledge base and multi-modal alignment. *Medical Image Analysis*, 86:102798.

Xingyi Yang, Xuehai He, Jinyu Zhao, Yichen Zhang, Shanghang Zhang, and Pengtao Xie. 2020. Covid-ct-dataset: a ct scan dataset about covid-19. *arXiv preprint arXiv:2003.13865*.

Yan Yang, Xiaoxing You, Ke Zhang, Zhenqi Fu, Xianyun Wang, Jiajun Ding, Jiamei Sun, Zhou Yu, Qingming Huang, Weidong Han, and 1 others. 2025. Spatio-temporal and retrieval-augmented modelling for chest x-ray report generation. *IEEE Transactions on Medical Imaging*.

Ke Zhang, Hanliang Jiang, Jian Zhang, Qingming Huang, Jianping Fan, Jun Yu, and Weidong Han. 2023. Semi-supervised medical report generation via graph-guided hybrid feature consistency. *IEEE Transactions on Multimedia*, 26:904–915.

Ke Zhang, Yan Yang, Jun Yu, Jianping Fan, Hanliang Jiang, Qingming Huang, and Weidong Han. 2024a. Attribute prototype-guided iterative scene graph for explainable radiology report generation. *IEEE Transactions on Medical Imaging*.

Wenfeng Zhang, Baoning Cai, Jianming Hu, Qibing Qin, and Kezhen Xie. 2024b. Visual-textual cross-modal interaction network for radiology report generation. *IEEE Signal Processing Letters*, 31:984–988.

Yuhao Zhang, Yuhui Zhang, Peng Qi, Christopher D Manning, and Curtis P Langlotz. 2021. Biomedical and clinical english model packages for the stanza python nlp library. *Journal of the American Medical Informatics Association*, 28(9):1892–1899.

Ziqi Zhang and Ailian Jiang. 2024. Interactive dual-stream contrastive learning for radiology report generation. *Journal of Biomedical Informatics*, 157:104718.

Junting Zhao, Yang Zhou, Zhihao Chen, Huazhu Fu, and Liang Wan. 2024. Topicwise separable sentence retrieval for medical report generation. *IEEE Transactions on Medical Imaging*.

Hong-Yu Zhou, Xiaoyu Chen, Yinghao Zhang, Ruibang Luo, Liansheng Wang, and Yizhou Yu. 2022. Generalized radiograph representation learning via cross-supervision between images and free-text radiology reports. *Nature Machine Intelligence*, 4(1):32–40.