CPR-RAG: Clinical Prior-Regularized Retrieval for Anatomy-Aware 3D CT Report Generation

Anonymous ACL submission

Abstract

Generating radiology reports from 3D volumetric data remains challenging due to the difficulty of grounding fine-grained pathologies within high-dimensional scans. While retrieval-augmented generation (RAG) offers a potential solution, standard approaches struggle with visual-semantic ambiguity and often introduce irrelevant "normal" context that dilutes pathological signals. To address this limitation, we introduce *CPR-RAG*, a model-agnostic RAG framework that enhances organ-level grounding by integrating clinical priors into the retrieval process. Specifically, we propose a *clinical prior-regularized re-ranking* module that leverages corpus-derived co-occurrence statistics to align retrieved candidates with latent disease distributions, ensuring clinical consistency beyond mere visual similarity. Furthermore, we employ *clinical relevance context refinement* to selectively filter out boilerplate normal descriptions, thereby maximizing the information density of the evidence provided to the generator. Extensive experiments on the RadGenome-ChestCT benchmark demonstrate that *CPR-RAG* significantly improves clinical efficacy across state-of-the-art radiology report generation models. Human evaluation further confirms that our approach achieves superior factual correctness, completeness, and utility compared to the existing models.

1 Introduction

Recent large language models (LLMs) have demonstrated remarkable capabilities in general text generation (OpenAI, 2023; Grattafiori et al., 2024). However, its application in clinical settings remains challenging due to strict safety requirements, especially for generating radiology reports from high-dimensional 3D volumetric modalities such as CT and MRI (Thirunavukarasu et al., 2023).

Radiology report generation from 3D volumes is a section-structured, multi-organ reasoning task. Given that clinical reports are organized into distinct anatomical structures (e.g., lung, mediastinum, pleura), the model requires localizing clinical findings and associating them with their specific anatomical structures. Consequently, the model faces a challenging grounding task that requires the model to link specific textual descriptions to potential visual evidence scattered across the high-dimensional 3D volume (Hamamci et al., 2024a; Tu et al., 2024).

The challenge of visual-text grounding is further exacerbated by the morphological heterogeneity of findings (Litjens et al., 2017; Raghu et al., 2019). Unlike generic objects, lesions present diverse variations in shape, texture, and size, often mimicking normal physiological structures. Accurately identifying these anomalies requires expert-level pattern recognition to distinguish fine-grained pathological cues from the dominant normal background (Esteva et al., 2019).

Despite recent progress, medical vision language models (VLMs) still struggle with long-context grounding in 3D volumes. Compressing high-dimensional scans into compact representations can attenuate fine-grained regional evidence and exacerbate visual-semantic misalignment (Bai et al., 2024; Lin et al., 2024). Moreover, radiology report generation often exhibits normality bias under severe normal/abnormal imbalance, leading the generator to default to generic "safe" templates (e.g., "No acute abnormality") rather than precise, evidence-grounded descriptions (Zhang et al., 2020; Miura et al., 2021).

Retrieval-augmented generation (RAG) is a natural strategy to improve grounding by providing relevant external references (e.g., similar cases or reports) as additional context (Ranjit et al., 2023; Endo et al., 2021). However, conventional retrieval approaches using medical imaging face a fundamental challenge due to visual-semantic misalignment, where visually similar patterns do not nec-

essarily imply identical clinical etiology. For example, pneumonia and atelectasis can both present as pulmonary opacities (Irvin et al., 2019; Wang et al., 2017), yet require different clinical interpretations and management. These errors are particularly problematic in multi-organ report generation scenarios. A retrieved case that does not perfectly match the query across all anatomical structures due to visual-semantic misalignment can lead the generator into hallucinating.

To address these challenges, we introduce **CPR-RAG** (**C**linical **P**rior-**R**egularized **R**AG), a plug-and-play framework that improves organ-level grounding for 3D report generation. Instead of relying solely on visual similarity, our approach integrates clinical regularity into the retrieval process. Specifically, *CPR-RAG* combines prior-based re-ranking to enforce clinical consistency and a context refinement mechanism to maximize the information density of the retrieved context to minimize hallucination. This strategy ensures that the generator is guided by high-density pathological evidence rather than noisy or irrelevant priors.

Our contributions are summarized as follows:

- We propose *CPR-RAG*, a model-agnostic, organ-centric retrieval framework that can be integrated into the existing 3D VLM report generators in a plug-and-play fashion.
- We introduce *clinical prior-regularized reranking* to address visual-semantic misalignment by leveraging corpus-level statistics, and *clinical relevance context refinement* to filter out boilerplate normal findings. This strategy effectively concentrates the information density of pathological cues, thereby significantly improving context relevance for the generator.
- Extensive experiments demonstrate that our method significantly improves clinical efficacy across the baselines (i.e., CT2Rep, M3D, RadFM). Furthermore, human evaluation confirms that *CPR-RAG* achieves higher correctness, completeness, and utility with fewer hallucinations compared to baseline method.

2 Related Work

2.1 Visual-Conditioned Radiology Report Generation

Radiology report generation is formally a visual-conditioned text generation task, translating visual biomarker into natural language descriptions. Common approaches adopt an encoder-decoder architecture conditioned on visual embeddings (Chen et al., 2020; Liu et al., 2021). In this framework, a visual encoder projects the input image into a sequence of continuous embeddings, which serve as soft prompts guiding an autoregressive LLM to generate the report.

While effective for 2D modalities (e.g., X-rays) (Lee et al., 2025), extending this paradigm to 3D volumetric data (e.g., CT, MRI) presents distinct representational constraints. Unlike 2D images, a 3D volume contains dense spatial information across hundreds of slices. Recent 3D-specialized VLMs, such as RadFM (Wu et al., 2025) and CT2Rep (Hamamci et al., 2024a), employ 3D-ViTs or resampling modules to project volumetric features into serialized token sequences aligned with the LLM.

However, relying solely on fixed-length visual tokens creates an information bottleneck, often attenuating fine-grained details. This limitation inevitably causes factual hallucinations (Ji et al., 2023), highlighting the necessity for explicit, non-parametric grounding via RAG.

2.2 Retrieval-Augmented Generation in Radiology

RAG has emerged as a promising paradigm to address hallucinations by grounding generation on external clinical evidence (Lewis et al., 2020). In medical imaging, conventional approaches typically leverage contrastive vision-language models (e.g., CLIP, BioViL) to retrieve existing radiology reports based on global image-text alignment (Endo et al., 2021; Yan et al., 2024). While effective for 2D X-rays where clinical pathologies remain discernible in planar projections, directly applying this global-level retrieval strategy to 3D volumetric imaging introduces two critical misalignments.

In particular, representing high-dimensional 3D volumes with a single global embedding obscures fine-grained local features, leading to the retrieval of morphologically similar yet pathologically different diagnoses. Second, and more critically, is the issue of multi-organ entanglement. Since 3D CT reports typically describe multiple anatomical structures simultaneously, retrieving a full report may include irrelevant descriptions of other organs, such as templated normal findings. Such noise dilutes the information density within the prompt, causing the LLM to overlook the target abnormality

due to the "lost-in-the-middle" phenomenon (Liu et al., 2024).

Recent studies have attempted to address these challenges. Mao et al. (2025) proposed CT-Agent, an agentic framework with memory-based retrieval. However, its iterative tool-execution process incurs prohibitive inference latency and computational overhead compared to end-to-end pipelines. Similarly, RadIR (Zhang et al., 2025) introduced multi-grained retrieval at scale. However, it focuses on the retrieval task itself and does not validate its efficacy for full report generation. These limitations underscore the necessity for an efficient retrieval-generation pipeline.

3 CPR-RAG: Clinical Prior-Regularized Retrieval-Augmented Generation

We propose *CPR-RAG*, a retrieval-augmented generation framework specialized for 3D CT radiology report generation. Since the CT radiology report contains multiple findings across different anatomical regions, such as heart, lungs, and mediastinum, we design *CPR-RAG* to operate in an *organ-centric* manner: for each organ, the model retrieves clinically relevant precedents and generates a localized finding description to complete the report.

As shown in Figure 1, the framework comprises three stages: (1) *anatomy-conditional representation learning* that disentangles global 3D features into organ-specific embeddings via learnable queries; (2) *clinical prior-regularized retrieval*, which integrates visual affinity with corpus-derived co-occurrence statistics to prioritize clinically consistent candidates; and (3) *clinical relevance context refinement*, which minimizes semantic noise by filtering out boilerplate normal descriptions, thereby maximizing the information density of the context provided to the generator.

3.1 Anatomy-Conditional Representation

Because 3D CT scans cover many organs, a single global feature vector can blur small, localized abnormalities into normal anatomy. To enable fine-grained retrieval without requiring segmentation masks, we extract organ-specific embeddings using a query-based cross-attention mechanism.

Organ-specific Embeddings. Let $\mathbf{X} \in \mathbb{R}^{N\times d}$ denote the sequence of visual tokens computed by a frozen 3D visual encoder, where N is the number of tokens and d is the feature dimension. We introduce a learnable query matrix $\mathbf{Q} = [\mathbf{q}_1, \ldots, \mathbf{q}_{|\Omega|}]^\top \in \mathbb{R}^{|\Omega|\times d}$, where each query $\mathbf{q}_i$ corresponds to the i-th anatomical region in Ω ={Heart, Lung, Mediastinum, Pleura, Trachea/Bronchi}. We compute organ-specific embeddings via cross-attention, using $\mathbf{Q}$ as queries and $\mathbf{X}$ as keys/values:

$$\mathbf{E} = \mathrm{CrossAttn}(\mathbf{Q}, \mathbf{X}, \mathbf{X}), \tag{1}$$

where $\mathbf{E} = [\mathbf{e}_1, \ldots, \mathbf{e}_{|\Omega|}]^\top \in \mathbb{R}^{|\Omega|\times d}$, and $\mathbf{e}_i$ is the organ-specific embedding for organ i.

Training via Auxiliary Supervision. To encourage $\mathbf{e}_i$ to encode pathology-relevant semantics, we attach an organ-specific multi-label classifier $C_i(\cdot)$ for each $i \in \Omega$. Let $\mathcal{L}$ denote the universal set of abnormality labels, and $\mathcal{L}_i$ is the subset of labels specific to organ i. For a training sample k, let $\mathbf{y}_i^{(k)} \in \{0,1\}^{|\mathcal{L}_i|}$ represent the multi-hot ground-truth label vector. Given the embedding $\mathbf{e}_i^{(k)}$, the classifier predicts the disease probability $\mathbf{p}_i^{(k)} = C_i(\mathbf{e}_i^{(k)}) \in [0,1]^{|\mathcal{L}_i|}$. We optimize the framework using a binary cross-entropy objective:

$$Loss_{\text{aux}} = \sum_k \sum_{i\in\Omega} \mathrm{BCE}\big(\mathbf{p}_i^{(k)}, \mathbf{y}_i^{(k)}\big). \tag{2}$$

This auxiliary task aligns the organ-specific embeddings with spatially relevant clinical evidence and provides the predicted disease distributions utilized in Sections 3.2 and 3.3.

Organ-Level Index Construction. We build an offline organ-level index $\mathcal{I}$ from the training corpus. For each training sample k and organ $i \in \Omega$, we extract the embedding $\mathbf{e}_i^{(k)}$ and parse the corresponding radiology report $\mathbf{t}_i^{(k)}$. We store the pair $(\mathbf{e}_i^{(k)}, \mathbf{t}_i^{(k)})$ in $\mathcal{I}$, enabling retrieval of organ-specific descriptions at inference time.

3.2 Clinical Prior-Regularized Retrieval

Standard retrieval relying solely on visual embedding similarity often yields noisy results due to the visual ambiguity in medical imaging. To address this problem, we propose a two-stage retrieval strategy that retrieves organ-specific candidates followed by clinical prior-based re-ranking.

Organ-Level Initial Retrival. Given a query volume, we obtain organ embeddings $\{\mathbf{e}_i^{(q)}\}_{i\in\Omega}$ and retrieve candidates independently per organ. For each organ i, we compute cosine similarity between

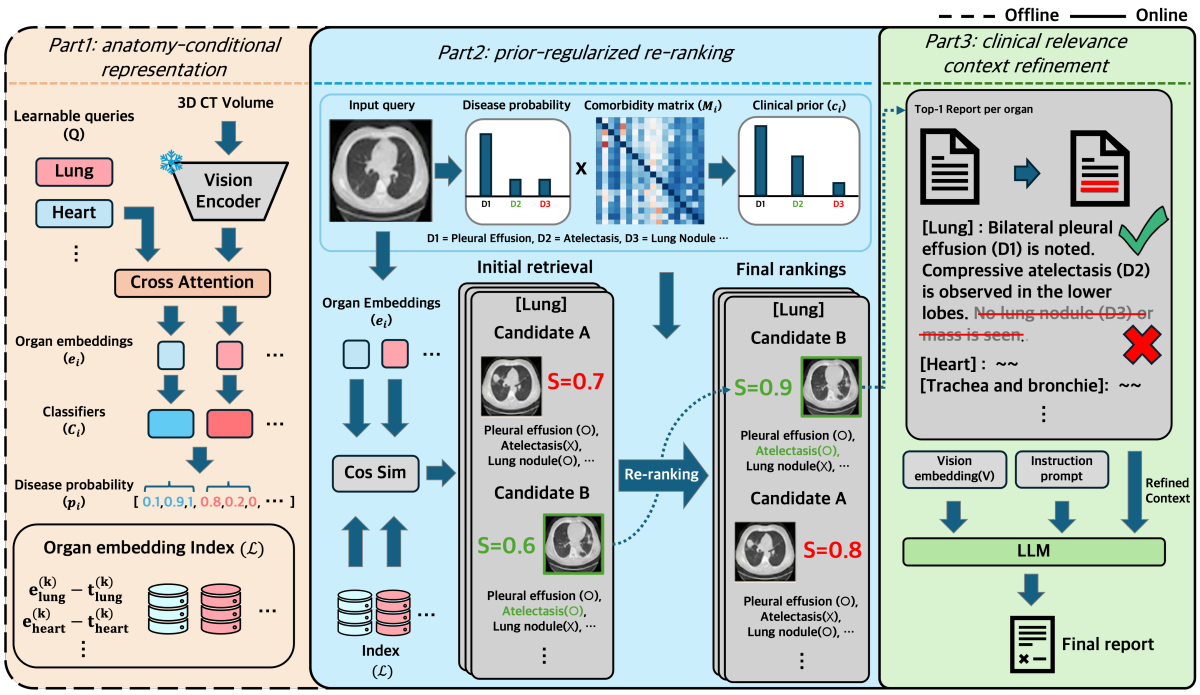

Figure 1: **Overview of the CPR-RAG framework.** First, we train a cross-attention perceiver based on frozen vision encoder from an existing VLM to learn *anatomy-conditional representations* and build an offline organ-level index $\mathcal{I}$. Second, given input image token $\mathbf{e}_i^{(q)}$ as query, we retrieve visually similar candidates and refine them via *clinical prior-regularized re-ranking* using on corpus-derived comorbidity prior M_i for each organ. Finally, we filter out clinically irrelevant contexts (e.g., "No lung nodule or mass is seen ") to maximize information density, and generate final radiology report based on pretrained existing VLM using input image token and refined prompts.

$\mathbf{e}_i^{(q)}$ and all indexed embeddings $\mathbf{e}_i^{(k)}$ in $\mathcal{I}$, and collect the top-K candidates:

$$\mathcal{D}_i = \text{TopK}\Big(\cos(\mathbf{e}_i^{(q)}, \mathbf{e}_i^{(k)})\Big). \tag{3}$$

Although visually similar, the set $\mathcal{D}_i$ may contain clinically inconsistent cases that share image texture but differ in underlying pathology. We therefore re-rank $\mathcal{D}_i$ using a hybrid objective (Eq. 6) that prioritizes candidates that are both visually similar and clinically consistent with the query.

Corpus-Derived Comorbidity Prior. We distill co-occurrence statistics from training reports into a comorbidity prior matrix $\mathbf{M_i} \in [0,1]^{|\mathcal{L}_i|\times|\mathcal{L}_i|}$, where $M_i(j,k) = P(f_j \mid f_k)$ denotes the conditional probability of observing finding f_j given f_k:

$$M_i(j,k) = \frac{\text{Count}(f_j, f_k) + \alpha}{\text{Count}(f_j) + \alpha|\mathcal{L}_i|}, \tag{4}$$

where α denotes smoothing parameter, f_j and f_k denotes disease labels.

Clinical Prior-Regularized Re-ranking. For organ i, we estimate the query-specific disease probability $\mathbf{p}_i$ using the auxiliary classifier trained in Section 3.1. We then propagate this distribution through the prior:

$$\mathbf{c}_i = \mathbf{p}_i^\top \mathbf{M}_i. \tag{5}$$

For each retrieved candidate d in the initial top-K, characterized by its organ-specific embedding $\mathbf{e}_i^{(d)}$ and multi-hot label vector $\mathbf{l}_d \in \{0,1\}^{|\mathcal{L}_i|}$, we compute the refined score:

$$S(q,d) = \cos(\mathbf{e}_i^{(q)}, \mathbf{e}_i^{(d)}) + \lambda \cdot (\mathbf{c}_i \cdot \mathbf{l}_d), \tag{6}$$

where λ controls the contribution of the prior-aligned semantic term.

3.3 Clinical Relevance Context Refinement

Even after re-ranking, the retrieved texts often contain boilerplate normal findings (e.g., “no acute abnormality”, “The heart size is within normal limits.”) alongside pathological descriptions. Injecting such irrelevant normal descriptions can mislead the generator, causing it to overlook visual evidence of abnormalities.

To address this problem, we propose *clinical relevance context refinement*, which selectively prunes retrieved contexts based on their ground-truth pathology labels. Given top-ranked retrieved

candidate $t_i^{(d^*)}$ for organ i, we refine the text by filtering out the description associated with normal findings, thereby retaining only clinically significant pathological evidence. Finally, the refined text $\mathcal{R}_i$ is obtained as:

$$\mathcal{R}_i = \begin{cases} t_i^{(d^*)} & \text{if } y_i^{(d^*)} \neq \text{Normal}, \\ \varnothing & \text{otherwise.} \end{cases} \tag{7}$$

3.4 Anatomy-Aware Prompting & Generation

We construct a structured anatomy-aware prompt by concatenating the refined texts in a fixed order. The retrieval prompt $\mathcal{P}$ is formed as follows:

$$\mathcal{P} = \left\{ [\text{TAG}]_i : \mathcal{R}_i \right\}_{i \in \Omega}, \tag{8}$$

where TAG_i is an organ header denoting the anatomical region (e.g., [Lung]:). If the context is filtered out, $\mathcal{R}_i$ is set to $\varnothing$.

First, the visual tokens $\mathbf{X}$ are projected into visual embeddings $\mathbf{V}$ via the compressor $g(\cdot)$. Here, $g(\cdot)$ denotes a *model-specific* token compressor that reduces the token length and projects features to the LLM embedding dimension.

Finally, the generator creates the report W conditioned on the visual embeddings $\mathbf{V}$, the instruction prompt Inst, and the retrieval prompt $\mathcal{P}$:

$$P(W \mid \mathbf{V}, \text{Inst}, \mathcal{P}) = \prod_{t=1}^{T} P(w_t \mid w_{<t}, \mathbf{V}, \text{Inst}, \mathcal{P}) \tag{9}$$

The detailed prompt utilized in our study is described in Appendix G.

4 Experiments

4.1 Data Setup

Dataset & Partitioning. We evaluate our framework on RadGenome-ChestCT (Zhang et al., 2024), which contains 25,692 chest CT volumes from 21,303 patients paired with radiology reports and anatomy-aware textual annotations. As no public test set is provided, we use the official validation split ($N = 1,564$) as a held-out test set. The official training split ($N = 24,128$) is further partitioned into training (90%) and internal validation (10%) subsets at the patient level to prevent data leakage. We use the 18 binary abnormality labels defined in CT-RATE (Hamamci et al., 2024b) as our unified label space for organ-specific auxiliary supervision and downstream evaluation.The detailed label description is provided in Appendix A.

Visual Input Processing. We process the 3D CT volumes using a standardized pipeline consisting of intensity clipping, min–max normalization, and spatial resampling to a unified resolution. Detailed preprocessing steps are provided in Appendix B.

4.2 Experimental Setup

Backbone Architectures. We utilize the pre-trained 3D visual encoders from three state-of-the-art models: CT2Rep (Hamamci et al., 2024a), RadFM (Wu et al., 2025), and M3D (Bai et al., 2024). To ensure a fair comparison, we employ the language decoder as Llama-3.1-8B-Instruct (Grattafiori et al., 2024) for all experimental settings. To demonstrate the model-agnostic versatility of *CPR-RAG*, we employ both the visual encoders and the LLM in frozen, pre-trained states. We restrict training to only the lightweight projection modules and LoRA adapters (Hu et al., 2022). This strategy prevents the model from simply memorizing training data via massive weight updates, suggesting that performance gains are primarily driven by the effective utilization of retrieved context.

RAG Configuration. We utilize the training corpus as the retrieval knowledge base, strictly excluding the query study to prevent data leakage. Retrieval operates at the organ level using FAISS (Johnson et al., 2019). To improve robustness, we incorporate both context dropout and vision modality dropout during training (Yoran et al., 2024; Wang et al., 2023). Specific hyperparameters for retrieval (K, λ) and dropout rates are described in Appendix B.

Training Details. The model is optimized using AdamW (Loshchilov and Hutter, 2019) with a cosine decay schedule. Training is performed with BF16 mixed precision using DeepSpeed (Rajbhandari et al., 2020) ZeRO-2 optimization. Training hyperparameters (learning rate, batch size, etc.) are listed in Appendix B.

4.3 Evaluation Metrics

Text Generation Quality. We employ standard natural language generation (NLG) metrics to evaluate linguistic quality: BLEU-n (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and ROUGE-L (Lin, 2004).

Clinical Efficacy. Following the CT-RATE protocol (Hamamci et al., 2024b), we use the RadBERT

Table 1: **Overall performance on the RadGenome-ChestCT dataset.** We compare state-of-the-art baselines with and without our proposed framework. + *CPR* denotes the integration of our *CPR-RAG*. The *Oracle* row serves as an empirical upper bound where the evidence is selected by ground-truth labels. The best non-oracle results per metric are reported in **bold** text. The micro-averaged clinical efficacy and standard text generation quality metrics are reported.

Model	Clinical efficacy [%]			Text generation quality					
	Precision	Recall	F1	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
CT2Rep	28.71	15.53	20.16	0.593	0.356	**0.250**	**0.198**	0.410	**0.306**
CT2Rep + CPR	29.77	22.32	25.51	0.554	0.333	0.230	0.176	0.429	**0.306**
M3D	27.95	18.65	22.37	0.571	0.328	0.219	0.165	0.400	0.287
M3D + CPR	**43.01**	31.93	36.65	0.565	0.331	0.222	0.167	0.424	0.302
RadFM	26.79	11.14	15.74	**0.606**	**0.361**	0.249	0.193	0.397	0.303
RadFM + CPR	40.09	**42.22**	**41.13**	0.533	0.313	0.208	0.154	**0.433**	0.293
RadFM + CPR (Oracle)	*44.91*	*45.39*	*45.15*	*0.555*	*0.333*	*0.227*	*0.171*	*0.440*	*0.300*

clinical labeler (Yan et al., 2022) to extract findings and evaluate micro-averaged Precision, Recall, and F1 scores. To assess the quality of the retrieved evidence, we focus on finding-level coverage. Specifically, recall@K measures the proportion of ground-truth abnormalities covered by the union of labels in the top-K retrieved documents, while precision@K quntifies the ratio of relevant findings among those retrieved. Detailed definitions are provided in Appendix C.

Human Evaluation. To assess clinical validity, we conduct a blinded side-by-side evaluation with a board-certified radiologist on randomly selected cases ($N = 100$) from test set. The radiologist evaluate generated reports in terms of *completeness*, *correctness*, and *utility*. Detailed evaluation protocols and the specific scoring rubrics are provided in Appendix D.

5 Results and Analysis

5.1 Overall Performance

We evaluate the efficacy of our framework by applying it to three state-of-the-art 3D CT report generation backbones: CT2Rep, M3D, and RadFM. Table 1 presents a quantitative comparison between the standard baselines and our proposed settings (*Baseline + CPR*) on the RadGenome-ChestCT test set. To assess the empirical upperbound of our strategy, we report the *Oracle* performance, where evidence selection is guided by ground-truth labels.

Clinical Utility. Integrating the *CPR-RAG* framework yields consistent improvements in clinical efficacy across all backbones. Most notably, *RadFM + CPR* achieves a F1-score of 41.13, outperforming the *RadFM* baseline (15.74) by a substantial improvement of +25.39 points. We observe that the RadFM baseline suffers from extremely low recall (11.14), indicating a failure to identify sparse pathological findings. However, the *RadFM + CPR* improves the recall to 42.22. This suggests that the retrieved context provides necessary semantic guidance, enabling the recovery of sparse findings that the visual encoder failed to detect.

Linguistic Fluency Regarding NLG metrics, we observe a trade-off pattern BLEU and METEOR. As shown in Table 1, incorporating *CPR-RAG* improves METEOR and ROUGE-L, with the exception of BLEU. These results suggest that the proposed framework enables the capture of correct pathological semantics, even if the descriptions do not match the reference text word-for-word.

Comparison with Oracle. To evaluate upperbound of *CPR-RAG*, we compare *RadFM + CPR* against the *Oracle* setting. Our framework recovers approximately 91% of the Oracle performance, in terms of clinical efficacy. These results suggest that the propsoed *corpus-derived comorbidity prior* provides a robust proxy for ground-truth knowledge and approximates ideal evidence selection without inference-time labels.

5.2 Impact of Comorbidity Prior on Re-ranking

To investigate the efficacy of our *corpus-derived comorbidity prior*, we analyze the recall@1 performance across varying candidate pool sizes $K \in \{1, \ldots, 200\}$. We compare three strategies derived from the scoring function $S(q, d) = \cos(\mathbf{e}_i^{(q)}, \mathbf{e}_i^{(d)}) + \lambda(\mathbf{c}_i \cdot \mathbf{l}_d)$: 1) *visual similarity*

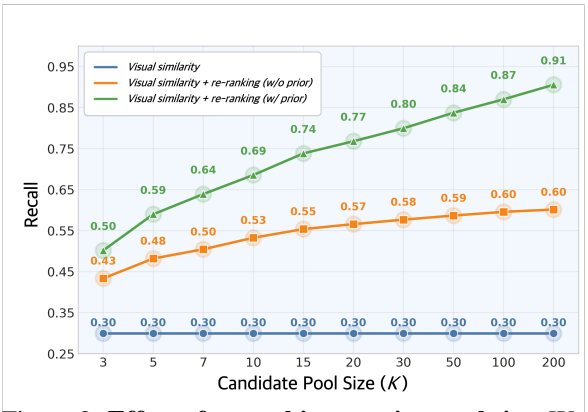

Figure 2: **Effect of re-ranking varying pool size.** We ablate *RadFM+CPR*; which relies on *visual similarity* alone, *CPR-RAG* without prior (*visual similarity + re-ranking (w/o prior)*), and *CPR-RAG* with prior (*visual similarity + re-ranking (w/ prior)*), in terms of recall@1.

(blue), which relies solely on visual similarity ($\lambda = 0$); 2) *visual similarity + re-ranking (w/o prior)* (orange), utilizing raw classifier outputs ($\mathbf{c}_i = \mathbf{p}_i$) without the matrix $\mathbf{M}$; and 3) *visual similarity + re-ranking (w/ prior)* (green), which incorporates the comorbidity prior ($\mathbf{c}_i = \mathbf{p}_i^\top \mathbf{M}_i$).

As shown in Figure 2, the *visual similarity* baseline remains invariant at 0.30, as the top-1 ranking is determined solely by the initial query regardless of pool expansion. The *re-ranking w/o prior* strategy improves performance but saturates at 0.60, suggesting that re-ranking based on raw predictions is limited by classifier noise and sparsity. In contrast, the *re-ranking w/ prior* method demonstrates robust scalability, reaching 0.91 at K=200. This indicates that the comorbidity prior effectively compensates for missing direct evidence by inferring latent pathological associations, thereby recovering relevant candidates even from deep within the retrieval pool.

Table 2: **Ablation study on the clinical relevance context refinement module.** Experiments are conducted using the RadFM backbone. We report micro-averaged clinical efficacy and standard text generation quality metrics. The best values are highlighted in **bold**.

Model	Clinical efficacy [%]			Text generation quality		
	Precision	Recall	F1	BLEU-4	METEOR	BERTScore
RadFM+RAG (w/o refinement)	26.84	27.69	27.26	0.1460	0.4191	0.893
RadFM+RAG (w/ refinement)	**40.09**	**42.22**	**41.13**	**0.1544**	**0.4334**	**0.896**
RadFM+RAG (Oracle) (w/o refinement)	*35.59*	*42.79*	*38.86*	*0.1402*	*0.4276*	*0.894*
RadFM+RAG (Oracle) (w/ refinement)	*44.91*	*45.38*	*45.15*	*0.1712*	*0.4398*	*0.897*

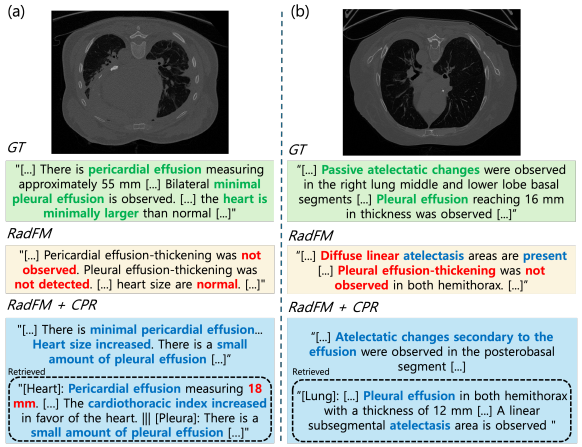

Figure 3: **Qualitative comparison of RadFM vs. *RadFM + CPR*.** Green indicates the ground-truth findings, red marks erroneous statements, and blue highlights correct predictions. Dashed boxes denote the retrieved reference evidence. Two cases are shown: (a) pericardial effusion with cardiomegaly and (b) atelectasis with pleural effusion.

5.3 Effectiveness of Context Refinement

To investigate the effectiveness of filtering, we compare the *CPR-RAG* with and without the *clinical relevance context refinement* module based on RadFM backbone. As shown in Table 2, incorporating *clinical relevance context refinement* always improves performance, not only clinical efficacy but also linguistic fluency. In particular, *CPR-RAG (w/ refinement)* improves the F1-score to 41.13 (+13.87), precision to 40.09 (+13.25), and recall to 42.22 (+13.87). Surprisingly, *CPR-RAG (w/ refinement)* even improves the performance in the *Oracle* setting, where ground-truth labels are utilized for retrieval. These results indicate that minimizing irrelevant normal findings improves the information density of the prompt and generation performance depends on the pathological evidence.

5.4 Qualitative & Human Evaluation

Case Studies. Figure 3 presents two scenarios cases to qualitatively compare between the RadFM and *RadFM + CPR*. In case (a), the RadFM fails to capture both pericardial effusion and the accompanying cardiomegaly, however, *RadFM + CPR* correctly identify both findings by leveraging a retrieved reference containing this specific pattern (bottom dashed box). A similar improvement is observed in Case (b), where the ground truth involves atelectasis complicated by pleural effusion. The RadFM partially detects the atelectasis but fails to identify the concurrent effusion, whereas *RadFM*

Table 3: **Human evaluation results.** The evaluation is conducted by a board-certified radiologist on 100 randomly sampled cases from the test set. Reports are assessed on a 1–5 Likert scale across three aspects: completeness, correctness, and utility. The best scores are highlighted in **bold**.

Model	Completeness	Correctness	Utility
RadFM	2.48	2.30	2.76
RadFM + CPR	**2.93**	**2.45**	**3.09**

Table 4: **Analysis of report-level detection and missed diagnoses.** We compare the rate of generated reports containing no findings ("predicted as healthy") against the ground truth. Values are reported as count (%).

Model	Predicted as healthy	Missed diagnosis (False negative)	False alarm (False positive)
RadFM	748 (47.8%)	661(42.3%)	106(6.8%)
RadFM + CPR	98 (6.3%)	26 (1.7%)	120(7.7%)
CT2Rep	532 (34.0%)	466 (29.7%)	127(8.12%)
CT2Rep + CPR	217 (13.9%)	84 (5.4%)	60(3.8%)
M3D	429 (27.4%)	379 (24.2%)	143(9.1%)
M3D + CPR	364 (23.3%)	211 (13.5%)	40(2.6%)

$*$ Test set (N=1,564) contains 192 (12.3%) healthy cases.

+ CPR successfully captures both findings. Collectively, these examples indicate that *CPR-RAG* utilizes retrieved clinical priors to compensate for visual ambiguity, ensuring that subtle, associated findings are not overlooked. Additional cases are presented in Appendix F.

Human Evaluation. In addition to case studies, we conduct a blinded human evaluation by a board-certified radiologist. In particular, we randomly sample 100 cases from the test set and the radiologist evaluates the reports using a 1–5 Likert scale: (1) *completeness*, measuring the extent to which necessary findings are captured without omission, and (2) *correctness*, assessing the absence of hallucinations or factual errors. (3) *utility*, evaluating the practical readiness of the generated report for clinical workflows (i.e., whether it reduces reporting time or requires rewriting). The detailed evaluation protocols are described in Appendix D.

We report the results in Table 3, showing that the *RadFM + CPR* impoves performance across all metrics compared to the RadFM. In particular, *RadFM + CPR* improves the completeness to 2.93, correctness to 2.45, and utility to 3.09. These results highlight that *CPR-RAG* reduces hallucination and missing findings. Moreover, the utility score indicates that while the generated reports still require physician review and editing, they serve as a helpful draft that effectively reduces the time and effort required for final documentation.

5.5 Analysis of Missed Diagnosis

Table 4 evaluates the report-level detection performance by comparing the rate of all-negative predictions against the ground truth. We observe that all baseline methods demonstrate a high misdiagnosis rate and can be improved by our framework. For example, RadFM classifies 47.7% of cases as all-negative, substantially exceeding the ground-truth prevalence of 12.3%. Surprisingly, the false negative rate (misdiagnosis) is 42.2% with RadFM, however, it drops to 1.7% with RadFM + CPR. Similar patterns are observed with CT2Rep and M3D, demonstrating that the retrieved context effectively prompts the generator to recover findings that are otherwise omitted by the visual encoder.

These findings are important because missing a disease (false negative) is much worse than a false alarm (false positive) in medical screening. In medical screening, the priority is to find every potential patient. Although our approach increases false alarms, this is an acceptable trade-off to ensure that no disease is overlooked. In a real-world workflow, where doctors review the final report, it is safer to be sensitive rather than missing diagnosis.

6 Conclusion

We propose *CPR-RAG*, a model-agnostic, organ-centric retrieval framework, that improves 3D radiology report generation by addressing visual-semantic misalignment and hallucination. In particular, we integrate clinical prior-regularized reranking into retrieval process to address visual-semantic misalignment in high-dimensional volumetric data. Additionally, we filter out irrelevant normal findings from retrived candidate context to maximize information density, which help to clinical efficacy of generated reports.

Our experiments on the RadGenome-ChestCT dataset demonstrate that CPR-RAG improves clinical efficacy across multiple baselines, including RadFM and CT2Rep. The results indicate that leveraging statistical priors allows the model to better identify sparse pathological evidence compared to reliance on visual similarity alone. Furthermore, the ablation studies suggest that maximizing information density by removing boilerplate text effectively guides the generator toward accurate finding descriptions.

Limitations

Despite the promising results, our framework faces several limitations to improve.

Limited External Validity. Although RadGenome-ChestCT includes scans acquired under diverse scanner and protocol settings, it is still derived from a single institution, resulting in relatively consistent reporting style and phrasing. This raises concerns about distribution shift when deploying CPR-RAG to multi-center datasets with different report templates, terminology, and documentation habits. Future work should validate our framework on external cohorts from multiple hospitals and countries, and investigate domain adaptation strategies to improve cross-site generalization.

Gap to Clinical Applicability. While CPR-RAG demonstrates substantial improvement (+25.39 F1) over baselines, the absolute Clinical F1 score ($\sim$41%) remains insufficient for fully autonomous clinical adoption. In safety-critical medical environments, strict adherence to factual accuracy is paramount, and hallucinations pose significant risks. Consequently, improving clinical efficacy is a prerequisite for real-world deployment.

Rigid Report Structuring. Furthermore, our current reporting flow does not reflect the dynamic prioritization of clinical practice. Radiologists typically prioritize acute or critical abnormalities to ensure immediate attention, whereas our system constructs prompts in a fixed, predefined sequence (e.g., Lung $\rightarrow$ Heart), disregarding the relative urgency of pathologies. This static ordering limits the system's ability to highlight the most clinically significant findings first. Extending the framework to handle such prioritization via dynamic planning remains an open challenge.

Dependency on Static Priors. From a modeling perspective, the reliance on comorbidity priors derived solely from training statistics inherently limits generalization to rare, long-tail pathologies. Moreover, the current co-occurrence matrix models only pairwise relationships ($A \rightarrow B$), simplifying the complex, high-order dependencies often observed in medicine. Incorporating external medical knowledge graphs (e.g., UMLS, SNOMED-CT) to capture sophisticated, neuro-symbolic causal relationships remains a crucial direction for future research.

Sensitivity to Error Propagation. Finally, the efficacy of the retrieval module is contingent on the accuracy of upstream organ-specific classifiers. Since the system utilizes predicted probabilities to guide context selection, miscalibrations can propagate through the pipeline, potentially filtering out relevant evidence or injecting noise. Enhancing the robustness of these guidance classifiers to minimize such error propagation is critical for stable deployment.

References

Fan Bai, Yuxin Du, Tiejun Huang, Max Q-H Meng, and Bo Zhao. 2024. M3d: Advancing 3d medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*.

Satanjeev Banerjee and Alon Lavie. 2005. 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*, pages 65–72.

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

Mark Endo, Rayan Krishnan, Viswesh Krishna, Andrew Y. Ng, and Pranav Rajpurkar. 2021. Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In *Proceedings of Machine Learning Research*, volume 158 of *Proceedings of the 6th Machine Learning for Health Conference (MLHC 2021)*, pages 209–219. PMLR.

Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. 2019. A guide to deep learning in healthcare. *Nature Medicine*, 25(1):24–29.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Ibrahim Ethem Hamamci, Sezgin Er, and Bjoern Menze. 2024a. Ct2rep: Automated radiology report generation for 3d medical imaging. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pages 476–486. Springer.

Ibrahim Ethem Hamamci, Sezgin Er, Enis Simsar, Alperen F Tabak, Baris Atac, Ayse Betul Kose, Ecem Masazade, Berke Sevgi, Zeynep Sevgi, Gozde Unal,

and 1 others. 2024b. Ct-rate: A high-fidelity 3d chest ct-report dataset for multi-modal foundation model pretraining. *arXiv preprint arXiv:2403.17834*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, and 1 others. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Seowoo Lee, Jiwon Youn, Hyungjin Kim, Mansu Kim, and Soon Ho Yoon. 2025. Cxr-llava: a multimodal large language model for interpreting chest x-ray images. *European Radiology*, pages 1–13.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9459–9474.

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

Jingyang Lin, Yingda Xia, Jianpeng Zhang, Ke Yan, Le Lu, Jiebo Luo, and Ling Zhang. 2024. Ct-glip: 3d grounded language-image pretraining with ct scans and radiology reports for full-body scenarios. *arXiv preprint arXiv:2404.15272*.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. 2017. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88.

Fenglin Liu, Changchang Yin, Xian Wu, Shen Ge, Ping Zhang, and Xu Sun. 2021. Contrastive attention for automatic chest x-ray report generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 269–280. Association for Computational Linguistics.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Yuren Mao, Wenyi Xu, Yuyang Qin, and Yunjun Gao. 2025. Ct-agent: A multimodal-llm agent for 3d ct radiology question answering. *arXiv preprint arXiv:2505.16229*.

Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. 2021. Improving factual completeness and consistency of image-to-text radiology report generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5288–5304. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. 2019. Transfusion: Understanding transfer learning for medical imaging. In *Advances in Neural Information Processing Systems*, volume 32.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE.

Mercy Ranjit, Gopinath Ganapathy, Ranjit Manuel, and Tanuja Ganu. 2023. Retrieval augmented chest x-ray report generation using openai gpt models. In *Proceedings of Machine Learning Research*, volume 219 of *Proceedings of the 8th Machine Learning for Health Conference (MLHC 2023)*, pages 650–666. PMLR.

Arun James Thirunavukarasu, Daniel Shu Wei Ting, Kallisvaran Elangovan, Laura Gutierrez, Ting Fang Tan, and David Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, and 1 others. 2024. Towards generalist biomedical ai. *NEJM AI*, 1(3):AIoa2300138.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Mohammed, Saksham Singhal, Subhojit Som, and 1 others. 2023. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19175–19186.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. 2025. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Nature Communications*, 16(1):7866.

An Yan, Yu Wang He, and Chun-Nan McAuley, Julian andXHsu. 2022. Radbert: Adapting transformer-based language models to radiology. *Radiology: Artificial Intelligence*, 4(4):e210258.

Sixing Yan, William K. Cheung, Ivor W. Tsang, Keith Chiu, Terence M. Tong, Ka Chun Cheung, and Simon See. 2024. Ahive: Anatomy-aware hierarchical vision encoding for interactive radiology report retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14324–14333.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. Making retrieval-augmented language models robust to irrelevant context. In *International Conference on Learning Representations (ICLR)*.

Tengfei Zhang, Ziheng Zhao, Chaoyi Wu, Xiao Zhou, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2025. Radir: A scalable framework for multi-grained medical image retrieval via radiology report mining. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, pages 508–518. Springer.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Jiayu Lei, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Radgenome-chest ct: A grounded vision-language dataset for chest ct analysis. *arXiv preprint arXiv:2404.16754*.

Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D. Manning, and Curtis Langlotz. 2020. Optimizing the factual correctness of a summary: A study of summarizing radiology reports. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5108–5120.

Appendix

A Disease Labels

We utilize the 18 binary abnormality labels inherited from the CT-RATE corpus (Hamamci et al., 2024b). Based on the RadGenome-ChestCT (Zhang et al., 2024) schema, these labels are mapped to five anatomical regions, as shown in Table 5.

Table 5: **List of abnormality labels.** Mapped to anatomical regions based on the dataset annotations. Note that *Medical material* is categorized as 'Others' as it comprises heterogeneous devices appearing across multiple anatomical regions (e.g., pacemakers, tubes, clips).

Anatomical Region	Abnormality Labels
Lungs	Lung Nodule, Lung Opacity, Atelectasis, Consolidation, Emphysema, Pulmonary Fibrotic Sequela, Mosaic Attenuation Pattern, Interlobular Septal Thickening
Heart	Cardiomegaly, Pericardial Effusion, Coronary Artery Wall Calcification, Arterial Wall Calcification
Trachea and Bronchie	Bronchiectasis, Peribronchial Thickening
Mediastinum	Lymphadenopathy, Hiatal Hernia
Pleura	Pleural Effusion
Others	Medical Material

B Implementation Details

Data Preprocessing. We preprocess collected 3D CT scans the following steps: 1) **Intensity Clipping:** Hounsfield Unit (HU) values are truncated to $[-1000, 1000]$. 2) **Normalization:** Voxel intensities are linearly scaled to $[0, 1]$. 3) **Resampling:** Volumes are resized to $256 \times 256 \times 64$ via trilinear interpolation.

Hyperparameters We provide the detailed hyperparameters used for training and our RAG module in Table 6.

C Retrieval Metrics

We formally define the retrieval metrics used in the main text. Let L_q be the set of ground-truth labels for a query, and $\hat{L}_K = \bigcup_{d \in \mathcal{R}_K} L(d)$ be the union

Table 6: **Hyperparameter Settings.**

Hyperparameter	Value
Training Optimization	
Optimizer	AdamW
Learning Rate	$5e^{-5}$
Weight Decay	0.01
Batch Size (per device)	4
Gradient Accumulation	4
Warmup Steps	100
Epochs	5
Precision	BF16 (DeepSpeed ZeRO-2)
PEFT (LoRA)	
Rank (r)	8
Scaling (α)	32
LoRA Dropout	0.05
Target Modules	Self-attention layers
RAG Module	
Retrieval Candidates (K)	100
Reranking Weight (λ)	0.3
Context Dropout (p)	0.25
Visual Dropout (p)	0.20
Similarity Metric	Inner Product (ℓ_2-norm)

of labels in the top-K retrieved documents.

$$\text{Recall@K} = \frac{|L_q \cap \hat{L}_K|}{|L_q|} \tag{10}$$

$$\text{Precision@K} = \frac{|L_q \cap \hat{L}_K|}{|\hat{L}_K|} \tag{11}$$

These metrics quantify the trade-off between coverage and noise in the retrieved context.

D Human Evaluation Details

Evaluation Protocol The human evaluation is conducted in a blinded, side-by-side manner. We randomly sampled $N = 100$ cases from the held-out test set. For each case, we provide the original ground-truth report to the board-certified radiologist to serve as the reference standard. Subsequently, two generated reports labeled anonymously as "Report A" and "Report B," are provided to the radiologist. To prevent potential bias, the assignment of models (e.g., RadFM and *RadFM + CPR*) to labels A and B is randomized for every case, ensuring that the evaluator remains unaware of the source of each report.

Scoring Rubric The radiologist evaluated each report using a 5-point Likert scale. The specific criteria for *Completeness*, *Correctness*, and *Utility* are defined in Table 7.

Table 7: **Human evaluation rubric.** The scoring criteria and definitions used by the board-certified radiologist are provided.

Criteria	Score & Definition
Completeness	**5 (Excellent):** Covers all critical and incidental findings present in the image. **3 (Fair):** Misses minor incidental findings but captures key pathologies. **1 (Poor):** Misses critical abnormalities or major pathologies.
Correctness	**5 (Accurate):** No factual errors or hallucinations. **3 (Acceptable):** Contains minor errors that do not alter the clinical diagnosis. **1 (Wrong):** Contains critical hallucinations or false positives.
Utility	**5 (Ready):** Can be used as a final report without modification. **3 (Helpful):** Requires minor editing but reduces reporting time. **1 (Useless):** Requires complete rewriting; potentially misleading.

E Analysis of organ-specific embeddings

To validate the anatomy-conditional learning, we visualize the organ-specific embeddings using t-SNE. As illustrated in Figure 4, the embeddings form distinct, well-separated clusters corresponding to anatomical regions (e.g., Lung, Heart, Mediastinum). This clear separation confirms that the learnable queries effectively disentangle the global 3D features into organ-specific representations, thereby minimizing cross-organ semantic interference during the retrieval phase.

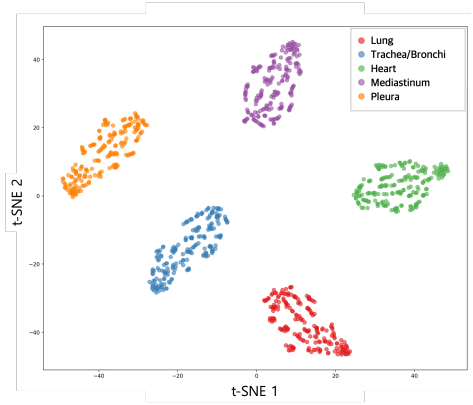

Figure 4: **Visualization of organ-specific embeddings.** The t-SNE visualization demonstrates that the learned embeddings successfully disentangle high-dimensional 3D features into distinct anatomical regions.

Success Case: Detection of Small Nodules (Case ID: `valid_1001_a_1`)

Target Finding: Detection of millimetric pulmonary nodules.

[Ground Truth (GT)]
"... **There are millimetric nonspecific nodules in both lungs.** ... No pathologically enlarged lymph nodes were observed."

[Baseline (RadFM)]
"... Aeration of both lung parenchyma is normal and **no nodular or infiltrative lesion is detected** ..."
→ *False Negative (misses nodules).*

[Retrieved Evidence (CPR; Top-1)]
Retrieved: "[LUNG]: A nonspecific nodule measuring 5×3 mm is observed at the laterobasal level."

[RadFM + CPR (Ours)]
"... No mass or infiltrative lesion ... **There are millimetric nonspecific nodules in both lungs.** ..."

Analysis: Retrieval provides a nodule-positive reference, enabling our model to recover GT-positive nodules that the baseline misses.

Failure Case: Retrieval-induced False Positive (Case ID: `valid_1142_a_1`)

Target Finding: Pulmonary nodules (should be **absent** in GT).

[Ground Truth (GT)]
"**No suspicious mass or nodular space-occupying lesion** was observed in the lung parenchyma."

[Baseline (RadFM)]
"... Aeration of both lung parenchyma is normal and **no nodular or infiltrative lesion is detected** ..."
→ *True Negative.*

[Retrieved Evidence (CPR; Top-1)]
Retrieved: "[LUNG]: **Millimetrically sized nonspecific nodules are observed** in both lungs."

[RadFM + CPR (Ours)]
"A few stable **non-specific pulmonary nodules with diameters less than 3 mm are observed** in both lungs."

Analysis: Noisy retrieved context introduces a positive nodule cue, which leads to a retrieval-induced false positive.

Figure 5: **Additional qualitative analysis. (Top)** Retrieval helps recover GT-positive nodules missed by the baseline. **(Bottom)** Noisy retrieval induces a false positive in the generated report.

F Additional Qualitative Analysis

In this section, we provide additional qualitative examples to offer a comprehensive view of our method's behavior. In Figure 5, the success case demonstrates how retrieving clinically relevant evidence allows the model to capture fine-grained details and incidental findings (e.g., specific nodule sizes) that are often overlooked by the baseline. Conversely, the failure case illustrates a limitation of RAG systems, where the retrieval of noisy or irrelevant context can mislead the generator.

G Prompt Details

We utilize the standard LLaMA-3.1 chat template to construct the input prompt for our CPR-RAG-based radiology report generation. The full prompt structure is illustrated in Figure 6.

Visual Prompt. Following prior multimodal LLM baselines, we represent the input CT scan as a fixed-length sequence of visual embeddings $\mathbf{V} = \{\mathbf{V}_i\}_{i=1}^{L}$, where $L = 32$ in our experiments. These visual embeddings are prepended to the textual chat prompt and serve as the image-conditioned prefix for the LLM (see Figure 6).

CPR-augmented Prompt Template (LLaMA-3.1 Chat Format)

[Visual Prompt]
$[\mathbf{V}_1][\mathbf{V}_2]\cdots[\mathbf{V}_{32}]$

[System]
<|begin_of_text|><|start_header_id|>system<|end_header_id|>
You are an expert radiologist. Generate a CT finding report.
Use the retrieved similar cases as important references for your diagnosis.
If similar cases show abnormalities, carefully check if the current scan has similar findings.
<|eot_id|>

[User + Retrieved Context]
<|start_header_id|>user<|end_header_id|>
Similar cases from database:
[LUNG]: {$\mathcal{R}_{\text{lung}}$}
[HEART]: {$\mathcal{R}_{\text{heart}}$}
[MEDIASTINUM]: {$\mathcal{R}_{\text{mediastinum}}$}
[PLEURA]: {$\mathcal{R}_{\text{pleura}}$}
[TRACHEA]: {$\mathcal{R}_{\text{trachea}}$}
Based on the CT image and the similar cases above, generate the finding report.
<|eot_id|>

[Assistant]
<|start_header_id|>assistant<|end_header_id|>
{Generated Report}

Figure 6: The prompt template used for CPR-augmented report generation. Visual embeddings $\mathbf{V}_i$ are prepended to the text prompt. Retrieved contexts $\mathcal{R}_{\text{organ}}$ are injected per anatomical region.

To ensure professional generation quality, we assign the role of an "expert radiologist" in the system instruction. Overall, the prompt consists of four main components:

- **Visual Prompt:** A sequence of visual embeddings $\mathbf{V}_1, \ldots, \mathbf{V}_{32}$ prepended to the input, providing image evidence to the LLM.
- **System Instruction:** Defines the task and persona (expert radiologist).
- **Retrieved Context:** We inject organ-specific retrieved similar cases into placeholders (e.g., $\mathcal{R}_{\text{lung}}$, $\mathcal{R}_{\text{heart}}$). Each placeholder is replaced with the top-1 retrieved text for the corresponding anatomical region.
- **Task Instruction:** Directs the model to generate the final finding report based on the CT image and the retrieved references.

Special tokens such as <|begin_of_text|> and <|start_header_id|> are used to strictly follow the instruction-tuning format of LLaMA-3.1.