SeMLI-RAG: Semantics-Guided Multimodal Late Interaction for Retrieval-Augmented Radiology Report Generation

Anonymous EMNLP submission

Abstract

The consultation of reference cases in clinical radiology reporting motivates integrating multimodal retrieval-augmented generation (RAG) into automated systems. However, traditional RAG approaches employ single-vector embeddings for indexing, resulting in loss of granular information. Moreover, a semantic gap exists whereby visually similar radiographs often encode distinct pathologies. In this paper, we propose a **Se**mantics-guided **M**ultimodal **L**ate **I**nteraction **RAG** (SeMLI-RAG) framework for radiology report generation. Our approach employs a weight-based late interaction strategy for retrieval to capture fine-grained radiographic features. Furthermore, we introduce an image-adaptive RAG to selectively leverage retrieved information, mitigating performance degradation caused by retrieval-induced noise. Experiments on three radiology benchmarks demonstrate that our framework yields consistent improvements over the backbone and RAG baselines across multiple evaluation dimensions. Further analysis shows that improving retrieval granularity and retrieval robustness can both benefit radiology report generation.

1 Introduction

Radiologists often consult similar historical cases during report writing to ground their interpretation and reduce diagnostic uncertainty. Retrieval-augmented generation (RAG) offers a natural computational analogue by leveraging external evidence to support report generation and mitigate hallucinations in large language models (LLMs) (Lewis et al., 2020; Amiraz et al., 2025). However, in radiology, this paradigm introduces two key challenges. First, many existing RAG approaches employ single-vector representations for retrieval, which may miss fine-grained semantic and spatial correspondences critical for diagnosis (Khattab and Zaharia, 2020; Huy et al., 2025). Second, a semantic gap exists between visual similarity and clinical meaning: radiographs that appear similar at the pixel level may correspond to distinct pathologies, leading to retrieval outputs that are clinically mismatched and introduce noise into generation (Oakden-Rayner et al., 2020; Schaefer-Prokop et al., 2008).

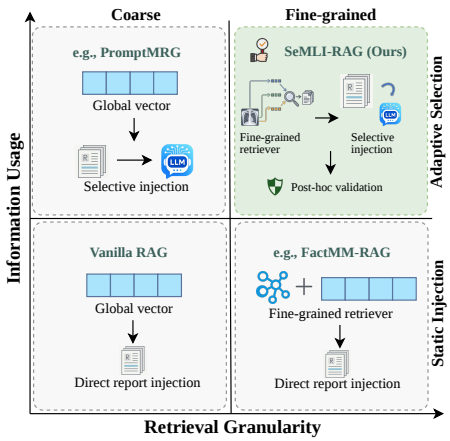


Figure 1: Comparison of representative RAG-based methods with our SeMLI-RAG, including Vanilla RAG, PromptMRG (Jin et al., 2024) and FactMM-RAG (Sun et al., 2025).

Existing multimodal RAG approaches for radiology address these challenges only partially. On the retrieval side, models such as MedCLIP (Wang et al., 2022) and CXR-CLIP (You et al., 2023) rely primarily on global single-vector representations. While these are effective for coarse alignment, they are insufficient for capturing localized pathological patterns, and theoretically limited in retrieval capability (Weller et al., 2025). On the generation side, prior work either directly injects retrieved reports into the prompt (Ranjit et al., 2023; Choi et al., 2025) or extracts auxiliary signals to guide generation more indirectly (Huang et al., 2025). As illustrated in Figure 1, these approaches differ along two axes —retrieval granularity and information usage—but none simultaneously achieves fine-grained retrieval, selective integration of evidence, and robustness when retrievals are clinically misaligned.

To address these limitations, we propose a **S**emantics-guided **M**ulti-modal **L**ate-**I**nteraction **RAG** (SeMLI-RAG) framework for radiology report generation. Our approach combines three components. First, we introduce a weighted late-interaction retriever that preserves fine-grained token- and region-level correspondences while dynamically emphasizing clinically informative image regions. Second, we design an image-adaptive RAG mechanism that employs image-conditioned latent queries to selectively extract useful information from retrieved reports, mitigating the impact of noisy or weakly aligned evidence. Finally, we incorporate a lightweight post-hoc validation mechanism that compares candidate reports generated with and without retrieval and selects the one better aligned with the input radiograph, improving overall robustness.

Experimental results on IU X-Ray (Demner-Fushman et al., 2016), ROCOv2 (Rückert et al., 2024), and MIMIC-CXR (Johnson et al., 2024) demonstrate that SeMLI-RAG consistently surpasses a strong backbone model as well as vanilla RAG baselines built upon recent medical retrieval models. The gains are particularly evident in clinical accuracy, where SeMLI-RAG consistently improves the MedCAT Concept F1 score (Kraljevic et al., 2021) across all three datasets and outperforms the backbone by a relative 5% in macro-F1 on MIMIC-CXR. These results indicate that improving retrieval granularity and robustness to noisy evidence is critical for reliable radiology report generation.

Our main contributions are as follows:

- We propose a semantics-guided and multimodal late-interaction (**SeMLI**) retriever, which retrieves both reference radiology images and reports based on fine-grained token- and region-level matching.
- We design an image-adaptive RAG framework with selective evidence use and post-hoc validation to improve robustness against noisy or clinically mismatched retrieval.
- We demonstrate on three benchmark datasets that **SeMLI-RAG** consistently improves over a strong backbone and outperforms conventional RAG. Qualitative case studies further show the effectiveness of our framework.

2 Related works

Radiology Report Generation is the task of automatically producing clinically coherent textual reports from medical images, such as chest X-rays, to assist radiologists in describing findings and impressions (Mamdouh et al., 2025). Existing methods fall mainly into two categories: conventional encoder-decoder frameworks that directly map visual features to reports, and recent LLM-based methods that leverage stronger language priors for more fluent and detailed generation (Wang et al., 2025). Earlier representative approaches typically adopt CNN- or Transformer-based visual encoders with recurrent or Transformer decoders to learn image-to-text generation in an end-to-end manner (Liu et al., 2021; Yang et al., 2021; Chen et al., 2020; Hou et al., 2021). More recently, LLM-based frameworks improve report quality by aligning radiology image features with frozen general-purpose language models (Huang et al., 2025; Wang et al., 2023). In parallel, retrieval-augmented based methods further enhance generation by introducing similar historical cases as auxiliary references (Jin et al., 2024; Sun et al., 2025), aiming to improve factual accuracy and clinical relevance.

Retrieval-Augmented Generation is a framework that combines parametric language models with an external retrieval system to access relevant information during generation. RAG is primarily motivated by mitigating hallucinations in LLMs while also improving factual foundation and access to up-to-date domain knowledge (Gao et al., 2023; Fan et al., 2024). Existing RAG methods can be broadly categorized into naive RAG, which directly retrieves and prepends external documents to the prompt (Lewis et al., 2020), and more advanced variants that introduce retrieval refinement (Chan et al., 2024), iterative retrieval (Jiang et al., 2023), adaptive selection (Asai et al., 2024), or modular designs (Gao et al., 2024) for better integration of evidence. In the medical domain, RAG has been increasingly explored to improve the reliability of clinical question answering and decision support by grounding LLM outputs against external biomedical knowledge sources (He et al., 2025).

3 Methodology

In this section, we present the overall framework of SeMLI-RAG, as illustrated in Figure 2, and detail its three primary modules: *i*) SeMLI retriever (3.1); *ii*) image-adaptive RAG (3.2); and iii) post-hoc

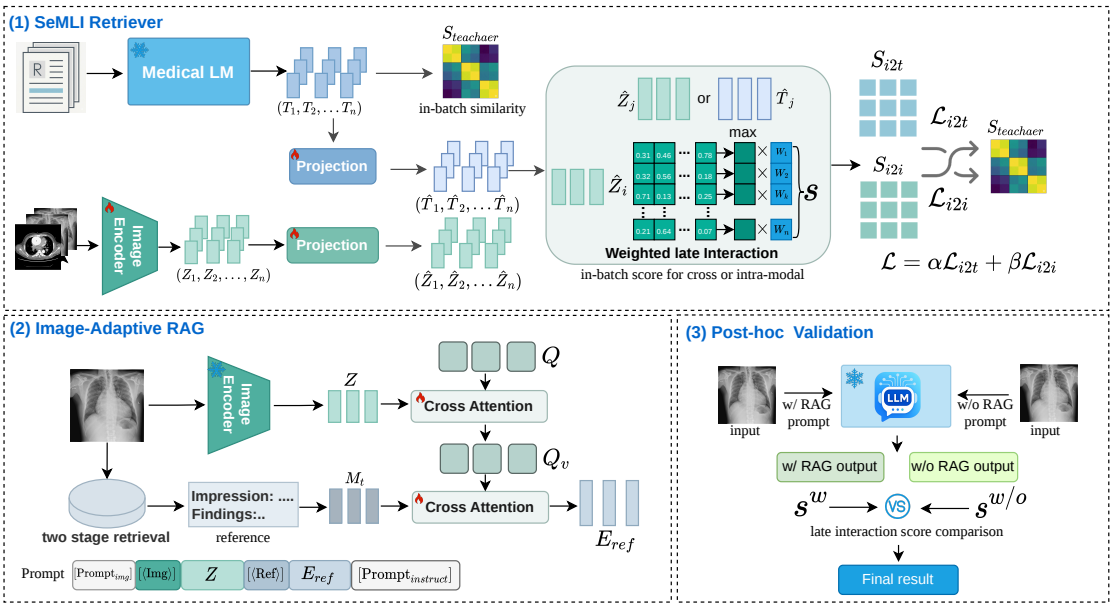


Figure 2: Overview of the proposed framework. (1) The SeMLI retriever performs fine-grained alignment between visual patches and token-level textual features via weighted late interaction. (2) The image-adaptive RAG module injects retrieved information through cross-attention. (3) The post-hoc validation selects the final report by comparing similarity scores of outputs with and without retrieval. Snowflake and flame icons denote frozen and trainable modules, respectively.

validation (3.3).

3.1 SeMLI Retriever

3.1.1 Dynamically weighted late interaction

For an input radiology image X_r, a pre-trained vision encoder $f_I(\cdot)$ is used to generate a group of vision features

$$Z = f_I(X_r), \tag{1}$$

where $Z \in \mathbb{R}^{q \times d_z}$, q is the number of patches and d_z is the image feature dimension.

To construct the clinical semantic space for our multimodal retrieval distillation, we process the corresponding radiology report r using a specialized, pre-trained medical language model $f_R(\cdot)$ as a text encoder to generate token-level embeddings

$$T = f_R(r), \tag{2}$$

where $T \in \mathbb{R}^{L \times d_t}$, L is the sequence length and d_t is the hidden dimension.

To align the modalities, we project both the visual features Z and textual features T into a shared low-dimensional latent space of length d_p (e.g., $d_p = 128$) using residual multi-layer perceptrons. This yields the L_2-normalized representations, denoted as $\hat{Z} \in \mathbb{R}^{q \times d_p}$ and $\hat{T} \in \mathbb{R}^{L \times d_p}$ respectively, which are subsequently used for fine-grained retrieval.

Traditional retrieval methods typically compress multimodal inputs into single-vector embeddings, inherently leading to the loss of granular information and theoretical performance limitations (Weller et al., 2025). Although multi-vector retrieval paradigms have achieved success in text retrieval (Khattab and Zaharia, 2020), applying them directly to vision tasks often treats all spatial regions equally (Reddy et al., 2025). This ignores the clinical reality that pathological findings are frequently highly localized within minor, yet critical, areas of a radiology image. To address this, our SeMLI retriever introduces a dynamically weighted late interaction scoring mechanism.

For image-to-image similarity, we formulate the weighted late interaction as follows:

$$s(X_i, X_j) = \frac{\sum_{k=1}^{q} W_k \cdot \max_l(\hat{Z}_{i,k} \cdot \hat{Z}_{j,l}^{\top})}{\sum_{k=1}^{q} W_k}, \tag{3}$$

where $\hat{Z}_{i,k}$ represents the projected feature of the k-th patch from the image X_i, and $\hat{Z}_{j,l}$ denotes the l-th patch from the image X_j. W_k is the semantic weight assigned to the k-th patch of X_i, dynamically highlighting the most informative regions, while $\max_l$ captures the highest similarity match for that patch across all spatial regions of X_j.

This weighted late interaction can be directly extended to image-to-text retrieval by replacing the patch embeddings of the target image, $\hat{Z}_{j,l}$, with the projected token embeddings of the report, $\hat{T}_{j,l}$. Specifically, $s(X_i, T_j)$ is obtained by matching each weighted visual patch of X_i to its most

similar textual token in T_j.

To dynamically determine the semantic weight W_k for each image patch, we introduce a set of learnable semantic anchors $A \in \mathbb{R}^{m \times d_a}$, which act as a continuous memory bank representing common radiological findings. To assign higher importance to visual patches that strongly align with these clinical concepts anchors, we project the image features Z via a multi-layer perceptron $P_w(\cdot)$ and calculate their maximum cosine similarity against all anchors. The final semantic weights are then derived through a temperature-scaled *Softmax* function:

$$
\begin{aligned}
r_{i,k} &= \max_{c \in \{1,\ldots,m\}} \left(\tilde{Z}_{i,k} \cdot \hat{A}_c^\top \right), \\
W_{i,k} &= \frac{\exp(r_{i,k}/\tau_w)}{\sum_{j=1}^{q} \exp(r_{i,j}/\tau_w)}, \quad k \in \{1, \ldots, q\},
\end{aligned}
\tag{4}
$$

where $\tilde{Z} = \text{Norm}(P_w(Z))$ and $\hat{A} = \text{Norm}(A)$ denote the L_2-normalized projected patch features and anchors, respectively, and τ_w is a temperature hyperparameter controlling the sharpness of the attention distribution.

3.1.2 Teacher–Student Distillation

To align clinical semantic information during multimodal retrieval, we introduce a teacher-student distillation paradigm. In this teacher–student setup, a pre-trained medical language model (e.g., CXR-BERT (Boecking et al., 2022)) provides semantic supervision.

Specifically, we compute the in-batch similarity matrix using the pre-trained teacher model to serve as the target distribution for distillation. A medical language model encodes reports into token-level features, where the `[CLS]` token embedding is used as the sentence representation. Pairwise cosine similarities among the batch (with size B) form a similarity matrix $S_{teacher} \in R^{B \times B}$. To compel the student retriever to focus on capturing fine-grained differences among highly similar reports, we filter the teacher's distribution by masking similarities below a threshold κ to zero, thereby eliminating noise from irrelevant in-batch negatives.

Then we compute loss based on image-to-text and image-to-image weighted late interaction. The cross-modal similarity matrix $S_{i2t} \in \mathbb{R}^{B \times B}$ is formulated such that each element

$$
S_{i2t}^{(i,j)} = s(X_i, T_j), \tag{5}
$$

represents the interaction score between the i-th image and the j-th report in a batch. Similarly, the intra-modal image-to-image similarity matrix S_{i2i} is calculated, where each element is defined as $S_{i2i}^{(i,j)} = s(X_i, X_j)$.

We train the student network to mimic this semantic space via *Cross-Entropy* (CE) loss. To ensure bidirectional alignment between vision and language, we calculate a symmetric cross-modal distillation loss

$$
\begin{aligned}
\mathcal{L}_{i \to t} &= \text{CE}(S_{i2t}/\tau, S_{teacher}) \\
\mathcal{L}_{t \to i} &= \text{CE}(S_{i2t}^\top/\tau, S_{teacher}^\top) \\
\mathcal{L}_{i2t} &= \frac{1}{2}(\mathcal{L}_{i \to t} + \mathcal{L}_{t \to i}),
\end{aligned}
\tag{6}
$$

where τ is the temperature scaling factor.

To enable our SeMLI retriever to perform visual retrieval that is highly consistent with clinical semantics, we also apply an intra-modal distillation loss

$$
\mathcal{L}_{i2i} = \text{CE}(S_{i2i}/\tau, S_{teacher}). \tag{7}
$$

The primary distillation objective for the retriever is formulated as the weighted sum of these two losses

$$
\mathcal{L} = \alpha \mathcal{L}_{i2t} + \beta \mathcal{L}_{i2i}, \tag{8}
$$

where α and β are hyperparameters used to balance the cross-modal and intra-modal learning objectives. To facilitate smooth gradient backpropagation, during the training phase, the non-differentiable $\max$ function in the late interaction calculation is replaced with its smooth approximation, the *LogSumExp* function.

3.1.3 Indexing and Retrieval

In real-world clinical practice, radiologists primarily rely on relevant historical radiology images for diagnostic reference. Furthermore, since large-scale medical datasets paired with high-quality textual reports are scarce, a pure image-based retrieval paradigm ensures much broader applicability across diverse clinical settings. Thus, while trained with image-to-text supervision, our retriever performs cross-patient image-to-image retrieval at inference time.

As directly computing the patch-level weighted late interaction across a large-scale database is computationally expensive and time-consuming, we adopt an efficient two-stage retrieval pipeline. During the indexing phase, we store the full projected visual patches $\hat{Z}$ and their corresponding semantic weights W. Simultaneously, to build an

efficient index, we compress the patches into a global representation z_{global} via weighted mean pooling: $z_{global} = \text{Norm}(\sum_k W_k \cdot \hat{Z}_k)$. During the retrieval phase, we first conduct a coarse search. The query image's global vector is used to perform a rapid maximum inner-product search (e.g., via FAISS) against the database to quickly filter out irrelevant samples and retrieve a small candidate pool. In the second stage, we perform fine-grained re-ranking. The precise weighted late interaction score $s(X_{query}, X_{candidate})$ is computed within this small pool using the preserved patch features $\hat{Z}$.

3.2 Image-Adaptive RAG

To effectively utilize the retrieved case without overwhelming the backbone model with raw, potentially noisy text, we introduce an image-adaptive RAG module. This module dynamically controls the focus and amount of information extracted from the reference report based on the current image.

Given a retrieved record R_{sim}, we encode the respective reference report using the input embedding layer of the frozen LLM and organize the resulting token embeddings as a textual memory bank $M_t \in \mathbb{R}^{L \times d_h}$, with L denoting the sequence length and d_h the LLM hidden dimension. To extract the most relevant clinical insights from M_t, we initialize a set of learnable latent queries $Q \in \mathbb{R}^{N_q \times d_h}$, where $N_q = 32$ is the number of query tokens. We design a two-stage sequential interaction mechanism using standard Transformer decoder layers, where the queries act as information probes:

1. **Vision Interaction**: This step aims to allow latent queries Q to read the radiology image and extract important visual features. Using a Transformer decoder, the queries Q cross-attend to current image features Z, yielding a set of vision-aware queries Q_v:

$$Q_v = \text{TransformerDecoder}_{vision}(Q, Z). \quad (9)$$

2. **Text Interaction**: The vision-aware queries Q_v then serve as the target sequence to cross-attend to the referenced text memory M_t, extracting supplementary clinical details E_{ref}:

$$E_{ref} = \text{TransformerDecoder}_{text}(Q_v, M_t). \quad (10)$$

Finally, we construct a unified multimodal continuous prompt to feed into the backbone LLM. The prompt is structured as follows:

$$\text{Prompt} = [\text{Prompt}_{img}] \oplus [\langle\text{Img}\rangle] \oplus Z \oplus [\langle\text{Ref}\rangle] \oplus E_{ref} \oplus [\text{Prompt}_{instruct}], \quad (11)$$

where $\oplus$ denotes sequence concatenation along the token dimension, and $[\cdot]$ represents the embedded text instruction templates. The backbone LLM is kept strictly frozen, and we exclusively optimize the learnable parameters within the image-adaptive RAG module.

3.3 Post-hoc Validation

Since projected image features $\hat{Z}$ are trained to align with semantic features, our framework naturally supports a post-hoc validation mechanism. As illustrated in Figure 2, we leverage the cross-modal capabilities of our trained SeMLI retriever to independently evaluate candidate outputs. Specifically, we collect two distinct candidate reports for each query image: one generated by the raw LLM backbone and another generated via our image-adaptive RAG module. Both candidate reports are encoded by the frozen medical language model into token-level embeddings. Finally, we compute the weighted late-interaction similarity scores for the outputs generated with and without image-adaptive RAG prompt, denoted as s^w and $s^{w/o}$, respectively. The output associated with the higher score is selected as the final report.

4 Experiment Setup

4.1 Dataset

We evaluate our SeMLI-RAG framework on three publicly available benchmarks: MIMIC-CXR (Johnson et al., 2024), IU X-Ray (Demner-Fushman et al., 2016), and ROCOv2 (Rückert et al., 2024).

MIMIC-CXR is a large-scale dataset consisting of 377,110 chest X-ray images paired with 227,827 corresponding free-text radiology reports. Similarly, IU X-Ray is a widely adopted smaller dataset comprising 7,470 chest X-ray images associated with 3,955 clinical reports. ROCOv2 is a diverse medical imaging dataset that contains a variety of imaging modalities—including MRI, CT, Ultrasound, and X-ray—and covers a comprehensive range of anatomical regions. Since these first two datasets are exclusively focused on chest radiography, we additionally evaluate our SeMLI-RAG on the ROCOv2 dataset to demonstrate its robustness and cross-domain generalizability.

4.2 Evaluation Metrics

We evaluate SeMLI-RAG using traditional natural language generation (NLG) metrics, semantic metrics, and medically-tailored evaluation metrics. Specifically, we employ BLEU-1, BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), and CIDEr (Vedantam et al., 2015) to measure lexical overlap and sequence fluency; BERTScore (Zhang et al., 2020), BLEURT (Sellam et al., 2020) to assess semantic equivalence; and MedCAT (Concept F1) (Kraljevic et al., 2021) to evaluate the clinical accuracy of the generated diagnostic entities.

We further evaluate the *clinical efficacy* of generated reports using the CheXbert labeler (Smit et al., 2020), which converts reports into 14 disease labels. We report Micro-F1 and Macro-F1 as the main metrics to assess diagnostic correctness. Since CheXbert is designed for chest X-ray reports, this evaluation is only conducted on MIMIC-CXR.

4.3 Baseline

We adopt R2GenGPT (Wang et al., 2023) as our backbone, as it provides a simple yet effective LLaMA-based framework for radiology report generation. We evaluate SeMLI-RAG against the backbone and several vanilla RAG baselines equipped with representative multimodal retrievers, including CLIP (Radford et al., 2021), MedCLIP (Wang et al., 2022), CXR-CLIP (You et al., 2023), and BiomedCLIP (Zhang et al., 2024). These retrievers cover general-domain, medical-domain, chest-X-ray-specific, and biomedical foundation model pretrained on large-scale image-text data. Detailed training settings are provided in Appendix A.1.

5 Evaluation Results

5.1 Overall Performance

Table 1 presents the experimental results on IU X-Ray, ROCOv2, and MIMIC-CXR. Overall, the SeMLI-RAG framework delivers consistent gains over a strong backbone in all three benchmarks, with the clearest improvements appearing in NLG and clinical metrics. In particular, the gains are most pronounced on ROCOv2, where BLEU-1 and ROUGE-L improve by 9.1% and 4.0% in relative terms, respectively, while MedCAT increases by 7.7% in relative terms. Even on the more challenging MIMIC-CXR benchmark, our framework still improves BLEU-1 and BLEU-4 by 6.6% and 5.6% in relative terms, together with a 2.9% relative gain in MedCAT.

We also compare SeMLI-RAG with published state-of-the-art (SOTA) methods, including PromptMRG (Jin et al., 2024) and FactMM-RAG (Sun et al., 2025), on overlapping datasets and shared metrics. The results show that SeMLI-RAG remains competitive with these methods. Since these methods compute clinical efficacy metrics using different label scopes and evaluation protocols, we do not list their CE scores in Table 1 for direct comparison.

Although the improvements on semantic metrics are more modest, they remain consistently positive across all datasets, indicating that our framework improves not only lexical generation quality but also semantic alignment and medical consistency. Given the strength of the backbone, such cross-dataset gains provide strong evidence for the effectiveness of SeMLI-RAG.

The clinical efficacy metrics on MIMIC-CXR further demonstrate the robustness of SeMLI-RAG. Although vanilla RAG with strong chest X-ray retrievers, such as MedCLIP and CXR-CLIP, can sometimes bring marginal gains on semantic metrics, these gains do not consistently translate into better medical accuracy. For example, on MIMIC-CXR, MedCLIP and CXR-CLIP achieve semantic scores that are relatively close to the backbone, but still show noticeable declines on MedCAT and clinical efficacy metrics. This suggests that vanilla RAG is more vulnerable to noisy retrievals or retrieved evidence that is only partially clinically relevant.

In contrast, SeMLI-RAG improves both semantic quality and clinical accuracy, achieving the best MedCAT score and further improving clinical efficacy performance by about 2.0% in Micro-F1 and 5.1% in Macro-F1 over the backbone. This trend is consistent with the dataset scale: on the smaller IU X-Ray benchmark, vanilla RAG remains competitive on NLG, semantic, and medical metrics, whereas on larger and more challenging datasets such as ROCOv2 and MIMIC-CXR, it becomes increasingly difficult for conventional RAG to provide consistent gains over an already strong backbone. Given that the backbone itself is a strong SOTA model, such limited or even negative gains from vanilla RAG are therefore unsurprising, while the consistent improvements of SeMLI-RAG further verify its effectiveness and robustness.

Table 1: Experimental results on IU X-Ray, ROCOv2, and MIMIC-CXR. Best results are highlighted in **bold** among our evaluated methods. Rows marked with † denote published SOTA results reported in prior papers. B1/B4: BLEU-1/4, R-L: ROUGE-L, C: CIDEr, BS: BERTScore, BT: BLEURT.

Dataset	Method	NLG				Semantic		Clinical		
		B1	B4	R-L	C	BS	BT	MedCAT	Micro-F1	Macro-F1
IU X-Ray	Backbone	0.477	0.166	0.367	**0.483**	0.670	0.394	0.305	–	–
	CLIP	0.461	0.150	0.354	0.369	0.652	0.400	0.304	–	–
	MedCLIP	0.471	0.154	0.371	0.393	0.665	0.396	0.301	–	–
	CXR-CLIP	0.475	0.162	0.373	0.429	0.668	0.395	0.305	–	–
	BiomedCLIP	0.465	0.148	0.360	0.354	0.656	**0.401**	0.300	–	–
	PromptMRG†	0.401	0.098	0.281	–	–	–	–	–	–
	SeMLI-RAG (ours)	**0.482**	**0.174**	**0.376**	0.476	**0.672**	0.399	**0.318**	–	–
ROCOv2	Backbone	0.187	0.029	0.176	0.124	0.569	0.307	0.150	–	–
	CLIP	0.181	0.015	0.171	0.061	0.548	0.266	0.081	–	–
	MedCLIP	0.190	0.017	0.176	0.065	0.553	0.272	0.089	–	–
	CXR-CLIP	0.189	0.022	0.171	0.094	0.563	0.283	0.113	–	–
	BiomedCLIP	0.201	0.023	0.181	0.095	0.568	0.287	0.114	–	–
	SeMLI-RAG (ours)	**0.204**	**0.033**	**0.183**	**0.133**	**0.579**	**0.331**	**0.162**	–	–
MIMIC-CXR	Backbone	0.331	0.108	0.296	0.130	0.595	0.306	0.272	0.444	0.254
	CLIP	0.307	0.069	0.241	0.047	0.573	0.293	0.163	0.286	0.170
	MedCLIP	0.312	0.072	0.248	0.049	0.586	0.310	0.173	0.311	0.205
	CXR-CLIP	0.318	0.083	0.281	0.086	0.590	0.308	0.212	0.361	0.221
	BiomedCLIP	0.308	0.068	0.245	0.047	0.583	0.305	0.165	0.308	0.197
	PromptMRG†	**0.398**	0.112	0.268	–	–	–	–	–	–
	FactMM-RAG†	–	–	**0.307**	–	0.561	–	–	–	–
	SeMLI-RAG (ours)	0.353	**0.114**	0.297	**0.137**	**0.607**	**0.324**	**0.280**	**0.453**	**0.267**

Table 2: I2I retrieval performance on CheXpert-5x200, reported as mean P@K.

Model	P@1	P@5	P@10
CLIP	0.205	0.222	0.229
BiomedCLIP	0.310	0.303	0.299
ConVIRT	0.394	0.383	0.363
CXR-CLIP	**0.427**	0.391	0.370
PMC-CLIP	0.332	0.314	0.298
SeMLI-Retriever	0.426	**0.399**	**0.391**

5.2 Retrieval Performance

To evaluate whether projected visual features support clinically relevant retrieval and effective cross-modal alignment, we conduct Image-to-Image (I2I) retrieval evaluation. Specifically, we follow the standard CheXpert 5×200 setting (Wang et al., 2022) to construct our query set, randomly sampling 200 image cases for each of the 5 distinct disease categories from the CheXpert dataset, yielding a total of 1,000 query images. To form the retrieval candidates, we utilize the MIMIC-CXR dataset by extracting image-report pairs that share the exact same 5 disease labels as our query set.

We compare our SeMLI retriever with representative radiology retrieval baselines, including CLIP (Radford et al., 2021), BiomedCLIP (Zhang et al., 2024), ConVIRT (Zhang et al., 2022), CXR-CLIP (You et al., 2023), and PMC-CLIP (Lin et al., 2023). We exclude MedCLIP from this benchmark because CheXpert is included in its training data, which would lead to unfair evaluation. As shown in Table 2, our SeMLI retriever achieves the best P@5 and P@10, while being essentially tied with CXR-CLIP at P@1, indicating that explicitly modeling token- and region-level information in our weighted late-interaction mechanism improves clinically relevant radiograph retrieval.

5.3 Ablation Study

We assess the impact of different components of the proposed method from both the retrieval and generation perspectives. As shown in Table 3, for retrieval, SeMLI-Retriever consistently achieves the best image-to-image retrieval precision, outperforming both global retrieval and vanilla late interaction on P@1 / P@5 / P@10, which confirms its ability to retrieve more relevant reference cases. For report generation, removing either image-adaptive RAG or post-hoc validation degrades performance. In particular, discarding image-adaptive RAG leads to important performance degradation across all metrics, demonstrating that adaptive retrieval-aware generation is the key contributor to performance gains. Meanwhile, removing post-hoc validation noticeably weakens the BLEU-1 and Macro-F1 metrics, indicating that this component improves clinical consistency and overall model quality. Together, these results verify the effectiveness of both SeMLI-Retriever and the

Table 3: Ablation study on the proposed method. We evaluate both the effectiveness of **SeMLI-Retriever** in image-to-image retrieval and the contribution of each component in **SeMLI-RAG** for report generation. Best results are highlighted in **bold**. B1/B4: BLEU-1/4, BS: BERTScore.

Retrieval Ablation (I2I)				**Generation Ablation**					
Setting	**P@1**	**P@5**	**P@10**	**Setting**	**B1**	**B4**	**BS**	**MedCAT**	**Macro-F1**
w/o late interaction	0.422	0.380	0.361	w/o adaptive RAG	0.288	0.066	0.541	0.162	0.254
w/o weight	0.400	0.396	0.389	w/o post-hoc validation	0.332	0.109	**0.609**	0.277	0.255
SeMLI-Retriever	**0.426**	**0.399**	**0.391**	**SeMLI-RAG**	**0.353**	**0.114**	0.607	**0.280**	**0.267**

Table 4: Case study. **GT Status** represents the ground-truth report. ✓ and × denote consistency and inconsistency with GT, respectively. ⇒ indicates the **final report selected by our post-hoc validation module** based on image-text alignment. Full reports for each case are provided in Appendix A.6.

Case	**Method**	**Edema**	**Effusion**	**Atelectasis**	**Cardiomegaly**
Case 1	**GT Status**	*Present*	*Bilateral*	*Bibasilar*	*Enlarged*
	Retrieved	✓	×	✓	✓
	Backbone	×	✓	✓	×
	⇒ **Image-adaptive RAG**	✓	✓	✓	✓
Case 2	**GT Status**	*Normal*	*Normal*	*Bibasilar*	*Enlarged*
	Retrieved	×	×	✓	×
	Backbone	✓	✓	✓	×
	⇒ **Image-adaptive RAG**	✓	✓	✓	×
Case 3	**GT Status**	*Interstitial*	*Normal*	*Normal*	*Enlarged*
	Retrieved	×	✓	✓	×
	⇒ Backbone	×	✓	✓	✓
	Image-adaptive RAG	×	✓	✓	×

full SeMLI-RAG framework.

5.4 Case Study

Table 4 presents three representative cases that illustrate the behavior of SeMLI-RAG under different retrieval conditions. In Case 1, the retrieved report provides two key findings, edema and cardiomegaly, that are missed by the backbone, and image-adaptive RAG successfully incorporates these clinically relevant observations, yielding the best overall agreement with the ground truth. In Case 2, the retrieved report introduces several incorrect abnormalities, including edema, effusion, and cardiomegaly. Nevertheless, similar to the backbone, image-adaptive RAG is not misled by these noisy cues and still preserves the correct normal edema/effusion status together with bibasilar atelectasis, demonstrating its robustness to retrieval noise. In Case 3, the retrieved report promotes an incorrect normal-heart pattern, which also affects the image-adaptive RAG output. In this case, however, the backbone prediction remains more consistent with the ground truth by correctly retaining cardiomegaly. Our post-hoc validation module identifies this difference through image-report alignment and selects the backbone output as the final report. Overall, these cases highlight three complementary advantages of our framework: i) effective use of clinically relevant retrieved evidence, ii) robustness against noisy retrievals, and iii) a post-hoc validation mechanism that provides a reliable fallback when RAG is less trustworthy.

6 Conclusion

In this paper, we address two key challenges in retrieval-augmented radiology report generation: the loss of fine-grained information in single-vector retrieval and the difficulty of effectively leveraging helpful references while suppressing noisy ones. To this end, we propose SeMLI-RAG, which combines a semantics-guided late-interaction retriever, an image-adaptive RAG module, and a post-hoc validation mechanism for improved robustness. Experiments on IU X-Ray, ROCOv2, and MIMIC-CXR show that SeMLI-RAG consistently improves over the backbone and vanilla RAG baselines. Further ablation and case analyses confirm the effectiveness of fine-grained retrieval, adaptive reference utilization, and robust final selection in improving radiology report generation.

Limitations

Although SeMLI-RAG achieves consistent improvements, its performance depends on the quality and coverage of the retrieval database. In practice, if the local case repository is small, noisily annotated, or distributionally different from the training data, the retrieved references may be less reliable, thereby limiting the advantage of retrieval augmentation.

In addition, the weighted late-interaction design introduces computational overhead, including higher indexing cost, larger memory footprint, and increased online retrieval latency than conventional dense retrieval. Thus, developing more efficient late-interaction indexing and search strategies, while preserving clinically relevant retrieval performance, would be a valuable direction for future work.

Ethics Statement

Our work uses publicly available radiology datasets, including MIMIC-CXR, IU X-Ray and ROCOv2, for research purposes only. In particular, MIMIC-CXR contains de-identified medical data and is distributed under controlled access. To obtain access, we completed the required training and signed a data use agreement, ensuring compliance with the dataset's privacy and security requirements.

References

Chen Amiraz, Florin Cuconasu, Simone Filice, and Zohar S. Karnin. 2025. The distracting effect: Understanding irrelevant passages in rag. *ArXiv*, abs/2505.06914.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

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

Chi-Min Chan, Chunpu Xu, Ruibin Yuan, Hongyin Luo, Wei Xue, Yike Guo, and Jie Fu. 2024. Rq-rag: Learning to refine queries for retrieval augmented generation. *arXiv preprint arXiv:2404.00610*.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang 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, Online. Association for Computational Linguistics.

Kyoyun Choi, Byungmu Yoon, Soobum Kim, and Jonggwon Park. 2025. Leveraging llms for multimodal retrieval-augmented radiology report generation via key phrase extraction. *ArXiv*, abs/2504.07415.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):474–477.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 6491–6501, New York, NY, USA. Association for Computing Machinery.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jin Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *ArXiv*, abs/2312.10997.

Yunfan Gao, Yun Xiong, Meng Wang, and Haofen Wang. 2024. Modular rag: Transforming rag systems into lego-like reconfigurable frameworks. *ArXiv*, abs/2407.21059.

Jiawei He, Boya Zhang, Hossein Rouhizadeh, Yingjian Chen, Rui Yang, Jin Lu, Xudong Chen, Nan Liu, and Douglas Teodoro. 2025. Retrieval-augmented generation in biomedicine: A survey of technologies, datasets, and clinical applications. *arXiv preprint arXiv:2505.01146*.

Daibing Hou, Zijian Zhao, Yuying Liu, Faliang Chang, and Sanyuan Hu. 2021. Automatic report generation for chest x-ray images via adversarial reinforcement learning. *IEEE Access*, 9:21236–21250.

Chentao Huang, Guangli Li, Xinjiong Zhou, Yafeng Ren, and Hongbin Zhang. 2025. DDGIP: Radiology report generation through disease description graph and informed prompting. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3884–3894, Albuquerque, New Mexico. Association for Computational Linguistics.

Ta Duc Huy, Duy Anh Huynh, Yutong Xie, Yuankai Qi, Qi Chen, Phi Le Nguyen, Sen Kim Tran, Son Lam Phung, Anton van den Hengel, Zhibin Liao, Minh-Son To, Johan W. Verjans, and Vu Minh Hieu Phan. 2025. Seeing the trees for the forest: Rethinking

weakly-supervised medical visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.

Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. 2024. Promptmrg: diagnosis-driven prompts for medical report generation. AAAI'24/IAAI'24/EAAI'24. AAAI Press.

Alistair Johnson, Tom Pollard, Roger Mark, Seth Berkowitz, and Steven Horng. 2024. MIMIC-CXR Database. *PhysioNet*. Version 2.1.0.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 39–48, New York, NY, USA. Association for Computing Machinery.

Zeljko Kraljevic, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio, Leilei Zhu, Amos A Folarin, Angus Roberts, Rebecca Bendayan, Mark P Richardson, Robert Stewart, Anoop D Shah, Wai Keong Wong, Zina Ibrahim, James T Teo, and Richard J B Dobson. 2021. Multi-domain clinical natural language processing with MedCAT: The medical concept annotation toolkit. *Artif. Intell. Med.*, 117:102083.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023: 26th International Conference, Vancouver, BC, Canada, October 8–12, 2023, Proceedings, Part VIII*, page 525–536, Berlin, Heidelberg. Springer-Verlag.

Fenglin Liu, Shen Ge, and Xian Wu. 2021. Competence-based multimodal curriculum learning for medical report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3001–3012, Online. Association for Computational Linguistics.

Dima Mamdouh, Mariam Attia, Mohamed Osama, Nesma Mohamed, Abdelrahman Lotfy, Tamer Arafa, Essam A. Rashed, and Ghada Khoriba. 2025. Advancements in radiology report generation: A comprehensive analysis. *Bioengineering*, 12(7).

Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Re. 2020. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. CHIL '20, New York, NY, USA. Association for Computing Machinery.

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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

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

Arun Reddy, Alexander Martin, Eugene Yang, Andrew Yates, Kate Sanders, Kenton Murray, Reno Kriz, Celso M De Melo, Benjamin Van Durme, and Rama Chellappa. 2025. Video-colbert: Contextualized late interaction for text-to-video retrieval. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19691–19701.

Johannes Rückert, Louise Bloch, Raphael Brüngel, Ahmad Idrissi-Yaghir, Henning Schäfer, Cynthia S. Schmidt, Sven Koitka, Obioma Pelka, Asma Ben Abacha, Alba Garc'ıa Seco de Herrera, Henning Müller, Peter Horn, Felix Nensa, and Christoph M. Friedrich. 2024. ROCOv2: Radiology objects in context version 2, an updated multimodal image dataset. *Scientific Data*, 11(1).

Cornelia Schaefer-Prokop, Ulrich Neitzel, Henk W Venema, Martin Uffmann, and Mathias Prokop. 2008. Digital chest radiography: an update on modern technology, dose containment and control of image quality. *European radiology*, 18(9):1818–1830.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

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

Liwen Sun, James Jialun Zhao, Wenjing Han, and Chenyan Xiong. 2025. Fact-aware multimodal retrieval augmentation for accurate medical radiology report generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 643–655, Albuquerque, New Mexico. Association for Computational Linguistics.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation . In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, Los Alamitos, CA, USA. IEEE Computer Society.

Xinyi Wang, Grazziela Figueredo, Ruizhe Li, Wei Emma Zhang, Weitong Chen, and Xin Chen. 2025. A survey of deep-learning-based radiology report generation using multimodal inputs. *Medical Image Analysis*, 103:103627.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. 2023. R2gengpt: Radiology report generation with frozen llms. *Meta-Radiology*, 1(3):100033.

Zifeng Wang, Zizhao Wu, Dinesh Agarwal, and Jimeng Sun. 2022. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3876–3887. PMID: 39144675; PMCID: PMC11323634.

Orion Weller, Michael Boratko, Iftekhar Naim, and Jinhyuk Lee. 2025. On the theoretical limitations of embedding-based retrieval. *arXiv preprint arXiv:2508.21038*.

Xingyi Yang, Muchao Ye, Quanzeng You, and Fenglong Ma. 2021. Writing by memorizing: Hierarchical retrieval-based medical report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5000–5009, Online. Association for Computational Linguistics.

Kihyun You, Jawook Gu, Jiyeon Ham, Beomhee Park, Jiho Kim, Eun K. Hong, Woonhyuk Baek, and Byungseok Roh. 2023. Cxr-clip: Toward large scale chest x-ray language-image pre-training. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 101–111. Springer Nature Switzerland.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. 2024. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI*, 2(1).

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. 2022. Contrastive learning of medical visual representations from paired images and text. In *Machine learning for healthcare conference*, pages 2–25. PMLR.

A Appendix

A.1 Implementation Details

SeMLI-RAG: We use R2GenGPT as the backbone, which consists of a visual encoder, a lightweight feature-alignment module, and a LLaMA decoder for radiology report generation. Specifically, it adopts microsoft/swin-base-patch4-window7-224 as the vision encoder and meta-llama/Llama-2-7b-chat-hf as the text decoder. We use microsoft/BiomedVLP-CXR-BERT-specialized as a frozen medical language model for retrieval supervision. The shared retrieval space is set to 128 dimensions, with $m = 64$ learnable semantic anchors A. The image-adaptive RAG module contains 32 learnable latent queries and two Transformer decoder layers. Training is carried out in three stages: backbone training, retriever training, and RAG training. In the backbone training stage, the vision encoder and the feature alignment module are trainable, while the text decoder and the medical language model remain frozen. In the retriever training and RAG training stages, the vision encoder is frozen. Throughout all stages, the text decoder and the medical language model are kept frozen. We use AdamW with cosine learning-rate decay and 5% warm-up. All experiments are conducted on a single NVIDIA A100 GPU.

Baseline: For the vanilla RAG baseline built upon the R2GenGPT backbone model, the most similar report retrieved is directly concatenated into the textual prompt as auxiliary reference, together with its similarity score when available, and fed into the decoder without additional adaptive filtering or retrieval-aware fusion. To enable parameter-efficient adaptation, we fine-tune the LLM using LoRA, with rank $r = 16$ and scaling factor $\alpha = 16$ for IU X-Ray, and rank $r = 32$ and scaling factor $\alpha = 64$ for ROCOv2 and MIMIC-CXR, using a LoRA dropout of 0.1. For a fair comparison, the vanilla RAG baseline uses the same training epochs and decoding hyperparameters as SeMLI-RAG.

A.2 Further analysis of Retrieval performance

The CheXpert 5×200 dataset considers five findings: Atelectasis, Cardiomegaly, Consolidation, Edema, and Pleural Effusion. For each finding, 200 studies with the corresponding exclusive label were sampled from MIMIC-CXR. Table 5 reports the statistics, where Matched denotes the number of candidate studies for each finding in MIMIC-CXR, MR denotes the match rate among all 227,827 source studies, and SR denotes the sampling rate.

Table 5: Statistics for the CheXpert 5×200 dataset. MR: match rate. SR: sampling rate.

Condition	Matched	MR	SR
Atelectasis	5,253	2.3%	3.8%
Cardiomegaly	7,358	3.2%	2.7%
Consolidation	805	0.4%	24.8%
Edema	2,883	1.3%	6.9%
Pleural Effusion	4,720	2.1%	4.2%

To further assess the cross-modal retrieval capability of our method, we conducted an image-to-text (I2T) retrieval evaluation on the CheXpert-5×200 dataset, where each image was used as a query to retrieve candidate reports. As shown in Table 6, SeMLI-Retriever achieves the best performance across all metrics, consistently outperforming all baseline models on P@1, P@5, and P@10. These results indicate that SeMLI-Retriever learns stronger cross-modal alignment between query images and candidate reports. Such alignment is particularly important for our post-hoc validation stage, as it enables more reliable image-report matching and improves the robustness of the retrieval and verification pipeline.

Table 6: I2T retrieval results on CheXpert-5x200, reported as mean P@K.

Model	**P@1**	**P@5**	**P@10**
CLIP	0.191	0.189	0.188
BiomedCLIP	0.345	0.330	0.317
ConVIRT	0.414	0.410	0.407
CXR-CLIP	0.320	0.308	0.307
PMC-CLIP	0.311	0.308	0.301
SeMLI-Retriever	**0.449**	**0.434**	**0.423**

A.3 Projected Image Features Dimension

We further investigate the impact of the projected feature dimension. As shown in Figure 3, increasing the projected feature dimension improves image-to-text retrieval performance, while bringing limited or even negligible gains for image-to-image retrieval. This is likely due to increasing the feature dimension enhances the semantic capacity of the representations, thereby benefiting cross-modal semantic alignment, which is inherently more complex, while introducing redundancy for intra-modal similarity. Considering the additional memory and

Table 7: Storage, latency, and retrieval performance comparison on CheXpert-5×200. Storage denotes the feature storage required for the entire candidate database and is independent of the candidate pool size.

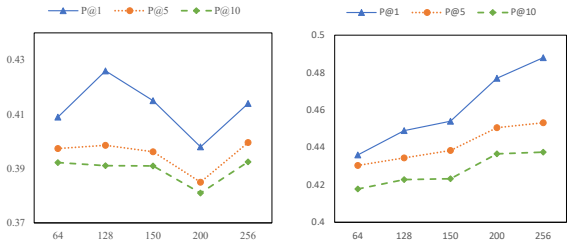
Method	Feature Dim	Pool Size	Storage	Latency	P@10
Global vector	1024	–	8.08 MB	1.93 ± 0.28 ms	0.3606
SeMLI Retriever	128×49	50	25.95 MB	2.07 ± 0.04 ms	0.3715
SeMLI Retriever	128×49	100	25.95 MB	3.46 ± 0.29 ms	0.3859
SeMLI Retriever	128×49	200	25.95 MB	6.35 ± 0.64 ms	0.3874
SeMLI Retriever	128×49	300	25.95 MB	9.20 ± 0.08 ms	0.3871

Table 8: Prompt templates used by the backbone, the vanilla RAG baseline, and SeMLI-RAG.

Method	Prompt template
Backbone	`Human: <Img> [Image Features] Generate a comprehensive and detailed diagnosis report for this radiology image. Assistant:`

Vanilla RAG baseline	`Human: <Img>[Image Features] You are given a chest xray image. Here is a clinically similar case for reference.`
	`Similar case report: {retrieved report}`
	`Retrieval similarity score: {score} (if available)`
	`Please generate a comprehensive and detailed diagnosis report for the current radiology image.`
	`Do not copy the similar case verbatim. Focus on the current image and use the similar case only as auxiliary reference. Assistant:`

SeMLI-RAG (ours)	`[Human: <Img>]` $\oplus$ `[Image Features]` $\oplus$ `[Reference: <Ref>]` $\oplus$ E_{ref} $\oplus$ `[</Ref>]` $\oplus$ `[Generate a comprehensive and detailed diagnosis report for this radiology image. Assistant:]`



(a) I2I Retrieval Results (b) I2T Retrieval Results

Figure 3: Comparison of retrieval performance across different dimensions (dim) of projected image features $\hat{Z}$. (a) Image-to-Image (I2I) results; (b) Image-to-Text (I2T) results.

computation overhead incurred by larger dimensions, we adopt 128 as a balanced choice.

A.4 Retrieval Storage and Latency

We further evaluate the index storage, retrieval latency, and retrieval performance of the proposed SeMLI Retriever on the CheXpert-5×200 benchmark. We compare SeMLI Retriever with a simple global-vector retrieval baseline, where the visual encoder feature is directly used to build a FAISS index. As shown in Table 7, the global-vector baseline requires the smallest index storage and achieves the lowest retrieval latency, since each image is represented by a single feature vector and retrieval only involves a k-nearest-neighbor search over global representations.

In contrast, SeMLI Retriever stores all projected image features for each candidate image. Specifically, each image is represented by 49 patches, each with a 128-dimensional projected feature. For 1,000 candidate images, this local feature storage increases the storage cost from 8.08 MB to 25.95 MB, resulting in approximately a 3.2× increase in storage cost. However, this additional storage enables the second-stage semantic-weighted late interaction, which improves retrieval performance over the global-vector baseline.

The retrieval latency of SeMLI Retriever is determined by both the first-stage global retrieval and the second-stage local reranking. The first stage performs efficient FAISS search using the global image representation. The second stage reranks the top-K_c candidates using weighted late interaction. For each query, the local reranking cost is approximately

$$O(K_c \cdot N_q \cdot N_c \cdot d),$$

where K_c is the candidate pool size, N_q and N_c denote the numbers of query and candidate patches, and d is the local feature dimension. In our setting, $N_q = N_c = 49$ and $d = 128$ are fixed, so the reranking cost grows approximately linearly with the candidate pool size K_c.

Empirically, increasing the candidate pool size

Table 9: Comparison of retrieved reports, backbone outputs, and Image-adaptive RAG outputs against ground-truth radiology reports. Orange denotes key ground-truth findings. Green indicates content consistent with the ground truth, and red indicates conflicting or unsupported findings or suspicions. Uncolored text denotes neutral additional description. ✓ indicates the output selected by the post-hoc validation stage.

Case	Ground-truth Report	Retrieved Similar Case	Backbone Output	Image-Adaptive RAG Output
1	*Impression:* Pulmonary edema with consistent with moderate to large pleural effusions and adjacent atelectasis. Superimposed infection cannot be excluded. *Findings:* Large bilateral pleural effusions are present with adjacent opacities most consistent with compressive atelectasis. Cephalization and indistinct appearance of the pulmonary vasculature are consistent with pulmonary edema. Heart size appears enlarged. Mediastinal contours are stable.	*Impression:* Findings suggest mild edema and possible developing opacity at the right lung base. *Findings:* The heart is again mildly to moderately enlarged. Cardiac, mediastinal, and hilar contours are stable. There is no definite pleural effusion or pneumothorax. In addition to new central vascular fullness, mild vascular pulmonary edema is suggested. A focal right basilar opacity may represent atelectasis or possibly developing pneumonia.	*Impression:* Moderate bilateral pleural effusions with bibasilar opacities, likely atelectatic; superimposed infection cannot be excluded. *Findings:* There are moderate bilateral pleural effusions with bibasilar opacities, likely atelectatic; superimposed infection cannot be excluded. Cardiac silhouette is difficult to assess because of low lung volumes. No acute osseous abnormality is identified.	✓ *Impression:* Moderate to severe pulmonary edema with bilateral pleural effusions and bibasilar atelectasis. *Findings:* There is moderate to severe pulmonary edema with bilateral pleural effusions and bibasilar atelectasis. No pneumothorax is seen. Cardiac silhouette appears mildly enlarged. Mediastinal contour is unremarkable.
2	Impression: Low lung volumes. Findings: Lung volumes are low. No pleural effusion or pneumothorax is detected. Bibasilar atelectasis is present. There is mild left ventricular enlargement. Bilateral rib fractures are noted.	*Impression:* Diffuse airspace and interstitial opacities, worsened from prior comparisons but improved relative to an earlier examination; differential considerations include pulmonary hemorrhage, pulmonary edema, and interstitial fibrosis. *Findings:* Cardiomediastinal contours are stable and within normal limits. The lungs demonstrate diffuse airspace and interstitial opacities. Blunting of the costophrenic angles and subtle obscuration of the hemidiaphragms suggest bibasilar atelectasis and small bilateral pleural effusions. There is no pneumothorax.	*Impression:* Low lung volumes with bibasilar atelectasis. *Findings:* There are low lung volumes. Bibasilar atelectasis is noted. No definite focal consolidation is seen. No pneumothorax is seen. Cardiomediastinal silhouette is unremarkable. No acute osseous abnormality is identified.	✓ *Impression:* Low lung volumes with bibasilar atelectasis. *Findings:* There are low lung volumes. Bibasilar atelectasis is seen. No pleural effusion or pneumothorax is seen. No definite acute osseous abnormality is identified.
3	Impression: Mild cardiomegaly and mild interstitial edema. Findings: Heart remains mildly enlarged. There is mild interstitial pulmonary edema. No large pleural effusion. Eventration of the right hemidiaphragm is noted. No focal consolidation suggestive of pneumonia.	*Impression:* No acute cardiopulmonary process. *Findings:* Low lung volumes. There is no pleural effusion or focal consolidation. No pneumothorax. Heart size is normal. There is no pulmonary edema.	✓ *Impression:* No acute cardiopulmonary abnormality. *Findings:* Heart size is mildly enlarged. Pulmonary vasculature is normal. Lungs are clear. No focal consolidation or pleural effusion is present. No pneumothorax is present. Eventration of the right hemidiaphragm is again noted.	*Impression:* No acute cardiopulmonary abnormality. *Findings:* Heart size is normal. Pulmonary vasculature is normal. Lungs are clear. No pleural effusion is seen. No pneumothorax is seen. Eventration of the right hemidiaphragm is noted.

improves retrieval performance, but the improvement becomes marginal once the pool size exceeds 100, while the retrieval latency continues to increase. Therefore, we choose a candidate pool size around 100 as a practical trade-off between retrieval accuracy and efficiency.

A.5 Prompt Templates

To clarify the differences between the backbone, the vanilla RAG baseline, and our SeMLI-RAG framework, we summarize their prompt templates in Table 8.

A.6 The Reports of the Case Study

We present the full report contents in Table 9, which serve as a supplementary illustration to the case selection results shown in Table 4.