ANATORAG: ANATOMY-GUIDED RETRIEVAL FOR CHEST X-RAY REPORT GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Given radiographic images, chest X-ray report generation aims to produce clinically accurate textual findings. A common recipe is retrieval-augmented generation (RAG), which takes prior reports as reference. However, the reports in such a corpus come from other patients, so their findings are plausible yet unverified for the target image, and report-level retrieval cannot separate the compatible findings from the rest. Yet retrieving at statement level does not resolve this: the retrieved text remains unverified against the target image and largely redundant. We introduce AnatoRAG, which treats retrieved evidence as a hypothesis to be verified and compressed rather than text to be copied. First, AnatoRetriever uses coarse anatomical masks to guide statement retrieval, and a finding classifier to discard statements that contradict the target image. Second, the Anatomy Evidence Adapter condenses the surviving statements and image-label predictions into compact anatomical representations, which a multimodal LLM fuses with visual features to generate the report. Extensive experiments on MIMIC-CXR show that AnatoRAG achieves 65.1 CheXbert-14 micro-F1, surpassing the most competitive method, *i.e.*, RADAR, by 2.4. Trained only on MIMIC-CXR, AnatoRAG also transfers to the unseen CheXpert Plus and IU X-Ray datasets, yielding 60.4 and 54.6 CheXbert-14 micro-F1, respectively. The improvement is concentrated on findings whose evidence must be recomposed across anatomical regions, indicating that verification and compression, rather than retrieval alone, drive the gain.

1 INTRODUCTION

Chest X-ray report generation aims to produce radiology reports that describe clinically relevant findings from chest radiographs. Multimodal large language models can generate fluent reports (Hyland et al., 2023; Bannur et al., 2024; Zhang et al., 2025), but fluency does not ensure that every statement is supported by the radiograph (Delbrouck et al., 2022; Yu et al., 2023). Retrieval-augmented generation reduces this risk by sourcing such language from the reports of visually similar training studies (Zhang et al., 2024; Sun et al., 2025). Because these studies belong to other patients, the retrieved text is plausible but never guaranteed to hold for the current radiograph. The central question is therefore which retrieved statements describe the current study.

Whole-report retrieval answers that question at the wrong granularity. A report may contain one relevant statement together with unrelated or contradictory findings of the other patient, and all statements enter the generator as a single evidence unit. This problem becomes sharper as the number of findings increases because no single training report is guaranteed to exhibit the combination the query presents. For the 1,748 test queries with at least two positive reference labels, the top-1 training report retrieved by BioMedCLIP recovers only 32.8% of those labels on average and matches the complete set for 1.6%.

The evidence itself is rarely the bottleneck. For 99.4% of queries with any positive label, the required label set can be composed from statements available in the training corpus, and a median of two evidence statements can compose the label set. Figure 1 develops this motivation across three corpus statistics. Panel (b) shows that distinct training label combinations become more numerous as positive-label count grows, while panel (c) shows that fewer studies share each combination. Panel (d) asks how many matches remain for each test query: complete-label-set matches collapse with additional findings, whereas within-anatomy matches remain abundant. Together, these observations identify

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

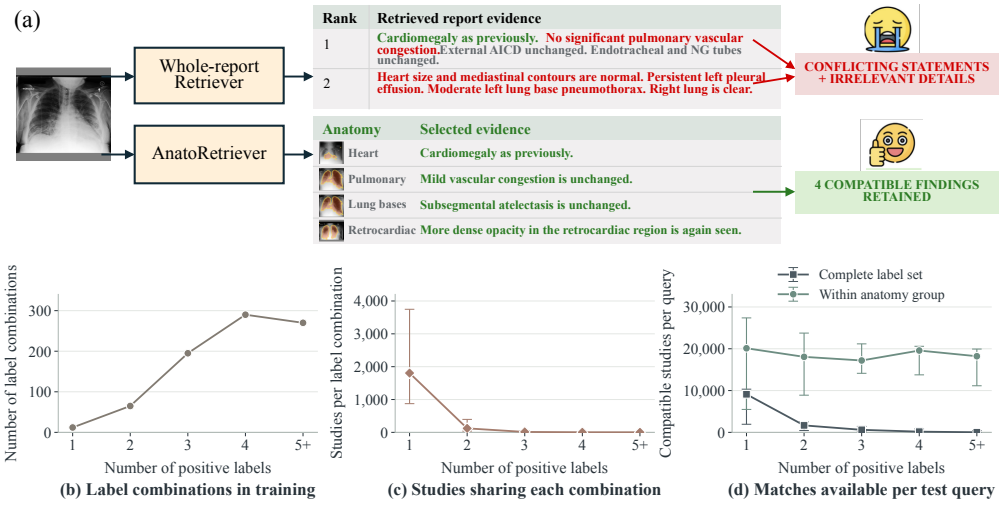


Figure 1: Report-level retrieval admits whole reports, so a supported finding arrives alongside contradictory and irrelevant statements; AnatoRetriever selects anatomy-linked evidence independently. Panel (a) traces one query radiograph through both retrieval units. Panels (b)–(d) use the number of positive CheXbert labels on the horizontal axis. Panel (b) counts distinct label combinations in the training set. Panel (c) reports the median number of training studies sharing each exact combination. Panel (d) reports median matches per test query: the dark curve requires the complete label set to match, while the green curve matches labels within each active anatomical group and averages across those groups. Whiskers in panels (c)–(d) denote the 25th–75th percentiles.

access to a compatible combination of statements, rather than the absence of relevant language, as the bottleneck. Existing retrievers return whole reports (Zhang et al., 2024; Sun et al., 2025), sentences (Endo et al., 2021), or phrases (Park et al., 2026). Recent work further conditions sentence retrieval on detected anatomical regions (Sloan et al., 2026). Smaller units improve access to individual findings, but similarity alone cannot determine whether a statement agrees with the current image. Such retrievers also return overlapping statements that restate the same finding, diluting the evidence presented to the generator. RADAR instead filters observations from a draft report (Hou et al., 2025). We separate two earlier decisions: finding admission determines which labels the image supports, and statement selection retrieves compatible text that covers those labels without redundancy. Table 1 compares the resulting retrieval units.

AnatoRAG implements this separation before report drafting. AnatoRetriever organizes training statements in six anatomical families, forms one anatomy-guided query per label, admits image-supported findings, and selects compatible evidence without redundancy. The Anatomy Evidence Adapter then maps variable-length selected text into fixed anatomical slots, reducing interference from raw retrieval text before generation. Table 3 compares the fixed-slot interface with raw-text conditioning under the same backbone. Tables 1 and 2 present scores for selected evidence and generated reports separately, so that the effect of evidence selection can be separated from that of generation. The evaluation measures retrieved evidence and generated reports separately, compares report-level and anatomy-linked evidence under one backbone, and tests the resulting system on two external cohorts. Our primary contributions are:

- We introduce AnatoRAG, which verifies each retrieved statement against the target image before compressing it for the generator, rather than concatenating whatever similarity-based retrieval returns. AnatoRetriever issues anatomy-guided label queries over a training-only Clinical Evidence Bank and admits a candidate statement only when the target image independently supports its finding, rejecting unsupported statements instead of down-ranking them. The Anatomy Evidence Adapter then compresses the surviving statements onto fixed anatomical slots, so the generator reads a constant-size, region-ordered representation rather than a variable-length statement set.
- On the Findings section of the MIMIC-CXR test set, AnatoRAG (BLIP-3) reaches 65.1 CheXbert-14 micro-F1, 34.9 RadGraph F1 and 27.7 BLEU-4. Separately, holding the BLIP-3 backbone, the training set and the decoding fixed, adding AnatoRetriever and the Anatomy Evidence Adapter raises CheXbert-14 macro-F1 from 40.3 to 50.8. The gain concentrates on findings whose evidence must be recomposed across anatomical regions.

2 Related Work

Report generation. Multimodal language models generate chest X-ray reports by adapting general vision–language architectures to radiology (Hyland et al., 2023; Bannur et al., 2024; Zambrano Chaves et al., 2025; Zhang et al., 2025). Multimodal visual interfaces improve image–text alignment, while structured generation methods organize the report around clinically meaningful regions or topics. For example, LLaVA-TA uses CXAS masks to generate anatomy-specific content (Cheng & Subramanian, 2026). These systems primarily change how visual features are represented or decoded. AnatoRAG instead changes the evidence supplied to the generator by selecting retrieved statements that describe the current study.

Retrieval units and admission. Retrieval units range from whole reports (Sun et al., 2025) and sentences (Endo et al., 2021) to key phrases (Park et al., 2026) and localized abnormality proposals (Yan et al., 2026). STAR-3 detects anatomical regions and retrieves report sentences conditioned on clinical indications and longitudinal changes between current and prior images (Sloan et al., 2026). Whole reports preserve discourse but couple a relevant finding to every co-occurring statement in the retrieved study; smaller units reduce this coupling but must be assembled into a coherent evidence set. RADAR instead selects observations from a draft report (Hou et al., 2025). Granularity alone therefore does not determine whether retrieved text is valid for the current image: similarity can still admit a statement that negates an active finding or introduces an inactive one. Self-RAG asks the generator to assess relevance and support during generation (Asai et al., 2024). AnatoRetriever performs this check before generation by separating finding admission from compatibility- and coverage-aware statement selection.

Structure of the retrieved evidence. BLIP-2 and xGen-MM pool input representations before language generation (Li et al., 2023; Xue et al., 2024). MAIRA-2 grounds generated findings in image regions (Bannur et al., 2024), and region-aware medical MLLMs train on region-level tasks (Wang et al., 2025). These approaches preserve spatial information in visual features or generation targets. AnatoRAG uses CXAS masks (Seibold et al., 2022) earlier in the pipeline: each mask constrains a retrieval query, and each selected evidence statement retains the corresponding anatomical association. The Anatomy Evidence Adapter then maps variable-length statements into family-specific slots rather than presenting the selected evidence as one raw-text sequence. Thus, anatomy determines both where evidence is sought and where the evidence enters the generator, while assertion compatibility determines whether a candidate may be used.

3 Method

AnatoRAG follows a retrieval-then-generation scheme (Zhang et al., 2024; Sun et al., 2025), but whole-report similarity cannot determine which statements describe the current study. AnatoRetriever first converts the current radiograph into anatomy-guided label queries, admits only image-supported findings, and selects compatible evidence statements from a training-only bank. This retrieval stage is followed by anatomy-structured generation. Before generation, the Anatomy Evidence Adapter converts variable-length selected text into fixed anatomical slots, limiting interference from raw retrieved text while preserving anatomical source information and calibrated image-label guidance.

Figure 2 shows the complete information flow. The inputs are a current radiograph I_{cur}, an optional prior radiograph I_{pri}, and a clinical prompt P; the output is a Findings report Y. The prompt contains available non-image context; in the BLIP-3 experiments the prompt also contains available prior Findings. AnatoRetriever forms queries from I_{cur} alone. When available, I_{pri} receives a separate prediction pass and also enters the generator through the native visual interface. AnatoRetriever is trained before language-model adaptation and remains frozen during report-generation training, so discrete evidence selection receives no gradient from the generator.

3.1 AnatoRetriever: Anatomy-Guided Retrieval

Anatomy-guided query formation. A visually adapted encoder (Pérez-García et al., 2025) extracts patch features H_{cur} from the whole current image. For each of the 14 labels ℓ, a baseline query $q_\ell^{(0)}$ attends to predefined spatial regions. An anatomy-guided residual further attends to patches permitted

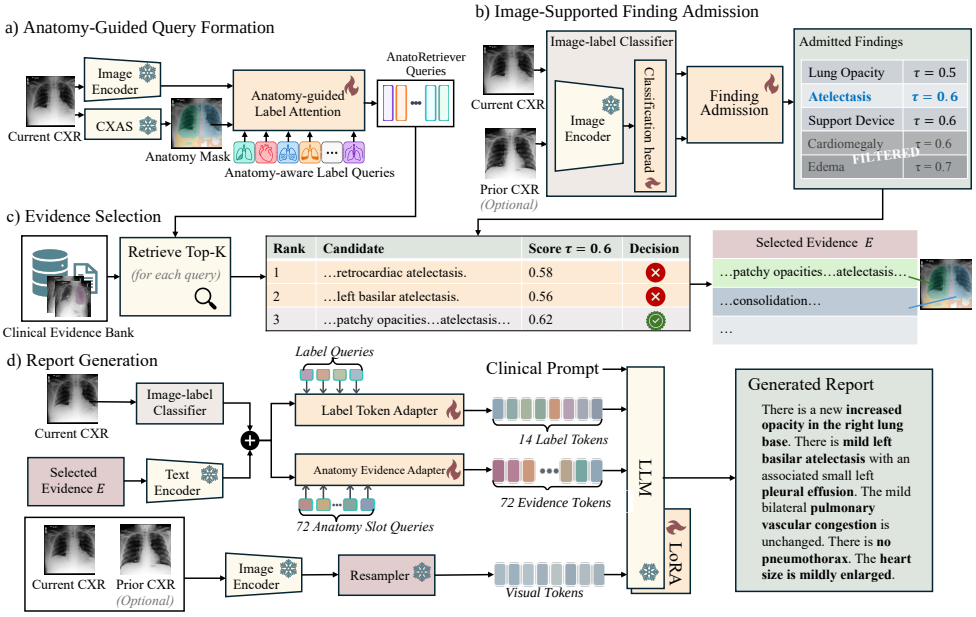


Figure 2: AnatoRAG uses AnatoRetriever to form anatomy-guided label queries, admit image-supported findings, and select compatible evidence statements from the training-only Clinical Evidence Bank. The Anatomy Evidence Adapter then maps selected statements and calibrated image-label scores into fixed evidence and label tokens, followed by report generation with native visual tokens and the clinical prompt. Numeric scores and thresholds illustrate the decision rule; evaluation uses the training-fitted per-label thresholds in Eq. 2. Flames and snowflakes mark trainable and frozen modules.

by a label-specific mask M_ℓ:

$$q_\ell = \mathrm{norm}\Big(q_\ell^{(0)} + \alpha_\ell W_\ell \,\mathrm{AttnPool}(H_{\mathrm{cur}}, M_\ell)\Big)\,. \tag{1}$$

Here W_ℓ maps attended features to the evidence-embedding space, α_ℓ scales the residual, and norm denotes ℓ_2 normalization. Attention pooling retains the highest-weighted permitted patches for each label. The anatomy mask restricts attention over whole-image features and does not provide pixel-level lesion annotation. Anatomy-guided query formation prevents a globally similar report from defining the evidence for every finding.

Image-supported finding admission. The query q_ℓ determines which statements rank highly, whereas finding admission determines which labels may be retrieved. A semantic residual head reads global image features, label queries, and mask availability to correct direct image-classification logits. At inference, direct scores are composed in a fixed order with class-specific heads, global and regional neighbours, anatomical subviews, and a frozen image-label classifier (Hou et al., 2025). Appendix Table S3 gives the active branch for each label. The resulting admission score r_ℓ defines

$$A = \{\ell : r_\ell \geq \tau_\ell\}, \tag{2}$$

where τ_ℓ is a label-specific threshold fitted on the training split and fixed for evaluation. The threshold acts on the composed score; r_ℓ is an admission score rather than a calibrated probability. Admission prevents retrieval from introducing evidence for findings unsupported by the current image.

The retrieval components are trained in stages. Whole-image visual adaptation first uses 14-label supervision derived from training reports. With the visual model fixed, baseline queries learn to match training evidence statements through a contrastive objective. The anatomy-guided residual then refines the queries using matched statements and hard textual contrasts, while an anchor limits drift from $q_\ell^{(0)}$. Finally, with visual and query features fixed, the semantic residual head learns logit corrections Δ from report-derived labels:

$$\mathcal{L}_{\mathrm{query}} = \mathcal{L}_{\mathrm{contrast}} + \lambda_{\mathrm{anc}}\mathcal{L}_{\mathrm{anchor}}, \qquad \mathcal{L}_{\mathrm{gate}} = \mathcal{L}_{\mathrm{WBCE}} + \lambda_{\mathrm{rank}}\mathcal{L}_{\mathrm{rank}} + \lambda_{\mathrm{res}}\,\mathrm{mean}(\Delta^2). \tag{3}$$

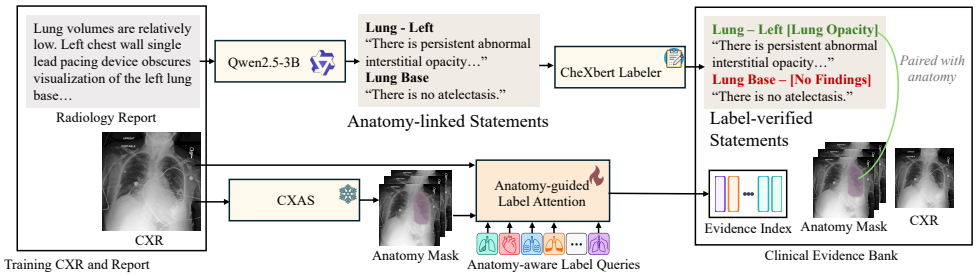


Figure 3: Training-report Findings are decomposed into independently selectable evidence statements. Qwen extracts anatomy-linked statements, CheXbert verifies report-derived labels, and a frozen text encoder embeds retained entries. Six anatomical families organize the Clinical Evidence Bank, while retrieval and candidate eligibility remain label-specific. Text, masks, and icons are illustrative.

The first objective describes residual-query training; the baseline query uses the contrastive term alone. Separately trained class-specific heads and visual-adaptation branches contribute to the frozen score-composition procedure.

Anatomy-linked evidence selection. Figure 3 shows construction of the Clinical Evidence Bank from training-report Findings. Each sentence is decomposed into JSON records whose atomic text retains polarity, uncertainty, temporality, anatomy, laterality, and exact supporting spans. A deterministic parser rejects invalid schemas and spans before CheXbert checks the labels of each statement and source report. Each bank entry j retains statement text t_j, a normalized text embedding v_j, the source patient and anatomical family, and explicitly affirmed and negated labels F_j^+ and F_j^-. Absence from F_j^+ does not imply negation; F_j^- records stated absence. Six anatomical families organize the entries, while each individual label defines retrieval eligibility.

For each $\ell \in A$, the query scores eligible statements by $s_{\ell j} = q_\ell^\top v_j$. Eligibility requires $\ell \in F_j^+$ and a source patient different from the query patient. The highest-scoring eligible entries form $\mathcal{C}_\ell$. A shortlisted statement is label-compatible when, under the CheXbert-14 inventory, the statement negates no active label and affirms no inactive label:

$$\mathcal{V}_\ell = \{j \in \mathcal{C}_\ell : F_j^- \cap A = \varnothing,\ F_j^+ \subseteq A\}. \tag{4}$$

Active labels are visited in a fixed order while U records labels already covered by selected statements. A label in U requires no additional selection. Otherwise, for a nonempty $\mathcal{V}_\ell$, AnatoRetriever chooses

$$j_\ell^* = \arg\max_{j \in \mathcal{V}_\ell} \left[s_{\ell j} + \lambda_{\text{cov}} |F_j^+ \setminus U| \right]. \tag{5}$$

The second term favours a compatible statement that also covers active labels not yet represented. Starting from $E = ()$ and $U = \varnothing$, AnatoRetriever appends a record only when the statement identifier and normalized text are new, followed by $U \leftarrow U \cup F_{j_\ell^*}^+$. One statement may therefore cover several labels, and evidence length varies across cases. An empty $\mathcal{V}_\ell$ triggers the most similar entry in $\mathcal{C}_\ell$; an otherwise empty E receives a neutral entry from another patient.

3.2 AnatoRAG: Anatomy-Structured Generation

Anatomy evidence adaptation. Admission scores r_ℓ select evidence and do not enter the generator as image probabilities. The frozen image-label classifier produces current-image logits that undergo class-wise calibration fitted on training data, yielding a separate image-label prior $\tilde{\mathbf{p}}$. Calibration and the classifier remain fixed during adapter and language-model training.

Selected records E supply text, source, and anatomical-family information to a shared evidence encoder. Learned anatomy-slot queries read contextual encoder states through family-structured cross-attention. Each of the six families contributes one header and 11 content positions, producing $Z_E \in \mathbb{R}^{72 \times d}$, where d is the generator token width. A separate label branch combines pooled evidence states with $\tilde{\mathbf{p}}$ to produce $Z_S \in \mathbb{R}^{14 \times d}$. The branches share evidence states but use distinct projections and gates. The Anatomy Evidence Adapter is

$$(Z_E, Z_S) = \text{Adapter}(E, \tilde{\mathbf{p}}), \tag{6}$$

with the fixed interface $Z = [Z_E; Z_S]$. Converting variable-length text into family-specific slots prevents long or frequent statements from dominating the generator input and preserves the anatomical organization of selected evidence.

Report generation. The generator reads Z together with native visual tokens for available images and the clinical prompt. Training separates adapter initialization from generator adaptation. With the generator frozen, the evidence interface is initialized using

$$\mathcal{L}_{\text{init}} = \lambda_{\text{aux}}(\mathcal{L}_{\text{label}} + \lambda_{\text{soft}}\mathcal{L}_{\text{soft}}) + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}}. \tag{7}$$

The label and soft-target terms preserve clinical content in the evidence tokens, while $\mathcal{L}_{\text{rec}}$ reconstructs selected evidence from the anatomy-slot positions. Generator adaptation then updates the adapter and language LoRA through teacher-forced report generation:

$$\mathcal{L}_{\text{gen}} = -\sum_{t=1}^{T} \log p_\theta(y_t \mid y_{<t}, I_{\text{cur}}, I_{\text{pri}}, P, Z)\,. \tag{8}$$

During generator adaptation, $\mathcal{L}_{\text{init}}$ is disabled and the adapter is updated solely through $\mathcal{L}_{\text{gen}}$. The image encoder, image-label classifier, and calibration remain fixed; report-derived targets supervise training and are absent at inference. Controlled report-generation experiments initialize the released XGen-MM/BLIP-3 backbone (Xue et al., 2024); Appendix E gives checkpoint and adaptation details.

4 Experiment

We train and evaluate on the Findings section of MIMIC-CXR (Johnson et al., 2019). The main comparison uses 2,461 Findings queries from 285 patients in the official test split, whose patients are disjoint from training. Both indexes are built from training data only, and candidates from the query patient are excluded before Top-k selection. Generation tables report one trained checkpoint per configuration; the paired retrieval analyses report patient-bootstrap intervals where shown.

Metrics. We report the metric families used by the systems we compare against (Hou et al., 2025; Zhang et al., 2025): BLEU-1, BLEU-4, and ROUGE-L for lexical overlap, and RadGraph F1, RadCliQ-v0, and CheXbert F1 for clinical content. We report micro- and macro-averaged CheXbert F1 over both the 14- and 5-label inventories, counting uncertain observations as negative. All labelers score the text a system produces. Scores are percentages except RadCliQ-v0; – marks an unreported value.

4.1 Statement Retrieval

Table 1 is a retriever-only comparison: no report generator is called. For each test image, a method returns reports, sentences, phrases, or evidence statements; CheXbert labels the retrieved text and compares those labels with the reference Findings. The superscript identifies the 14- or 5-label inventory. Macro-F_1 weights each label equally, whereas micro-F_1 pools decisions across labels and cases. The "Retrieved unit" column identifies the text unit being scored, rather than a generated-report format. We show in Table 1 that AnatoRetriever reaches 63.0 CheXbert-14 micro-F_1 and 50.3 macro-F_1 on the 2,461-query evaluation. The published report-, sentence-, and phrase-retrieval systems are shown under the source protocols and provide context across retrieval units rather than a controlled ranking. Bank coverage separates availability from selection. An exact label-set match exists for 99.4% of positive queries, and a median of two evidence statements can compose one. Nevertheless, the most similar report recovers less than one third of the target labels and exactly matches only 1.6% of multi-label queries. Across 1,748 such queries, AnatoRetriever yields 38.3 points higher mean recall and 8.0 points higher exact matching than report-level Top-1. The bank already contains the evidence; the retrieval challenge is to assemble the evidence without importing incompatible findings.

4.2 Report Generation

We show in Table 2 a controlled BLIP-3 comparison and a separate Libra-3B integration. Under the controlled BLIP-3 protocol, AnatoRAG raises CheXbert-14 macro/micro-F_1 by 10.5/8.6 points and RadGraph F_1 by 2.1, while RadCliQ-v0 improves from 2.74 to 2.61. The Libra-3B integration reaches 42.9/57.3 CheXbert-14 macro/micro-F_1, 32.3 RadGraph F_1, and 2.71 RadCliQ-v0. Because

Method	Retrieved unit	^{14}Ma	^{5}Ma	^{14}Mi	^{5}Mi
Published systems					
BiomedCLIP (Zhang et al., 2024)	Report	23.3	<u>41.8</u>	44.6	52.9
FactMM-RAG (Sun et al., 2025)	Report	–	–	–	<u>60.5</u>
CXR-RePaiR-Select (Endo et al., 2021)	Sentences	27.4	–	–	–
RA-RRG (Park et al., 2026)	Phrases	<u>42.5</u>	–	<u>58.8</u>	–
AnatoRetriever	Evidence statements	**50.3**	**58.7**	**63.0**	**64.9**

Table 1: Retriever-only comparison on MIMIC-CXR; the generator is not invoked. Published systems retain the corresponding source evaluation protocols, while AnatoRetriever is evaluated on 2,461 queries. Superscripts 14 and 5 denote the label inventory; Ma and Mi denote macro- and micro-averaging. Values are percentages, uncertain observations count as negative, and higher is better. Bold and underline mark the highest and second-highest reported value in each column.

Model	Lexical metrics			Clinical metrics					
	B-1	B-4	R-L	RG-F_1	$CliQ_0^{\downarrow}$	^{14}Ma	^{5}Ma	^{14}Mi	^{5}Mi
Published systems									
MAIRA-1 (Hyland et al., 2023)	39.2	14.2	28.9	24.3	3.10	38.6	47.7	55.7	56.0
MAIRA-2 (Bannur et al., 2024)	46.5	23.4	38.4	<u>34.6</u>	<u>2.64</u>	41.6	50.4	58.1	59.1
LLaVA-Rad (Zambrano Chaves et al., 2025)	38.1	15.4	30.6	–	–	39.5	47.7	57.3	57.4
Libra-7B (Zhang et al., 2025)	51.3	24.5	36.7	32.9	2.70	40.4	53.8	55.9	60.1
Libra-3B† (Zhang et al., 2025)	50.5	23.3	35.2	–	–	–	–	–	–
RA-RRG (Park et al., 2026)	–	8.0	24.9	26.7	–	41.7	–	58.5	–
RADAR (Hou et al., 2025)	50.9	<u>26.2</u>	<u>39.7</u>	<u>34.6</u>	**2.61**	<u>46.0</u>	<u>56.7</u>	<u>62.7</u>	<u>65.3</u>
Generators without retrieval									
Libra-3B‡ (Zhang et al., 2025)	50.4	23.2	34.7	30.2	2.90	37.5	44.7	54.8	54.6
BLIP-3 backbone (Xue et al., 2024)	48.6	24.6	38.3	32.8	2.74	40.3	49.3	56.5	57.7
Retrieval with evidence adapter									
AnatoRAG (Libra-3B)	<u>51.6</u>	25.3	37.5	32.3	2.71	42.9	50.8	57.3	57.3
AnatoRAG (BLIP-3)	**54.8**	**27.7**	**40.1**	**34.9**	**2.61**	**50.8**	**59.8**	**65.1**	**67.5**

Table 2: Report generation on MIMIC-CXR Findings. Ma and Mi are macro- and micro-averaged CheXbert F_1 over the 14- and 5-label inventories; every column is a percentage except $CliQ_0$, and an unmarked column is higher is better. † marks values displayed by the official Libra repository, and ‡ marks our evaluation of the released Libra-3B checkpoint without retrieval. Published rows retain the corresponding source protocols and provide context. The BLIP-3 rows form the controlled comparison; the Libra-3B rows document checkpoint integration and are not an equal-budget ablation. Bold and underline mark the best and second-best reported value in each column; tied values receive the same mark. Each owned configuration is one trained checkpoint.

the Libra-3B evidence projections, adapter, and language LoRA are adapted jointly, the released Libra-3B row is a checkpoint reference rather than an equal-budget ablation.

Within the controlled BLIP-3 comparison, macro-F_1 improves more than micro-F_1 for both label inventories, so the gain persists when labels receive equal weight. The separate Libra-3B integration shows that the evidence interface can also be implemented with Libra's native temporal connector. Comparing the retrieved text with the BLIP-3 report shows where the measured score differences appear. On the fourteen-label inventory, the backbone starts at 40.3/56.5 macro/micro-F_1, AnatoRetriever supplies evidence at 50.3/63.0, and the generated report reaches 50.8/65.1. Thus, most of the observed macro difference is already present in the selected evidence, while the generated report adds a further 2.1 micro-F_1 points. The five-label scores follow the same ordering: 49.3/57.7 for the generator without retrieval, 58.7/64.9 for retrieved evidence, and 59.8/67.5 for the report. These stage-wise scores locate the measured changes without treating automated metric differences as a causal decomposition.

Method (RG-F_1, $CliQ_0{}^{\downarrow}$)	**Findings**
Reference Findings:	**Lines and Tubes:** Stable right IJ line tip position. **Lungs:** Low lung volumes with mild worsening of pulmonary edema. **Pleura:** Small left pleural effusion. **Mediastinum:** Stable cardiomegaly. **Bony thorax:** No change. 
Backbone: 21.1 / 2.975	**Lungs:** Well inflated with interval improvement in pulmonary edema. **Pleura:** There is no pleural effusion or pneumothorax. **Mediastinum:** There is no cardiomegaly. Mediastinal silhouette is within normal limits. **Bony thorax:** No interval change.
Report-level RAG: 18.7 / 2.892	**Lungs:** Persistent moderate pulmonary edema. No lobar consolidation. **Pleura:** There is no pleural effusion. **Mediastinum:** No cardiomegaly. Mediastinal silhouette is within normal limits. **Bony thorax:** No interval change.
AnatoRAG (BLIP-3): 45.8 / 1.851	**Lungs:** Persistent low lung volumes with unchanged mild pulmonary edema. **Pleura:** There is a small left pleural effusion. No pneumothorax. **Mediastinum:** Stable mild cardiomegaly and mediastinal silhouette. **Bony thorax:** No interval change.

Figure 4: Case I, test index 1671: the reference, the backbone, report-level RAG, and AnatoRAG (BLIP-3) on one study. Green marks reference-supported content; red marks contradictions, unsupported additions, and attribute mismatches. Uncolored text is not adjudicated by the reference.

4.3 Qualitative Evidence-to-Report Trace

Figure 4 illustrates the systems' different outputs on one report. The backbone reverses the reference findings for pleural effusion and cardiomegaly. Report-level RAG recovers pulmonary edema but repeats both reversals, whereas the complete AnatoRAG (BLIP-3) system recovers all reference findings, including low lung volume, edema, left pleural effusion, cardiomegaly, and the unchanged bony thorax. The second case is a failure case. The backbone omits chronic right-rib deformities and pleuroparenchymal scarring, while report-level RAG adds a pleural effusion and left-rib deformity absent from the reference. AnatoRAG (BLIP-3) restores the right-rib deformity and retains the negative pleural finding, but AnatoRAG (BLIP-3) does not state the reference pleuroparenchymal scarring and instead attributes the costophrenic-angle appearance to pleural thickening. Together, the cases illustrate a successful output and a remaining failure mode in which label-compatible retrieved evidence is not expressed as the reference finding.

4.4 Ablations

Effect of evidence configuration. We show in Table 3 how the retrieval path and evidence delivery vary across five configurations. Within each retrieval path, the raw and adapted rows isolate the delivery change. Comparisons between report-level and anatomy-linked rows additionally change the retriever and therefore measure the complete evidence-source systems. Raw report evidence adds only 2.1 CheXbert-14 micro-F_1 points and worsens RadCliQ-v0 from 2.74 to 2.76. Adapting the same evidence raises macro-F_1 from 42.1 to 46.4 and improves RadCliQ-v0 to 2.66, a difference associated with structured delivery. Raw anatomy-linked evidence statements likewise improve CheXbert F_1 but reduce BLEU-4 and RadGraph F_1 and worsen RadCliQ-v0. Adaptation reverses all three changes and adds 7.5 macro-F_1 points over the same evidence statements. Between the adapted rows, the anatomy-linked retrieval path yields 4.4 higher macro-F_1, 2.0 higher micro-F_1, and 1.0 higher RadGraph F_1. These differences are consistent with the two complete retrieval paths supplying different evidence and with adaptation improving how the generator expresses the evidence. The crossed metric response shows why retrieval quality alone is insufficient. Raw evidence statements improve CheXbert F_1 but coincide with lower RadGraph F_1 and worse RadCliQ-v0; adaptation preserves the label gains and improves both relation-sensitive metrics. The difference is larger for anatomy-linked evidence statements: adding slots raises BLEU-4, RadGraph F_1, and CheXbert-14 macro-F_1 by 3.6, 3.6, and 7.5 points, compared with 1.2, 1.0, and 4.3 for report evidence. This pattern is consistent with structured delivery helping the generator use disjoint evidence statements; the comparison does not directly measure relation or discourse preservation.

Effect on external cohorts.

Table 4 shows that AnatoRAG improves every column over BLIP-3 on both cohorts. On CheXpert Plus, generation also exceeds retrieval-only AnatoRetriever in all four columns; on IU X-Ray, retrieval-only AnatoRetriever remains stronger in all four columns, although the micro-F_1 gaps are

Method	Report evidence	Anatomy evidence	Adapter slots	B-4	^{14}Ma	^{5}Ma	^{14}Mi	^{5}Mi	RG-F_1	CliQ$_0^{\downarrow}$
No evidence	✗	✗	✗	24.6	40.3	49.3	56.5	57.7	32.8	2.74
Report-level evidence	✓	✗	✗	25.3	42.1	52.4	58.6	60.2	32.9	2.76
Report-level evidence + adapter	✓	✗	✓	<u>26.5</u>	<u>46.4</u>	<u>56.7</u>	<u>63.1</u>	<u>65.1</u>	<u>33.9</u>	<u>2.66</u>
Anatomy-linked evidence	✗	✓	✗	24.1	43.3	53.6	59.6	61.0	31.3	2.78
AnatoRAG (BLIP-3)	✗	✓	✓	**27.7**	**50.8**	**59.8**	**65.1**	**67.5**	**34.9**	**2.61**

Table 3: Evidence-configuration ablation with one BLIP-3 generator and training protocol. The ✓/✗ columns denote report evidence, anatomy evidence, and adaptation; metrics follow Table 2. Bold and underline mark the best and second-best result.

Method	CheXpert Plus				IU X-Ray			
	^{14}Ma	^{5}Ma	^{14}Mi	^{5}Mi	^{14}Ma	^{5}Ma	^{14}Mi	^{5}Mi
AnatoRetriever	35.9	45.8	58.8	55.4	36.8	38.0	55.0	45.2
RADAR	<u>36.2</u>	<u>41.7</u>	<u>54.1</u>	<u>52.4</u>	**32.5**	–	**54.6**	–
BLIP-3	26.6	33.6	49.0	42.9	24.5	<u>31.5</u>	<u>49.8</u>	<u>38.2</u>
AnatoRAG (BLIP-3)	**36.5**	**48.1**	**60.4**	**58.8**	<u>32.1</u>	**33.5**	**54.6**	**45.1**

Table 4: External-cohort CheXbert F_1 for retrieval-only AnatoRetriever and generated reports.

small. The contrast locates the remaining cross-domain loss after evidence selection and suggests that evidence expression transfers less consistently than evidence availability.

5 Discussion

Separate retrieval and generation scores locate differences across admission, selection, and evidence delivery. The stages remain coupled: a missed admission is never searched, a false admission can retrieve unsupported text, and the IU X-Ray result shows that selected evidence can disappear during generation. The aligned CheXbert, RadGraph, and RadCliQ optimum of the complete configuration further shows gains beyond lexical overlap.

Where the gains arise. The retrieval analysis links the motivation to the measured gains: as positive-label count grows, exact report-level matches become sparse, whereas anatomy-linked statements remain available across anatomical groups. This difference explains why AnatoRetriever establishes most of the CheXbert gain before generation. Table 3 then isolates delivery: adapting anatomy-linked evidence statements improves every reported metric over raw evidence statements. Selection therefore determines the available content, while adaptation determines how consistently the selected content appears in the report. The IU X-Ray gap further shows that transfer at selection does not guarantee equally faithful expression, which motivates reporting retrieval-only and generation scores separately.

6 Conclusion

AnatoRAG treats retrieved statements as hypotheses to be verified rather than text to be copied. AnatoRetriever filters anatomy-linked evidence against the target image, and the Anatomy Evidence Adapter compresses selected statements and label scores into structured tokens for generation. This design addresses the compositional mismatch of report-level retrieval by assembling evidence across anatomical regions without importing an entire, only partially matched report. AnatoRAG achieves 65.1 CheXbert-14 micro-F_1, 34.9 RadGraph F_1, and 27.7 BLEU-4 on MIMIC-CXR, while reaching 60.4 and 54.6 CheXbert-14 micro-F_1 on CheXpert Plus and IU X-Ray without target-domain training. These results show that selecting compatible evidence and expressing that evidence are distinct requirements. The IU X-Ray gap motivates stronger evidence-to-text alignment while retaining image-conditioned filtering.

Limitations As with other systems built on multimodal large language models, AnatoRAG remains susceptible to errors and biases inherited from its pretrained components, which may propagate to generated reports. Accordingly, the system is intended for research use and requires independent clinical validation before deployment.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on research methodology and experiments and to assist with result interpretation. We did not use generative AI tools to generate synthetic data, ground-truth annotations, or reported experimental scores; required-disclosure tasks involving theoretical claims or proofs are not applicable to this work. Additionally, we used generative AI tools for literature search, manuscript organization and language editing, editing auxiliary research code, and preparing and formatting LaTeX tables and scientific figures. The authors verified citations and numerical values against the original sources and experiment records, and reviewed and tested AI-assisted code, text, and artifacts. We take responsibility for the final content of this work, including text, claims, code, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work concerns retrospective research on de-identified chest radiographs and associated reports. Use of MIMIC-CXR and any external evaluation data is subject to the applicable access terms and data-use agreements. Retrieved text remains associated with source records for research auditing. This provenance is not exposed as patient-identifying information in released examples. Clinical labelers and report generators can reproduce dataset biases and produce unsupported findings. The system is a research prototype and is not intended to replace radiologist interpretation or to support clinical decisions without independent clinical validation. No claim of patient benefit follows from automated metric improvements.

DATA AVAILABILITY STATEMENT

MIMIC-CXR is distributed through PhysioNet under credentialed access: users must complete the required training and sign the project data-use agreement (Johnson et al., 2019). The agreement does not permit us to redistribute MIMIC-CXR images, reports, or derived records that retain sufficient content to reconstruct the source reports. We will instead release the complete pipeline for regenerating our structured artifacts from an authorized local copy, including the Qwen-based report-structuring code and prompts, output schemas and parsing rules, the CheXbert-based label extraction and filtering procedure, evidence-bank construction code, and all preprocessing and experiment parameters. IU X-Ray is available from the Indiana University Chest X-ray collection hosted by the U.S. National Library of Medicine Open-i (Demner-Fushman et al., 2016); Open-i asks researchers not to redistribute the collection outside their research group or organization. CheXpert Plus is available through the Stanford AIMI Shared Datasets portal under the Stanford Dataset Research Use Agreement (Chambon et al., 2024), which permits research use but prohibits redistribution of dataset copies. Our release will therefore contain code and documentation, but no images, reports, or record-level structured outputs from these datasets.

REPRODUCIBILITY STATEMENT

Section 3 specifies the separation between queries, admission scores, selection rules, and generator conditioning, and Section 4 specifies score aggregation, patient exclusion, and the control each comparison uses. Appendices B–E give bucket membership, the selection procedure, model dimensions, and training-stage provenance, and Appendix F records which rows we control and which are quoted from the source publications. The released code reconstructs the structured artifacts from an authorized local dataset copy and preserves patient identity across exclusion checks, the fixed label order, input/reference pairing, and the model-selection history.

REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2310.11511.

Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Pérez-García, Maximilian Ilse, Daniel C. Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, Anton Schwaighofer, Maria Wetscherek, Matthew P. Lungren, Aditya Nori, Javier Alvarez-Valle, and Ozan Oktay.

Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15016–15027, June 2023.

Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, Felix Meissen, Mercy Ranjit, Shaury Srivastav, Julia Gong, Noel C. F. Codella, Fabian Falck, Ozan Oktay, Matthew P. Lungren, Maria Teodora Wetscherek, Javier Alvarez-Valle, and Stephanie L. Hyland. MAIRA-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*, 2024. URL https://arxiv.org/abs/2406.04449.

Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven Q. H. Truong, Chu The Chuong, and Curtis P. Langlotz. CheXpert Plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats. *arXiv preprint arXiv:2405.19538*, 2024. URL https://arxiv.org/abs/2405.19538.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1439–1449, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.112. URL https://aclanthology.org/2020.emnlp-main.112/.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. Cross-modal memory networks for radiology report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5904–5914, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.459. URL https://aclanthology.org/2021.acl-long.459/.

Sheng Cheng and Devika Subramanian. Rethinking radiology report generation: From narrative flow to topic-guided findings. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=nV3SAjFlyv.

Francesco Dalla Serra, Chaoyang Wang, Fani Deligianni, Jeff Dalton, and Alison O'Neil. Controllable chest X-ray report generation from longitudinal representations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 4891–4904. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.325. URL https://aclanthology.org/2023.findings-emnlp.325/.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. Improving the factual correctness of radiology report generation with semantic rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4348–4360. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-emnlp.319. URL https://aclanthology.org/2022.findings-emnlp.319/.

Jean-Benoit Delbrouck, Pierre Chambon, Zhihong Chen, Maya Varma, Andrew Johnston, Louis Blankemeier, Dave Van Veen, Tan Bui, Steven Truong, and Curtis Langlotz. RadGraph-XL: A large-scale expert-annotated dataset for entity and relation extraction from radiology reports. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12902–12915, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.765. URL https://aclanthology.org/2024.findings-acl.765/.

Jean-Benoit Delbrouck, Justin Xu, Johannes Moll, Alois Thomas, Zhihong Chen, Sophie Ostmeier, Asfandyar Azhar, Kelvin Zhenghao Li, Andrew Johnston, Christian Bluethgen, Eduardo Pontes Reis, Mohamed S Muneer, Maya Varma, and Curtis Langlotz. Automated structured radiology report generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26813–26829, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1301. URL https://aclanthology.org/2025.acl-long.1301/.

Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016. doi: 10.1093/jamia/ocv080.

Mark Endo, Rayan Krishnan, Viswesh Krishna, Andrew Y. Ng, and Pranav Rajpurkar. Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In Subhrajit Roy, Stephen Pfohl, Emma Rocheteau, Girmaw Abebe Tadesse, Luis Oala, Fabian Falck, Yuyin Zhou, Liyue Shen, Ghada Zamzmi, Purity Mugambi, Ayah Zirikly, Matthew B. A. McDermott, and Emily Alsentzer (eds.), *Proceedings of Machine Learning for Health*, volume 158 of *Proceedings of Machine Learning Research*, pp. 209–219. PMLR, 04 Dec 2021. URL https://proceedings.mlr.press/v158/endo21a.html.

Tomer Erez, Moshe Kimhi, Chaim Baskin, and Ehud Rivlin. CXR-Retrieve: Compositional text-to-image retrieval in chest radiography, July 2026. URL https://arxiv.org/abs/2607.27779.

Debanjan Goswami, Ronast Subedi, and Shayok Chakraborty. MediVLM: A vision language model for radiology report generation from medical images. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 10287–10304, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.544. URL https://aclanthology.org/2025.findings-emnlp.544/.

Wenjun Hou, Yi Cheng, Kaishuai Xu, Heng Li, Yan Hu, Wenjie Li, and Jiang Liu. RADAR: Enhancing radiology report generation with supplementary knowledge injection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26366–26381, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1279. URL https://aclanthology.org/2025.acl-long.1279/.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Shih-Cheng Huang, Liyue Shen, Matthew P. Lungren, and Serena Yeung. GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3942–3951, 2021. doi: 10.1109/ICCV48922.2021.00391. URL https://openaccess.thecvf.com/content/ICCV2021/html/Huang_GLoRIA_A_Multimodal_Global-Local_Representation_Learning_Framework_for_Label-Efficient_Medical_ICCV_2021_paper.html.

Stephanie L. Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Mercy Ranjit, Anton Schwaighofer, Fernando Pérez-García, Valentina Salvatelli, Shaury Srivastav, Anja Thieme, Noel C. F. Codella, Matthew P. Lungren, Maria Teodora Wetscherek, Ozan Oktay, and Javier Alvarez-Valle. MAIRA-1: A specialised large multimodal model for radiology report generation. *arXiv preprint arXiv:2311.13668*, 2023. URL https://arxiv.org/abs/2311.13668.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Q. H. Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis P. Langlotz, and Pranav Rajpurkar. RadGraph: Extracting clinical entities and relations from radiology reports. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/c8ffe9a587b126f152ed3d89a146b445-Abstract-round1.html.

Pengcheng Jiang, Cao Xiao, Minhao Jiang, Parminder Bhatia, Taha Kass-Hout, Jimeng Sun, and Jiawei Han. Reasoning-enhanced healthcare predictions with knowledge graph community retrieval. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2410.04585.

Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. PromptMRG: Diagnosis-driven prompts for medical report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 2607–2615, 2024. doi: 10.1609/aaai.v38i3.28038. URL https://ojs.aaai.org/index.php/AAAI/article/view/28038.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6 (1):317, 2019. doi: 10.1038/s41597-019-0322-0. URL https://doi.org/10.1038/s41597-019-0322-0.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81. Association for Computational Linguistics, 2004. URL https://aclanthology.org/W04-1013/.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318. Association for Computational Linguistics, 2002. doi: 10.3115/1073083.1073135.

Jonggwon Park, Byungmu Yoon, Soobum Kim, and Kyoyun Choi. RA-RRG: Multimodal retrieval-augmented radiology report generation with key phrase extraction. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 5029–5048, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.247. URL https://aclanthology.org/2026.findings-acl.247/.

Fernando Pérez-García, Harshita Sharma, Sam Bond-Taylor, Kenza Bouzid, Valentina Salvatelli, Maximilian Ilse, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Matthew P. Lungren, Maria Teodora Wetscherek, Noel Codella, Stephanie L. Hyland, Javier Alvarez-Valle, and Ozan Oktay. Exploring scalable medical image encoders beyond text supervision. *Nature Machine Intelligence*, January 2025. ISSN 2522-5839. doi: 10.1038/s42256-024-00965-w. URL https://doi.org/10.1038/s42256-024-00965-w.

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. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Constantin Seibold, Simon Reiss, Saquib Sarfraz, Matthias A. Fink, Victoria Mayer, Jan Sellner, Moon Sung Kim, Klaus H. Maier-Hein, Jens Kleesiek, and Rainer Stiefelhagen. Detailed annotations of chest x-rays via CT projection for report understanding. In *Proceedings of the 33rd British Machine Vision Conference*, 2022. URL https://bmvc2022.mpi-inf.mpg.de/58/.

Phillip Sloan, Edwin Simpson, and Majid Mirmehdi. Spatio-temporal and clinical conditioning for fine-grained radiology report retrieval, 2026. URL https://arxiv.org/abs/2607.02024.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1500–1519, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.117. URL https://aclanthology.org/2020.emnlp-main.117/.

Liwen Sun, James Jialun Zhao, Wenjing Han, and Chenyan Xiong. 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)*, pp. 643–655, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.28. URL https://aclanthology.org/2025.naacl-long.28/.

Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7433–7442, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Tanida_Interactive_and_Explainable_Region-Guided_Radiology_Report_Generation_CVPR_2023_paper.html.

Lehan Wang, Haonan Wang, Honglong Yang, Jiaji Mao, Zehong Yang, Jun Shen, and Xiaomeng Li. Interpretable bilingual multimodal large language model for diverse biomedical tasks. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2410.18387.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. R2GenGPT: Radiology report generation with frozen LLMs. *arXiv preprint arXiv:2309.09812*, 2023. URL https://arxiv.org/abs/2309.09812.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. MedCLIP: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3876–3887. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.256. URL https://aclanthology.org/2022.emnlp-main.256/.

Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. MMed-RAG: Versatile multimodal RAG system for medical vision language models. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2410.13085.

Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, et al. xGen-MM (BLIP-3): A family of open large multimodal models, 2024. URL https://arxiv.org/abs/2408.08872.

Sixing Yan, Ziao Wang, Kejing Yin, William K. Cheung, Ka Chun Cheung, and Simon See. Learning self-critiquing mechanisms for region-guided chest X-ray report generation. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=6sOSwgCmpH.

Kihyun You, Jawook Gu, Jiyeon Ham, Beomhee Park, Jiho Kim, Eun Kyoung Hong, Woonhyuk Baek, and Byungseok Roh. CXR-CLIP: Toward large scale chest x-ray language-image pre-training. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pp. 101–111. Springer Nature Switzerland, 2023. URL https://arxiv.org/abs/2310.13292.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y. Ng, Curtis P. Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9):100802, 2023. doi: 10.1016/j.patter.2023.100802.

Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, et al. A clinically accessible small multimodal radiology model and evaluation metric for chest x-ray findings. *Nature Communications*, 16(1):3108, 2025. doi: 10.1038/s41467-025-58344-x. URL https://doi.org/10.1038/s41467-025-58344-x.

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, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. BiomedCLIP: A multimodal biomedical foundation model pretrained

from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2024. URL https://arxiv.org/abs/2303.00915.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

Xi Zhang, Zaiqiao Meng, Jake Lever, and Edmond S. L. Ho. Libra: Leveraging temporal images for biomedical radiology analysis. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 17275–17303, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.888. URL https://aclanthology.org/2025.findings-acl.888/.

Tianshuo Zhou, Sen Mei, Xinze Li, Zhenghao Liu, Chenyan Xiong, Zhiyuan Liu, Yu Gu, and Ge Yu. MARVEL: Unlocking the multi-modal capability of dense retrieval via visual module plugin. *arXiv preprint arXiv:2310.14037*, 2024. URL https://arxiv.org/abs/2310.14037.

Supplementary Material for
AnatoRAG: Anatomy-Guided Retrieval for Chest X-ray Report Generation

Contents

A Additional Related Work **17**

B Clinical Evidence Bank and Label Vocabulary **18**

B.1 Label Prevalence and Report Composability 19

B.2 Multi-label Matching 19

C Retriever Implementation and Selection Algorithm **20**

C.1 Candidate-Pool Depth 22

C.2 Coverage-Bonus Sensitivity 23

D Anatomy Evidence Adapter Token Budget **23**

E Training Stages and Conditioning Details **24**

F Comparison Protocol and Quoted Sources **26**

G Case Studies **27**

H Retriever Component Ablation Details **30**

I Retrieval by Reference Label Count **30**

J Additional Generation Metrics **31**

K Datasets and Evaluation Cohorts **34**

L Metric Definitions and Scoring Procedure **35**

Appendix organization. The supplement proceeds from literature and the training-only evidence bank to retriever and adapter implementation, followed by controlled comparisons, case studies, ablations, dataset protocols, and metric definitions. Retrieval-only evidence quality and end-to-end report generation remain separate throughout, and replay diagnostics are identified as such.

A Additional Related Work

This section places the retrieval unit, admission rule, anatomy constraint, and evidence interface in the wider literature.

Retrieval units and admission. Retrieval-augmented systems differ in both retrieval unit and selection rule. FactMM-RAG trains whole-report retrieval with factual agreement (Sun et al., 2025). RADAR filters observations from a draft using classification output (Hou et al., 2025). RA-RRG retrieves key phrases (Park et al., 2026), while MMed-RAG trains a medical retriever and selects passages (Xia et al., 2025). MediVLM studies image-to-report generation with a dedicated alignment stage (Goswami et al., 2025). KARE indexes biomedical knowledge-graph communities and retrieves the community summaries (Jiang et al., 2025). AnatoRetriever selects label-linked evidence statements using a separate admission decision and a compatibility-aware selector.

Self-RAG trains a generator to decide when to retrieve and to critique retrieved evidence and generated text during decoding (Asai et al., 2024). AnatoRetriever makes the admission and compatibility decisions before decoding, using the image-derived finding set. This placement lets us score retrieved text separately from generated reports.

CXR-RePaiR retrieves individual report sentences with contrastive image–text representations (Endo et al., 2021). CXR-Retrieve composes evidence for a specified finding configuration (Erez et al., 2026). MARVEL adds visual inputs to a dense text retriever outside the clinical domain (Zhou et al., 2024). Our evidence statements are indexed by label, allowing each candidate's assertions to be checked against the admitted label set.

Anatomy and representation. Region-guided models condition generation on anatomical regions (Tanida et al., 2023), predicted findings (Jin et al., 2024), or region-level self-critique (Yan et al., 2026). CXAS provides chest X-ray anatomical masks (Seibold et al., 2022). AnatoRAG uses these masks to constrain query attention and statement search before generation. The selected evidence statement retains the linked anatomical family for subsequent adaptation.

Contrastive image–text encoders include CLIP (Radford et al., 2021), GLoRIA (Huang et al., 2021), MedCLIP (Wang et al., 2022), CXR-CLIP (You et al., 2023), and BiomedCLIP (Zhang et al., 2024). RAD-DINO uses image-only pretraining (Pérez-García et al., 2025), and BioViL-T models temporal information (Bannur et al., 2023). Our retriever uses RAD-DINO image features and BioViL-T report embeddings. Both encoders remain frozen during retriever training and generator adaptation.

Generation interfaces. Earlier report generators model long-range image–text dependencies with memory (Chen et al., 2020; 2021), region representations (Tanida et al., 2023), clinical prompts (Jin et al., 2024), or temporal information (Dalla Serra et al., 2023). BLIP-2 introduced an interface between frozen visual and language models (Li et al., 2023). R2GenGPT, MAIRA-1, MAIRA-2, LLaVA-Rad, and Libra adapt multimodal generators to radiology (Wang et al., 2023; Hyland et al., 2023; Bannur et al., 2024; Zambrano Chaves et al., 2025; Zhang et al., 2025); low-rank adaptation reduces the number of trained language-model parameters (Hu et al., 2022). AnatoRAG changes the retrieved-evidence interface of a BLIP-3/XGen-MM backbone.

Table S1 compares the inference-time inputs used by recent report-generation systems. The table separates the current frontal and lateral views, longitudinal patient context, and text retrieved from other training studies. MAIRA-2 accepts an optional current lateral view in addition to the current frontal image; RA-RRG supports multiple current views by aggregating retrieved phrases. The Libra-3B configuration retains the released Libra visual and prompt interface and adds gated retrieved evidence. The BLIP-3 configuration uses the BLIP-3 current/prior visual interface, includes available prior Findings in the clinical prompt, and also requires retrieved evidence.

Evaluation and transfer. CheXbert supplies the 14-label finding inventory (Smit et al., 2020). RadGraph and RadGraph-XL represent report entities and relations (Jain et al., 2021; Delbrouck et al., 2024); RadCliQ combines clinical error signals (Yu et al., 2023). Structured report generation

Method	Current frontal	Current lateral	Prior frontal	Prior report	Clinical context	Retrieved evidence
MAIRA-1 (Hyland et al., 2023)	✓	✗	✗	✗	○	✗
MAIRA-2 (Bannur et al., 2024)	✓	○	○	○	○	✗
LLaVA-Rad (Zambrano Chaves et al., 2025)	✓	✗	✗	✗	○	✗
Libra (Zhang et al., 2025)	✓	✗	○	✗	○	✗
FactMM-RAG (Sun et al., 2025)	✓	✗	✗	✗	✗	✓
RADAR (Hou et al., 2025)	✓	✗	○	○	○	✓
LLaVA-TA (Cheng & Subramanian, 2026)	✓	✗	✗	✗	✗	✗
RadSCR (Yan et al., 2026)	✓	✗	✗	✗	✗	✓
RA-RRG (Park et al., 2026)	✓	○	✗	✗	✗	✓
AnatoRAG (Libra-3B)	✓	✗	○	✗	○	✓
AnatoRAG (BLIP-3)	✓	✗	○	○	○	✓

Table S1: Inference-time inputs of recent chest X-ray report generators. ✓ denotes a required input, ○ an input used when available, and ✗ an input not used by the reported configuration. Frontal and lateral are views from the current study; Prior report denotes the same patient's earlier report or Findings. Clinical context covers indication, history, technique, and comparison fields. Retrieved evidence is produced internally from other training studies and is required by each retrieval-augmented method shown.

and factuality scoring have also been studied with section-level targets and CheXprompt (Delbrouck et al., 2025; Zambrano Chaves et al., 2025). We report lexical metrics for comparison with published systems (Papineni et al., 2002; Lin, 2004; Zhang et al., 2020). Applying the same labeler to retrieved and generated text locates changes in clinical labels across the two stages.

MIMIC-CXR (Johnson et al., 2019) supplies training, development, and test data and fixes the report-section convention we follow. CheXpert Plus (Chambon et al., 2024) and IU X-Ray (Demner-Fushman et al., 2016) are used only as external transfer targets, with the evidence bank and all retrieval indexes left as trained on MIMIC-CXR. The eligibility rule and the resulting case counts accompany each transfer result.

B Clinical Evidence Bank and Label Vocabulary

The bank keeps one index per anatomical family, and laterality is retained as metadata and as text. A candidate can express several labels within the assigned family, and an entry may carry language that reaches past the named anatomical structure. No Finding entries are complete Findings sections and can mention several negative observations. Table S2 fixes the family-to-label key, and Figure S1 gives the size of each family.

The label order used for queries, scores, and metadata is: Atelectasis, Cardiomegaly, Consolidation, Edema, Enlarged Cardiomediastinum, Fracture, Lung Lesion, Lung Opacity, No Finding, Pleural Effusion, Pleural Other, Pneumonia, Pneumothorax, and Support Devices. Candidates are drawn from training reports only, and the findings selected for downstream training exclude the query patient. Cached report features are computed with a 128-token input limit, which truncates 54 entries, while complete candidate text remains available at output. Positive image–text pairs used to learn alignment can share a training study.

Statement-extraction contract. The extractor receives one target sentence and returns `{sentence_index, atoms, needs_context, review_reason}`, where each atom contains `atomic_text`, exact `evidence_spans`, `needs_review`, and `review_reason`. The prompt requires one independent clinical assertion per atom; preservation of polarity, uncertainty, temporal change, severity, count, laterality, anatomy, and device position; and forbids facts drawn from other sentences. Coordinated findings are split, while attributes of one observation remain together. Parsing verifies the exact key set, sentence index, Boolean fields, nonempty atoms, and character-for-character evidence spans. Malformed outputs are retried at most twice and otherwise rejected. Anatomy-family assignment follows Table S2, and evidence statement and source-report CheXbert labels determine F_j^+ and F_j^-. The supplementary source archive includes the complete extraction instruction and all few-shot

Family	Positive-label membership
Pulmonary parenchyma	Atelectasis, Consolidation, Edema, Lung Lesion, Lung Opacity, Pneumonia
Pleural	Pleural Effusion, Pleural Other, Pneumothorax
Cardiomediastinal	Cardiomegaly, Enlarged Cardiomediastinum
Osseous	Fracture
Support devices	Support Devices
Global normal	No Finding

Table S2: Saved training-only clinical evidence bank: the key from each anatomical family to the positive labels available to entries in that family. An entry belongs to the family assigned by the recorded anatomy and can be retrieved under any label listed for that family. Explicit negative labels, laterality, and source identifiers are retained alongside these positive memberships.

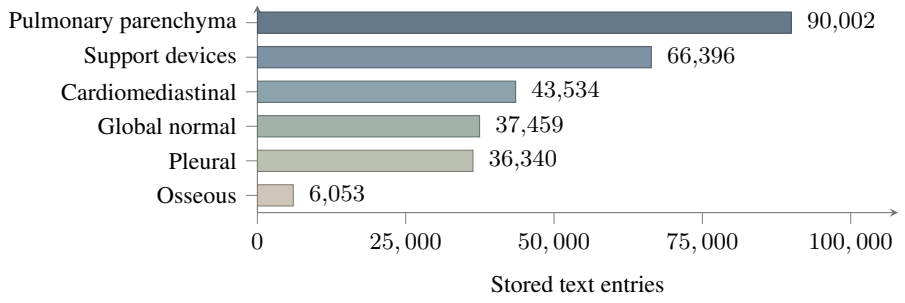


Figure S1: Entries per anatomical family in the saved clinical evidence bank, on a common linear scale. The bank is deliberately unbalanced: pulmonary parenchyma holds fifteen times the entries of the osseous family, which is why family partitioning buys separately searchable indices rather than balanced labels.

examples in `artifacts/evidence statement_extraction_prompt.txt` (SHA-256 `37aa8a79f92b9d7ba15213a218844b32d02437470bc637cd106d90cf485be5b2`).

B.1 Label Prevalence and Report Composability

A separate corpus census distinguishes original-report labels from retained-evidence statement labels. The corpus census reconstructs 162,954 available training-image records and counts each of 146,120 studies once after verifying agreement of labels across views. These studies cover 58,685 patients, all disjoint from the test patients. The report-derived targets are distinct from statement-label unions. Support Devices occurs in 28.3% of studies and Pleural Other in 1.9%, a 15.0-fold prevalence ratio. This census covers the image gallery used by AnatoRetriever; the BiomedCLIP control gallery contains 146,121 studies.

The pooled clinical evidence bank has a 32.8-fold difference between the largest and smallest positive-label counts. Imbalance also persists within the pulmonary and pleural families, with ratios of 10.6 and 14.9. Family partitioning therefore provides separately searchable evidence statements, rather than balancing the underlying labels. The study-level census finds that 63,743 of 146,120 training studies (43.6%) have positive labels in at least two clinical families. Excluding No Finding and Support Devices leaves 45,206 studies spanning multiple disease families: 30.9% of all studies and 48.6% of the 92,936 disease-positive studies. These counts describe co-occurrence, not retrieval failure. The report-derived targets also contain 165 studies (0.1%) with both No Finding and a disease label, showing potential label-state inconsistencies. Multi-positive evidence statements account for 5.4% of the bank. 35.7% of rows repeat a normalized text already present. This template reuse motivates text deduplication.

B.2 Multi-label Matching

This analysis uses the common scoring pass over the same 2,310 queries with at least one positive reference label. The strata contain 562, 712, 549, and 487 queries with one, two, three, and at least four labels. For reference-positive set T_i and output-positive set S_i, per-query label recall is

$r_i = |T_i \cap S_i|/|T_i|$. We average r_i over queries within each stratum. This is distinct from pooled micro-recall, class-macro recall, and the fraction of queries with $r_i = 1$. Exact-match (subset) accuracy is the mean of $\mathbf{1}[T_i = S_i]$, so a query counts only when the predicted set equals the reference set exactly. The 151 queries with no positive reference are excluded because per-case recall is undefined. Pointwise intervals use 2,000 patient-bootstrap resamples, recomputing each stratum's denominator.

Among 1,748 multi-label queries, mean recall is 32.8% for BiomedCLIP Rank-1 and 71.1% for AnatoRetriever. Exact-match accuracy is 1.6% and 9.6%. The paired differences are 38.3 points [36.4, 40.2] for mean recall and 8.0 points [6.3, 9.8] for exact matching. Mean recall increases with label count in these observations, while exact matching is lower for larger sets. AnatoRetriever also predicts more extra positives on multi-label queries (1.3 versus 1.0 per query), so the added recall arrives alongside more predicted positives rather than fewer. These intervals describe patient-sampling uncertainty.

Combination availability. The 146,120-study training corpus contains a positive-label set exactly matching the target for 2,297 of 2,310 nonempty queries (99.4%). The proportion remains 97.3% for four or more labels. The combinations are present in the corpus, so low matching has another source. This check uses report-derived labels; the BiomedCLIP control gallery contains one additional study. Exact positive sets also omit distinctions among absent, uncertain, and unmentioned findings.

A label-space search over the whole bank can compose every nonempty target set, even after excluding evidence statements whose stored explicit-negative labels intersect the target positives. The minimum uses a median of two evidence statements and at most five. Every singleton-positive pattern exists in the bank, making this a generous availability check: the analysis tests whether the bank holds the combinations, leaving gates, the $K = 64$ shortlist, ranking, deduplication, and rescoring of concatenated text to the retrieval experiments.

Removing No Finding and Support Devices from predictions and references, micro-F1 over the remaining 12 labels and all 2,461 queries is 33.8% for BiomedCLIP Rank-1 and 59.5% for AnatoRetriever (difference 25.7 points, paired patient-bootstrap 95% CI [23.9, 27.6]). Among 1,244 queries whose reference positives span at least two of four disease families, mean per-query recall is 30.8% versus 69.7%. Extra predicted positives average 0.8 versus 1.1 per query. These analyses rescore paired retrieval outputs under a restricted label set.

C Retriever Implementation and Selection Algorithm

Algorithms S1 and S2 trace one case from image to report and then expand the retriever itself. The section specifies the encoder outputs, anatomy-mask constraints, class-specific admission composition, and deterministic selection before analyzing candidate-pool depth and the coverage bonus.

Algorithm S1 Inference for one case

Input: current radiograph I_{cur}, optional prior I_{pri}, clinical prompt P; bank $\mathcal{B}$; frozen encoder and image-label classifier; generator
Output: the Findings report Y

Encode what the image branch returns
patch features H_{cur}, classifier logits s_ℓ, queries q_ℓ ▷ Eq. 1
Retrieve expanded in Algorithm S2
$A \leftarrow$ the admitted findings ▷ fused scores r_ℓ, Eq. 2
$E \leftarrow$ compatible evidence statements for the admitted findings ▷ Eq. 5
Generate what the language model is given
$\tilde{\mathbf{p}} \leftarrow$ the calibration of the same logits s_ℓ ▷ priors, not the scores r_ℓ
$(Z_E, Z_S) \leftarrow \text{Adapter}(E, \tilde{\mathbf{p}})$ ▷ Eq. 6
$Y \leftarrow$ the decoder on $[Z_I; Z; \text{Emb}(P)]$ ▷ $Z = [Z_E; Z_S]$
return Y

Encoder and shortlist configuration. RAD-DINO provides 768-dimensional patch features on a 37×37 grid and a 1,536-dimensional global representation concatenating CLS and mean patch features. Attention pooling retains the 16 highest-weighted permitted patches for each label. BioViL-

Label	Active admission branches after the direct score
Atelectasis	class head; global neighbours; frozen classifier
Cardiomegaly	direct score only
Consolidation	class head; global neighbours
Edema	class head; global neighbours; frozen classifier
Enlarged Cardiomediastinum	frozen classifier
Fracture	regional neighbours; osseous subview
Lung Lesion	regional neighbours
Lung Opacity	direct score only
No Finding	direct score only
Pleural Effusion	global neighbours; regional neighbours
Pleural Other	class head; global neighbours; regional neighbours; frozen classifier
Pneumonia	frozen classifier
Pneumothorax	class head; regional neighbours
Support Devices	global neighbours; regional neighbours; device subview

Table S3: Per-label branch participation in the fixed admission-score composition. Branches are applied in the order stated in Section 3.1; labels marked direct score only bypass all later branches. All thresholds are fitted on training data and frozen for evaluation.

T CXR-BERT produces 128-dimensional normalized evidence statement embeddings, and each label uses a shortlist width of $K = 64$.

Anatomy-mask construction. The baseline query head uses predefined spatial pooling. The CXAS residual uses nine cached anatomical channels on the 37×37 patch grid. A separate six-region mask-construction pathway exists at original image resolution. The query branch carries a dedicated text input. The permitted mask per label is a constraint on where a query may attend, not a disease segmentation: bilateral lung masks serve Atelectasis, Consolidation, Lung Opacity, and Pneumonia; Edema adds cardiac context and Lung Lesion adds mediastinal context; Cardiomegaly uses the heart; Enlarged Cardiomediastinum uses heart and mediastinum; Fracture uses the osseous and chest-wall mask; pleural findings use a lung-boundary band with diaphragm context; Pneumothorax additionally admits lung patches; Support Devices and No Finding use the full image. An empty permitted mask falls back to global attention.

Admission-score composition. The direct branch is supplemented by BCE residuals for Atelectasis and Edema, an asymmetric head for Pneumothorax, a hard-sampling LoRA branch for Consolidation, and a uniform-sampling branch for Pleural Other. The ordered score composition is class-wise prediction, global image neighbours, regional query neighbours, subview scores, and image-label classifier scores. Neighbour counts vary by class from 4 to 256. The image-label classifier in Figure 2, initialized from the RADAR auxiliary checkpoint, contributes to Enlarged Cardiomediastinum, Pleural Other, Pneumonia, Atelectasis, and Edema. The image-label classifier uses a SwinV2-Large visual backbone with empty auxiliary text input. These branches can require multiple forward passes.

To make the frozen admission score explicit, let $p_\ell^{(0)}$ be the direct score after any class-head replacement and let $e_{\ell b}$ be the score supplied by branch b. Global and regional neighbour stages use the convex update

$$p_\ell^{(b)} = (1 - \alpha_{\ell b}) p_\ell^{(b-1)} + \alpha_{\ell b} e_{\ell b}. \tag{9}$$

Subview and frozen-classifier stages first standardize both inputs with training-set statistics and then return a probability,

$$p_\ell^{(b)} = \sigma\left((1 - \alpha_{\ell b}) \frac{p_\ell^{(b-1)} - \mu_{\ell b}^p}{\sigma_{\ell b}^p} + \alpha_{\ell b} \frac{e_{\ell b} - \mu_{\ell b}^e}{\sigma_{\ell b}^e} \right). \tag{10}$$

Inactive branches leave the score unchanged. The final r_ℓ is the last active-stage output. All α, normalization statistics, neighbour reductions, and τ_ℓ are estimated from training records and then frozen; evaluation labels never enter score composition or selection.

Deterministic selection. Algorithm S2 gives the ordered procedure. Admission determines the active findings from the composed label scores and the corresponding thresholds. Selection then chooses evidence statements for those findings, falling back to the nearest shortlisted candidate when none is compatible and to a neutral entry when no evidence statement is selected. The procedure is

Algorithm S2 Admission and selection in the frozen retriever

Input: image I_{cur} of patient u_I; bank $\mathcal{B} = \{(v_j, F_j^+, F_j^-, u_j)\}$; per-label queries q_ℓ and thresholds τ_ℓ; shortlist width K; coverage bonus λ_{cov}
Output: the selected findings E, each with the source identifier, bucket, laterality, labels, similarity and rank

Admission which findings are in play
$r_\ell \leftarrow$ fixed class-specific score composition ▷ Section 3.1
$A \leftarrow \{\ell : r_\ell \geq \tau_\ell\}$

Selection which evidence statement carries each finding
$E \leftarrow ()$; $U \leftarrow \varnothing$ ▷ U: labels already covered
for $\ell \in A$ in the fixed label order **do**
  if $\ell \notin U$ **then** ▷ an earlier choice already covers this label
    $\mathcal{C}_\ell \leftarrow$ the K nearest evidence statements with $\ell \in F_j^+$ and $u_j \neq u_I$
    $\mathcal{V}_\ell \leftarrow \{j \in \mathcal{C}_\ell : F_j^- \cap A = \varnothing,\ F_j^+ \subseteq A\}$ ▷ Eq. 4
    if $\mathcal{V}_\ell \neq \varnothing$ **then**
      $j^\star \leftarrow \arg\max_{j \in \mathcal{V}_\ell} \left[q_\ell^\top v_j + \lambda_{\text{cov}} |F_j^+ \setminus U|\right]$
    else
      $j^\star \leftarrow \arg\max_{j \in \mathcal{C}_\ell} q_\ell^\top v_j$ ▷ **fallback**: nearest candidate
    end if
    $E \leftarrow E \parallel j^\star$; $U \leftarrow U \cup F_{j^\star}^+$ ▷ duplicates skipped
  end if
end for
if $E = ()$ **then**
  $E \leftarrow$ one neutral text drawn from another patient ▷ **fallback**
end if
return E

K	No compatible pool	Incompatible selected	Rank > 1 (%)	Mean label statements	Covered skips
1	418	408	0.0	2.9	81
4	71	65	6.3	2.9	196
16	1	0	9.6	2.8	426
64	0	0	12.2	2.7	651

Table S4: Shortlist depth and evidence selection over 7,335 active-label shortlists. No compatible pool counts shortlists containing no compatible candidate; Incompatible selected counts chosen evidence statements that fail the metadata check against the predicted active set. Rank > 1 is the fraction selected below first similarity rank. Mean statements excludes the 97 neutral fallbacks present at every K, and Covered skips counts labels already covered by an earlier choice.

deterministic given the bank, takes no gradient from the generator, and runs once per image before generation.

C.1 Candidate-Pool Depth

The fixed configuration reported in the main tables uses $K = 64$, $\lambda_{\text{cov}} = 0.06$, and 72 finding slots. These values define the reported checkpoints before the appendix replays and capacity comparisons. The analyses below vary one setting at a time while keeping the trained model and remaining retrieval or generation settings fixed, and the varied settings are not used for model selection. We replay the frozen candidate cache on 2,461 queries from 285 patients; Table S4 reports the result. The model, label scores, thresholds, bank, eligibility rules, and selector remain fixed, and reference labels are not used for selection.

At $K = 64$, 814 of 6,684 label-selected evidence statements (12.2%) originate below first rank. Increasing K from 1 to 64 removes 418 incompatible-only pools and reduces incompatible selected evidence statements from 408 to 0, which is what retaining alternatives for compatibility and coverage-based selection buys. $K = 16$ already avoids incompatible final selections because the only unavailable label at $K = 16$ is covered by an earlier choice. Predicted active-label coverage is 100.0% for every K because selection processes the predicted active set.

λ_{cov}	Statements	Covered labels	^{14}Ma	^{14}Mi
0.00	2.95	2.98	49.5	62.4
0.03	2.85	2.98	<u>49.7</u>	62.5
0.06	2.72	2.98	**49.8**	<u>62.7</u>
0.12	2.60	2.98	<u>49.7</u>	**62.8**

Table S5: Retriever-only coverage-bonus analysis on the 2,461 test queries using train candidates. Statements and active labels covered are means per query; a neutral fallback counts as neither. Ma and Mi are macro- and micro-averaged CheXbert-14 F_1 on the selected text and are percentages; higher is better. Bold and underline mark the best and second-best score in each F_1 column; tied values receive the same mark.

C.2 Coverage-Bonus Sensitivity

The coverage bonus λ_{cov} in Eq. 5 favours an evidence statement that covers active labels that earlier selections have not covered. We replay the frozen Top-64 candidate cache at four values of λ_{cov} while keeping admission scores, thresholds, compatibility checks, and the evidence bank fixed. This post-hoc test-set analysis characterizes the frozen selector and is not used to select the reported configuration. Table S5 evaluates the selected text directly; the report generator is not invoked.

The replay preserves 2.98 covered labels per query while reducing the selected text from 2.95 to 2.60 evidence statements as the bonus increases. Across the four values, macro-F_1 ranges from 49.5 to 49.8 and micro-F_1 from 62.4 to 62.8; $\lambda_{\text{cov}} = 0.06$ gives 49.8/62.7 with 2.72 evidence statements per query.

D Anatomy Evidence Adapter Token Budget

The anatomy-guided retriever supplies a variable-length sequence to the adapter. On the 2,461-case test set, the selected evidence statements contain 71.1 language-model tokens on average (99th percentile: 162; maximum: 193) before adaptation. We compare evidence-token budgets of 36, 72, and 144 while retaining the same selected evidence, 14 label slots, generator, and training protocol. The 72-token architecture is the reported configuration; the 36- and 144-token rows are post-hoc capacity ablations and do not select the main checkpoint. The evidence tokens are distributed evenly across the six anatomical families. The 72-token design assigns one header and 11 content positions to each family.

Table S6 shows that the 72-slot interface has the strongest score in every reported column. Relative to 72 slots, 36 slots reduce 14-label macro-F_1 by 4.5 points and RG-F_1 by 1.8 points, while 144 slots reduce every reported score.

Adapter tokens complement the backbone visual interface. The frozen visual encoder and resampler produce 128 native tokens per image, so a current–prior pair contributes 256 visual tokens alongside the evidence interface. Clinical-prompt tokens retain the natural sequence length. Family-wise allocation prevents one long evidence statement or frequent anatomy from consuming the evidence budget, while presence masks suppress absent families.

Evidence tokens	B-4	^{14}Ma	^{5}Ma	^{14}Mi	^{5}Mi	RG-F_1	CliQ$_0^{\downarrow}$
36	<u>25.8</u>	<u>46.3</u>	54.6	<u>62.5</u>	<u>61.8</u>	<u>33.1</u>	<u>2.68</u>
72	**27.7**	**50.8**	**59.8**	**65.1**	**67.5**	**34.9**	**2.61**
144	24.3	45.7	<u>55.0</u>	60.9	61.4	32.8	2.71

Table S6: Generation results as the adapter evidence-token budget changes. Fourteen label tokens are retained in every row, and all other settings are fixed. Ma and Mi are macro- and micro-averaged CheXbert F_1 over the 14- and 5-label inventories. All columns except CliQ$_0$ are percentages and higher is better; CliQ$_0$ is lower is better. Bold and underline mark the best and second-best result in each column.

E Training Stages and Conditioning Details

Training proceeds in four stages: statement-bank and visual preparation, query and gate learning, Anatomy Evidence Adapter initialization, and generator adaptation. Retrieval is fixed before language-model adaptation, so report-generation gradients do not change the discrete evidence set.

Statement-bank construction and visual adaptation. Statement-bank construction uses Qwen2.5-3B-Instruct with greedy decoding. The RAD-DINO adaptation used 162,954 stored training image records, a batch size of 72, and 2,263 updates. LoRA rank, scale, and dropout were 16, 32, and 0.05, respectively, with visual and classification-head learning rates of 2×10^{-5} and 2×10^{-4}. Weighted sampling with replacement and dropping the incomplete final batch produced 162,936 sampled positions. The RAD-DINO base is frozen in the query-learning and downstream stages, which retain the adapted RAD-DINO checkpoint.

Query and gate learning. Baseline and residual query training use matched training-report evidence statements and label-conditioned contrastive objectives. Side/entity mismatches provide hard textual contrasts, and an anchoring term limits drift from the baseline query. Visual LoRA is frozen during query-residual training. The semantic gate learns a residual over current-image logits with weighted BCE, a 0.1 pairwise ranking term, and a 0.01 squared-residual penalty. The semantic gate optimizer is AdamW at 2×10^{-4}. The class-wise BCE and asymmetric heads use a 0.005 residual penalty and a learning rate of 3×10^{-4}. The selected class-wise BCE and asymmetric-head checkpoints are epochs 6 and 2, respectively, within a configured 12-epoch budget.

Admission branches. Independent hard- and uniform-sampling LoRA branches start from the same direct parent and update 393,216 adapter parameters at 10^{-5}. A logical batch of 84 contains three positive and three negative training samples per label. The branch loss weights are 1.0 for focused-label BCE, 0.25 for all-label BCE, 1.0 for query anchoring, and 0.5 for soft-gate anchoring. The selected checkpoints are hard step 200 and uniform step 100, corresponding to 16,800 and 8,400 sampled positions with repetition.

Generation backbones and BLIP-3 implementation. Report-generation experiments initialize the released XGen-MM/BLIP-3 checkpoint with the accompanying tokenizer and image processor (Xue et al., 2024). Separate Libra-3B evaluation retains the native Temporal Alignment Connector and current/prior handling (Zhang et al., 2025).

For Libra-3B, 72 finding slots and 14 label slots are projected to the Libra token width and concatenated after the native current/prior visual interface. The Temporal Alignment Connector, image encoder, and base language weights remain frozen; evidence projections, adapter, and language LoRA use the report-generation objective in Eq. 8. The no-retrieval row retains the released connector, prompt, visual preprocessing, and decoding while omitting the evidence interface.

The following constants and schedule describe BLIP-3. The frozen visual encoder and resampler produce 128 tokens per image, ordered prior then current when available. The backbone defines the token count and ordering. Finding projections match the decoder width, and the 72/14 finding/label design remains separate from the native visual interface. The BLIP-3 implementation defines subsequent generation analyses, ablations, and transfer protocols unless another backbone is named.

Anatomy Evidence Adapter. The BLIP-3 anatomy evidence adapter receives 3,072-dimensional frozen LLM token embeddings and adds source identity plus sinusoidal position information before contextual encoding. The adapter uses a hidden width of 768, 12 attention heads, and two query cross-attention layers. Each family receives one header and 11 content slots. The 72 finding and

14 label slots remain separate from the 128 visual slots per image; presence masks suppress absent families. The selected interface equally mixes projected and native values, disables family semantic anchors, and caps the learned output scale at 0.20 across generation ablations.

Image-score source and calibration. The frozen image-label classifier, initialized from the RADAR auxiliary checkpoint, processes the current radiograph with empty text and returns logits s_ℓ, the final branch scores in Eq. 2. The retrieval gate uses the logits in class-wise fusion. Downstream guidance instead applies train-fitted calibration, $\hat{p}_\ell = \sigma(a_\ell s_\ell + b_\ell)$. Training records use five patient-disjoint calibration folds. Evaluation uses fixed coefficients fitted on training data. For No Finding, $\tilde{p}_{\mathrm{NF}} = \hat{p}_{\mathrm{NF}}(1 - \max_{\ell \in \mathcal{D}} \hat{p}_\ell)$, where $\mathcal{D}$ excludes No Finding and Support Devices. Other classes retain $\hat{p}_\ell$. Class reliability is also estimated from training data. These probabilities are cached inputs and distillation targets. The classifier and classifier calibration sit outside the adapter and language-model parameter groups.

Shared and branch-specific parameters. Text projection, contextual encoding, and bucket cross-attention produce states shared by finding and label outputs. The label branch combines an image-score adapter with a gated projection of pooled states without gradient detachment. Label identities, score projection, text projection, and gates differ from the finding-head projections and content gates. Both branches share one adapter module and optimizer during report-conditioned training. Subsequent language adaptation updates both branches alongside LoRA; the frozen-adapter condition freezes both. The image-label classifier remains outside the parameter group.

Adapter initialization. The adapter is initialized for three passes while the generator remains frozen. A reconstruction-only warm-up occupies the first 0.2 pass. The remaining 2.8 passes combine the clinical auxiliary objective with reconstruction replay, using weights 0.05 and 0.25, respectively. The clinical objective groups evidence-label and report-label supervision, frozen-image distillation, and the family-content, label-token, and evidence-presence terms summarized by $\mathcal{L}_{\mathrm{label}}$ and $\mathcal{L}_{\mathrm{soft}}$ in Eq. 7. The final initialization checkpoint is passed directly to generator adaptation without intermediate checkpoint selection.

Retrieved-text labels, target-report labels, and frozen image scores supervise adapter initialization only. Generator adaptation disables these initialization losses and updates the adapter solely through $\mathcal{L}_{\mathrm{gen}}$. At inference, the calibrated prior $\tilde{\mathbf{p}}$ remains an input to the finding and label branches together with the selected evidence; the frozen classifier logits are not passed directly to the generator.

Generator adaptation and decoding. The reported BLIP-3 generator is optimized for 2.5 epochs with a micro-batch size of 4. AdamW uses learning rates of 1×10^{-5} for the adapter and 3×10^{-5} for language LoRA. The visual encoder, resampler, base language weights, and adapter auxiliary heads remain frozen, while gradient checkpointing is enabled. During this language-adaptation stage, adapter dropout is disabled and the adapter is updated solely through the report-generation loss in Eq. 8. The complete selected text, clinical context, and target reports are retained without truncation. Table S7 summarizes these settings.

Setting	Value
Training duration	2.5 epochs
Micro-batch size	4
Adapter learning rate	1×10^{-5}
Language LoRA learning rate	3×10^{-5}
Language LoRA	all linear layers; rank 64; scale 128; dropout 0.1
Gradient checkpointing	Enabled
Frozen modules	Visual encoder, resampler, base language model, and adapter auxiliary heads
Adapter dropout	Disabled
Adapter auxiliary-loss weight	0
Native-token value mixture	0.5
Family semantic-anchor tokens	Disabled
Maximum adapter output scale	0.20
Target reports	Complete; no truncation
Decoding	Deterministic beam search; 5 beams; length penalty 2.0
Generation limits	24–512 new tokens; repeated 8-grams blocked

Table S7: Settings for the 2.5-epoch BLIP-3 language-adaptation stage. The adapter initialization losses are disabled at this stage, while the adapter remains trainable through report-generation cross-entropy.

F Comparison Protocol and Quoted Sources

This section separates controlled comparisons under one fixed backbone from source-reported retrieval values and matched retrieval-only visualization.

Controlled comparisons. Within each generation backbone we fix the initial checkpoint, the current and prior records, the prompt, the tokenizer, the visual preprocessing, and the decoding, and we control the language-adaptation and evidence-encoding budgets within each ablation. The whole-report retrieval control uses frozen BiomedCLIP (Zhang et al., 2024) image–text representations and returns two complete reports; AnatoRetriever uses the retriever of Section 3.1. Both draw candidates from the training bank under patient exclusion. Our controlled generation rows are the BLIP-3 control of Table 2 and every row of Table 3. The released Libra-3B checkpoint is evaluated separately in Table S13.

Source-reported retrieval context. Table S8 records the retrieval unit, report-section convention, and reported metrics for each cited system. The report-retrieval rows come from FactMM-RAG (Sun et al., 2025). CXR-RePaiR's evaluation covers 3,678 MIMIC-CXR studies (Endo et al., 2021). RA-RRG E9/E10 are phrase-retrieval configurations (Park et al., 2026). Aggregates a source does not report stay blank.

Matched retrieval visualization. Figure S4 compares the frozen whole-report control at one and two returned reports against AnatoRetriever on the four CheXbert columns; all three curves use one scoring pass on the same cohort. The evidence statements lead on every column, by 20.6 points on ^{14}Mi-F_1 up to 24.9 on ^{14}Ma-F_1 against the two-report setting, while returning a second report closes only 3.5 to 6.9 points of that gap. The panel is scored on retrieved text before generation.

Method	Output	CheXbert F_1			
		^{14}Ma	^{5}Ma	^{14}Mi	^{5}Mi
CLIP (Radford et al., 2021)	Report	–	–	–	34.1
GLoRIA (Huang et al., 2021)	Report	–	–	–	34.6
MedCLIP (Wang et al., 2022)	Report	–	–	–	53.9
CXR-CLIP (You et al., 2023)	Report	–	–	–	51.6
BiomedCLIP (Zhang et al., 2024)	Report	23.3	<u>41.8</u>	44.6	52.9
Med-MARVEL (Zhou et al., 2024)	Report	–	–	–	55.0
FactMM-RAG (Sun et al., 2025)	Report	–	–	–	<u>60.5</u>
CXR-RePaiR-2 (Endo et al., 2021)	Sentences	25.6	–	–	–
CXR-RePaiR-Select (Endo et al., 2021)	Sentences	27.4	–	–	–
RA-RRG E9 (Park et al., 2026)	Phrases	<u>42.5</u>	–	58.3	–
RA-RRG E10 (Park et al., 2026)	Phrases	42.3	–	<u>58.8</u>	–
AnatoRetriever	Evidence statements	**50.3**	**58.7**	**63.0**	**64.9**

Table S8: Retrieval results reported on MIMIC-CXR. All scores are percentages. The four metric columns are CheXbert F1 with an uncertain observation counted as negative, marked by inventory (^{14} for all fourteen conditions and ^{5} for the five-condition subset) and by averaging (Ma for macro, Mi for micro); all are higher is better. Sentence and phrase outputs are merged before scoring. Report-baseline values follow Sun et al. (2025), Table 2, and rows retain the corresponding source protocols. Bold and underline mark the highest and second-highest reported value in each column. – denotes an unreported metric.

G Case Studies

Reading guide. We record two complete cases from the 2,461-query test set, under the same configuration as the main tables: Findings-only output, one fixed decoding setting, and the stored output of every system for that query. Each table runs from the reference through the retrieved evidence to the generated reports. In generated reports and whole-report retrieval, green marks reference-supported content and red marks contradictions, unsupported additions, attribute mismatches, and evaluator-supplied [Missing: . . .] notes. In the AnatoRetriever row, green marks evidence compatible with the admitted finding category. AnatoRAG conditions jointly on this retrieved evidence and the current/prior visual tokens.

Case selection. Table S9 follows acute cardiopulmonary findings; Table S10 follows chronic right-rib and pleural findings while preserving acute negatives.

Table S9: Case I, test index 1671, Findings only. The case follows low lung volume, pulmonary edema, left pleural effusion, cardiomegaly, and the unchanged bony thorax. RG-F_1 and $CliQ_0$ are shown for generated reports.

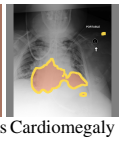
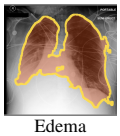
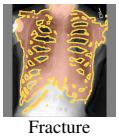
Output	Findings	Image
Reference Findings	**Lines and Tubes:** Stable right IJ line tip position. **Lungs:** Low lung volumes with mild worsening of pulmonary edema. **Pleura:** Small left pleural effusion. **Mediastinum:** Stable cardiomegaly. **Bony thorax:** No change.	
Backbone RG-F_1 21.1 $CliQ_0^{\downarrow}$ 2.975	**Lines and Tubes:** Right-sided internal jugular dialysis catheter terminates in the right atrium. Left-sided internal jugular catheter terminates in the left brachiocephalic vein. EKG leads overlie the chest wall. **Lungs:** Well inflated with interval improvement in pulmonary edema. **Pleura:** There is no pleural effusion or pneumothorax. **Mediastinum:** There is no cardiomegaly. Mediastinal silhouette is within normal limits. **Bony thorax:** No interval change.	
Report-level retrieval Top-2, Rank-1 then Rank-2	*Rank-1.* The lungs are clear besides streaky linear opacities which are likely atelectasis versus scarring. There is no focal consolidation, effusion, or edema. The cardiomediastinal silhouette is within normal limits. No acute osseous abnormalities. No free intraperitoneal air. *Rank-2.* Lung volume is low. Opacities are increased compared to prior, concerning for worsening pneumonia. There is no pulmonary edema. Cardiomediastinal silhouette is normal size. **[Missing:** small left pleural effusion**]**	 Rank-1  Rank-2
Report-level RAG RG-F_1 18.7 $CliQ_0^{\downarrow}$ 2.892	**Lines and Tubes:** Right-sided IJV dialysis catheter and left-sided IJV central venous catheter are unchanged in position. EKG leads overlie the chest wall. **Lungs:** Persistent moderate pulmonary edema. No lobar consolidation. **Pleura:** There is no pleural effusion. **Mediastinum:** No cardiomegaly. Mediastinal silhouette is within normal limits. **Bony thorax:** No interval change.	
AnatoRetriever	**Support Devices.** As compared to the previous radiograph, patient has received a new hemodialysis catheter introduced via the left internal jugular vein. **Cardiomegaly.** Continued enlargement of the cardiac silhouette with worsening congestive failure. **Edema.** Signs of pulmonary edema that might have minimally decreased in severity. **Lung Opacity.** More confluent lower lobe opacities have also worsened in the interval, and could be due to developing pneumonia in the appropriate clinical setting.	  Support Devices Cardiomegaly   Edema Lung Opacity
AnatoRAG (BLIP-3) RG-F_1 45.8 $CliQ_0^{\downarrow}$ 1.851	**Lines and Tubes:** Right IJ catheter terminates at the cavoatrial junction. Interval removal of left IJ catheter. EKG leads overlie the chest wall. **Lungs:** Persistent low lung volumes with unchanged mild pulmonary edema. **Pleura:** There is a small left pleural effusion. No pneumothorax. **Mediastinum:** Stable mild cardiomegaly and mediastinal silhouette. **Bony thorax:** No interval change.	

Table S10: Case II (failure case), test index 656, Findings only. AnatoRAG (BLIP-3) recovers the right-rib deformity but omits pleuroparenchymal scarring and attributes the costophrenic-angle appearance to pleural thickening.

Output	Findings	Image
Reference Findings	PA and lateral views of the chest were obtained. Multiple right rib deformities are noted along the right lateral rib cage. Areas of pleuroparenchymal scarring are noted in the underlying lung. Otherwise, the lungs appear clear bilaterally without focal consolidation, effusion, or pneumothorax. Cardiomediastinal silhouette is normal. Bony structures appear stable without definite signs of an acute fracture. No free air below the right hemidiaphragm is seen.	
Backbone RG-F_1 67.0 $CliQ_0^{\downarrow}$ 1.728	PA and lateral views of the chest were obtained. The lungs appear clear bilaterally without focal consolidation, effusion, or pneumothorax. No signs of CHF. Cardiomediastinal silhouette is normal. Bony structures appear intact. No free air below the right hemidiaphragm. **[Missing:** right-rib deformities and pleuroparenchymal scarring**]**	
Report-level retrieval Top-2, Rank-1 then Rank-2	*Rank-1.* The lungs are clear. Cardiomediastinal silhouette and hilar contours are unremarkable. There is no pleural effusion or pneumothorax. Clips are seen in bilateral breasts. There is no pleural effusion or pneumothorax. *Rank-2.* The lungs are clear. There is no evidence of pneumonia, pneumothorax, or pleural effusion. Cardiac silhouette is normal in size. The cardiac and mediastinal silhouettes are stable. Surgical clips are noted overlying both breasts. **[Missing:** right-rib deformities and pleuroparenchymal scarring**]**	 Rank-1  Rank-2
Report-level RAG RG-F_1 41.2 $CliQ_0^{\downarrow}$ 2.694	PA and lateral views of the chest were obtained. There is blunting of the right CP angle which likely indicates a small pleural effusion. There is no definite sign of pneumonia or pneumothorax. Old right rib cage deformity is noted. Cardiomediastinal silhouette is normal. No free air below the right hemidiaphragm. Old left upper rib cage deformity is also noted. **[Missing:** pleuroparenchymal scarring**]**	
AnatoRetriever	**Pleural Other.** There is biapical pleural thickening and thickening of the pleura along the minor fissure, similar to prior exams. **Fracture.** There are multiple old rib fractures, which appear unchanged.	 Pleural Other  Fracture
AnatoRAG (BLIP-3) RG-F_1 58.9 $CliQ_0^{\downarrow}$ 1.179	PA and lateral views of the chest were obtained. The lungs appear clear bilaterally without focal consolidation, effusion, or pneumothorax. Blunting of the right CP angle is stable from prior and likely reflects pleural thickening. Cardiomediastinal silhouette is normal. No pneumothorax or pleural effusion. Old right rib cage deformity noted. No acute bony abnormalities are seen. No free air below the right hemidiaphragm. **[Missing:** pleuroparenchymal scarring**]**	

Configuration	^{14}Ma	^{14}Mi	^{5}Ma	^{5}Mi
Full retriever	**49.8**	**62.7**	**58.5**	**64.6**
w/o query residual	<u>48.7</u>	<u>62.1</u>	<u>58.2</u>	<u>64.1</u>
w/o neighbor scores	46.3	59.4	57.1	63.4
w/o classifier	41.8	60.4	54.7	62.5
w/o compatibility	41.9	58.9	54.2	62.0
w/o coverage bonus (λ_{cov}=0)	42.2	58.5	54.0	61.9

Table S11: Cumulative retriever ablation on selected statement text before report generation. "w/o" denotes without. Ma and Mi denote macro- and micro-averaged CheXbert F_1 over the 14- and 5-label inventories. Reading downward, each row additionally removes the named component and retains all preceding removals; incremental deltas compare each row with the preceding configuration. Bold and underline mark the best and second-best value in each column.

H Retriever Component Ablation Details

Ablation protocol. Table S11 evaluates selected text before generation. Starting from the full retriever, removals accumulate from top to bottom in pipeline order. The first row restores the baseline regional query $q_\ell^{(0)}$; the next two remove the neighbor-score and image-classifier contributions to admission r_ℓ. The fourth removal bypasses the compatible set $\mathcal{V}_\ell$ while retaining candidate and patient eligibility, and the final row sets $\lambda_{\text{cov}} = 0$. Each row is read against the preceding configuration in the same replay.

Observed changes. The first three removals examine finding admission. Relative to the preceding row, removing the query residual reduces 14-label macro-F_1 by 1.1 points, removing neighbour scores reduces macro-F_1 by another 2.4 points, and removing the image-label classifier reduces macro-F_1 by another 4.5 points. Across the same sequence, micro-F_1 changes from 62.7 to 62.1, 59.4, and 60.4. The cumulative sequence therefore has the clearest effect on macro-F_1, which falls from 49.8 to 41.8 as the three admission branches are removed.

The final two removals examine statement selection. Bypassing compatibility reduces 14-label micro-F_1 from 60.4 to 58.9, and setting $\lambda_{\text{cov}} = 0$ further reduces micro-F_1 to 58.5. Compatibility filters statements against the active label set, while the coverage bonus favours statements that add active labels not yet represented. These rows complete the sequence from admission to compatible, coverage-aware selection.

Mapping to the training objectives. Eq. 3 maps to the ablation sequence as follows. The query objective $\mathcal{L}_{\text{query}}$ trains the anatomy-guided residual evaluated by w/o query residual, which restores $q_\ell^{(0)}$. The gate objective $\mathcal{L}_{\text{gate}}$ trains the semantic residual that contributes to the composed admission score r_ℓ and remains fixed while w/o neighbor scores and w/o classifier remove two complementary admission-evidence branches. After admission, w/o compatibility bypasses the compatible candidate set $\mathcal{V}_\ell$, and w/o coverage bonus sets $\lambda_{\text{cov}} = 0$ in statement selection. The five steps therefore follow the pipeline from anatomy-guided query formation, through admission-score composition, to compatibility- and coverage-aware selection.

I Retrieval by Reference Label Count

This section complements the aggregate availability analysis in Section B.2 by asking how retrieval changes as the reference label set grows. Figure S2 stratifies the paired retrieval outputs by the number of positive labels in each reference CheXbert-14 vector. Queries with no positive reference label are excluded because per-case positive-label recall is undefined.

BioMedCLIP Top-2 sequentially concatenates the Rank-1 report followed by the Rank-2 report. Adding the second report improves partial-credit metrics, with the Top-2 gain over Top-1 increasing from 2.2 to 12.0 points in micro-F_1 and from 8.7 to 14.8 points in mean per-case recall as the reference set grows. Subset accuracy exposes the accompanying precision cost: subset accuracy requires the predicted and reference positive-label sets to be identical, so either one missing label or one additional label makes the query incorrect. Top-2 is therefore slightly worse than Top-1 for single-label queries and reaches only 1.2% subset accuracy for five or more labels. AnatoRetriever improves both sides

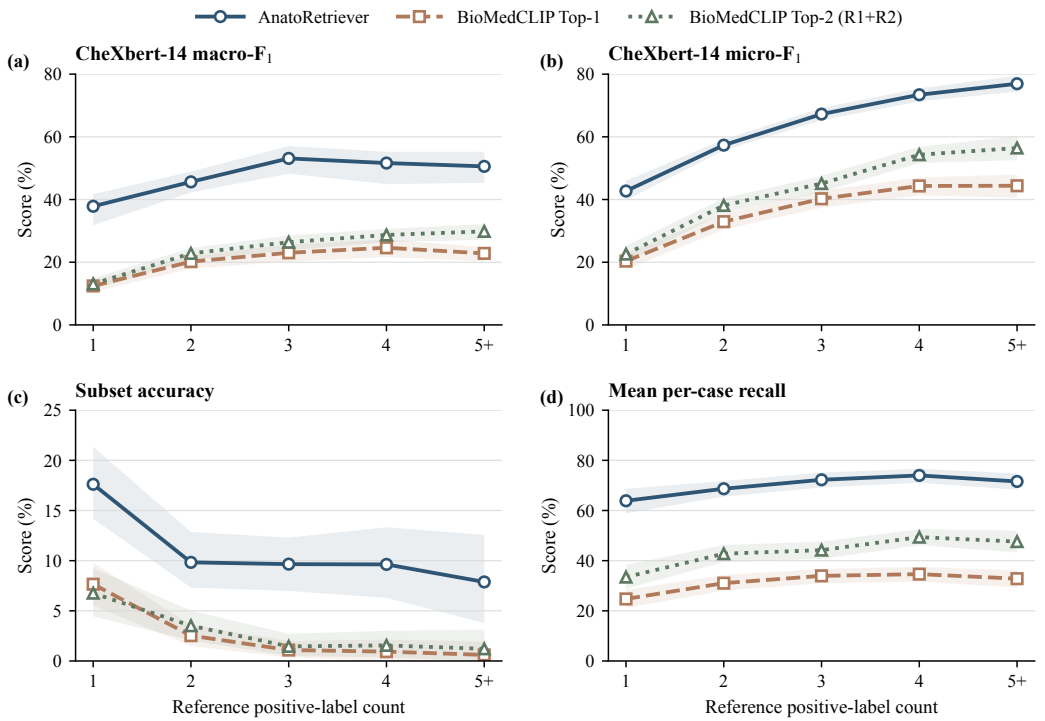


Figure S2: Retrieval performance by reference-positive CheXbert-14 label count. AnatoRetriever, BioMedCLIP Top-1, and BioMedCLIP Top-2 are evaluated on the same 2,310 positive-label queries; the five strata contain 562, 712, 549, 322, and 165 queries. Lines report scores within each stratum and shaded bands are patient-level 95% bootstrap intervals. Subset accuracy is the proportion of queries whose predicted and reference positive-label sets are identical. Mean per-case recall averages the fraction of reference-positive labels recovered for each query. BioMedCLIP Top-2 sequentially concatenates the Rank-1 report followed by the Rank-2 report (R1+R2), and the combined text is scored once.

of this trade-off and exceeds Top-2 by 19.1–22.1 points in micro-F_1 and by 24.0–30.4 points in mean per-case recall across the five strata, while retaining 7.9% subset accuracy in the five-or-more stratum.

J Additional Generation Metrics

Controlled evidence configurations. Table S12 adds lexical and factuality metrics, earlier specialist models, and RGER. Figure S3 compares the five evidence configurations across four CheXbert columns. Relative to the no-retrieval backbone, the full system improves ^{14}Ma-F_1 by 25.9%, ^{5}Ma-F_1 by 21.2%, ^{14}Mi-F_1 by 15.1%, and ^{5}Mi-F_1 by 17.0%. For our runs, RG-F1 averages exact entity-set and relation-set F1 within each report and then across reports; an empty/empty component scores zero. RGER averages recall from the same entity comparison. RadCliQ-v0 uses a separate, nonpercentage scale and is lower-is-better.

Relative-score visualization. Figure S4 presents report generation and matched retrieval in separate panels because only the retrieval panel shares one cohort and one labeler.

Published generation context. Table S12 adds lexical, graph-based, and clinical-factuality columns that are not jointly available in the main comparison.

Released Libra-3B checkpoint. Table S13 separates our evaluation of the released checkpoint from the values listed by the Libra repository.

LLaVA-TA reports one mask-conditioned finding per clinical topic and concatenates the findings into a report; the LLaVA-TA source does not report RadGraph F_1, RGER, or RadCliQ-v0.

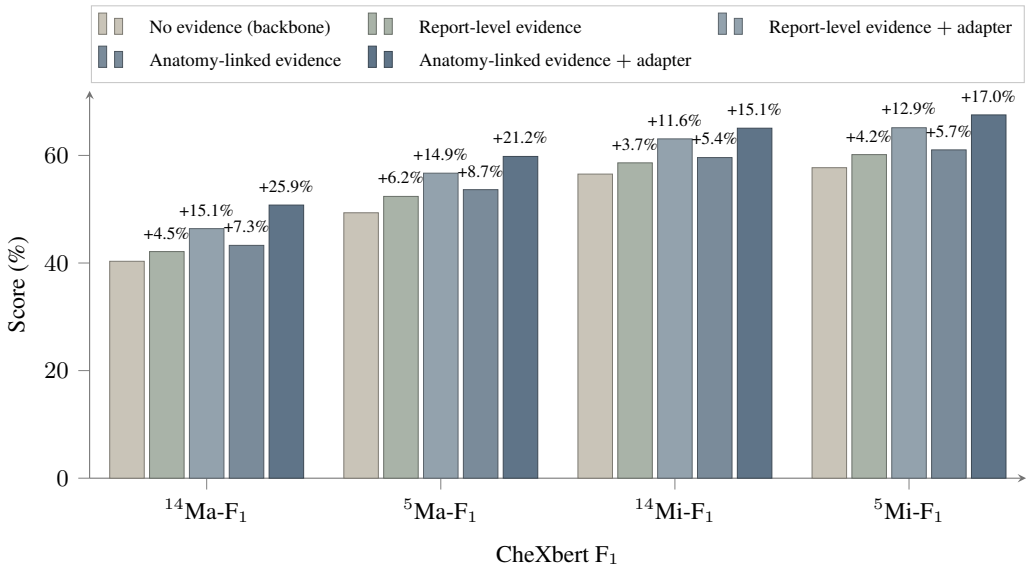


Figure S3: Every evidence configuration of Table 3 on the four CheXbert columns, drawn at the reported scores. There is one group per column and one bar per configuration, so a bar's height gives the configuration score in percent, and the number above the bar gives the percentage change from the no-evidence bar in the same group. The bare backbone carries no number because the remaining four bars use the backbone as the reference. The full system is the tallest bar in every group.

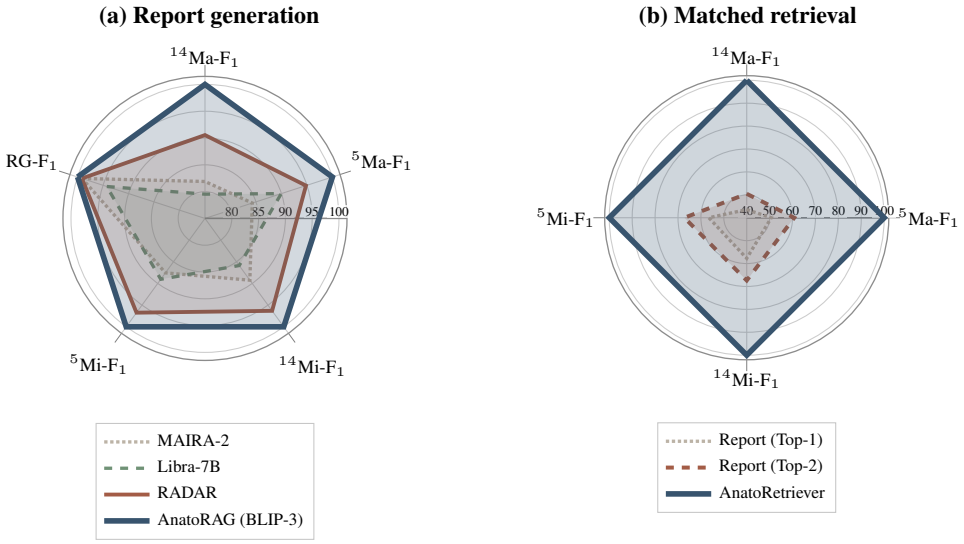


Figure S4: Relative clinical scores. Each radius is the displayed score divided by the best score on that spoke, so the outer ring is 100. **(a)** Published systems that report all five clinical columns in Table 2; this panel is descriptive across source protocols and is not a cross-protocol ranking. **(b)** Whole-report retrieval and AnatoRetriever on the same 2,461 queries under one frozen labeler; Report Top-2 sequentially concatenates Rank-1 and Rank-2 before scoring. Relative scaling supports within-spoke comparison; polygon area is not an aggregate metric.

CheXbert comparison. Table S14 reports the four CheXbert columns omitted from the preceding table, including the published scores for LLaVA-Rad (Zambrano Chaves et al., 2025) and RA-RRG (Park et al., 2026). A rule separates the published rows from AnatoRAG.

Model	B-1	B-4	R-L	RG-F1	RGER	$\text{CliQ}_0{}^{\downarrow}$
R2Gen (Chen et al., 2020)	35.3	10.3	27.0	–	–	–
R2GenCMN (Chen et al., 2021)	35.3	10.6	27.8	–	–	–
RGRG (Tanida et al., 2023)	37.3	12.6	26.4	–	–	–
PromptMRG (Jin et al., 2024)	39.8	11.2	26.8	–	–	–
R2GenGPT (Wang et al., 2023)	41.1	13.4	29.7	–	–	–
LLaVA-Rad (Zambrano Chaves et al., 2025)	38.1	15.4	30.6	–	29.4	–
MAIRA-1 (Hyland et al., 2023)	39.2	14.2	28.9	24.3	29.6	3.10
MAIRA-2 (Bannur et al., 2024)	46.5	23.4	38.4	<u>34.6</u>	<u>39.6</u>	<u>2.64</u>
Libra (Zhang et al., 2025)	<u>51.3</u>	24.5	36.7	32.9	37.6	2.70
Libra-3B† (Zhang et al., 2025)	50.5	23.3	35.2	–	37.5	–
RADAR (Hou et al., 2025)	50.9	<u>26.2</u>	39.7	<u>34.6</u>	39.3	**2.61**
RA-RRG (Park et al., 2026)	–	8.0	24.9	26.7	–	–
LLaVA-TA‡ (Cheng & Subramanian, 2026)	43.7	24.8	**42.6**	–	–	–
AnatoRAG (BLIP-3)	**54.8**	**27.7**	<u>40.1</u>	**34.9**	**40.3**	**2.61**

Table S12: Published generation results with AnatoRAG (BLIP-3) appended. BLEU-n (B-n), ROUGE-L (R-L), RadGraph F1, and RGER are percentages; RadCliQ-v0 uses the native lower-is-better scale. Bold and underline mark the best and second-best reported value in each column; tied values receive the same mark. – means unreported, † marks Libra repository values, and ‡ marks concurrent work.

Metric	Our evaluation‡	Libra repository†
BLEU-1	<u>50.4</u>	**50.5**
BLEU-4	<u>23.2</u>	**23.3**
ROUGE-L	<u>34.7</u>	**35.2**
RaTEScore	<u>57.0</u>	**61.1**
RG-ER	<u>35.4</u>	**37.5**
RadGraph F_1	30.2	–
RadCliQ-v0 ↓	2.90	–
CheXbert-14 micro-F_1	54.8	–
CheXbert-14 macro-F_1	37.5	–
CheXbert-5 micro-F_1	54.6	–
CheXbert-5 macro-F_1	44.7	–

Table S13: Evaluation of the released Libra-v1.0-3b checkpoint alongside scores reported in the Libra repository. † marks repository values and ‡ marks our checkpoint evaluation. Displayed values are rounded to one decimal and are percentages except RadCliQ-v0, for which lower is better. Where both sources report a metric, bold and underline mark the higher and lower value, respectively. RG-ER follows the repository's metric name and is distinct from RadGraph F_1. – means the repository does not report the metric.

Method	^{14}Ma	^{5}Ma	^{14}Mi	^{5}Mi	RG-F1	B-4	R-L
MAIRA-1	38.6	47.7	55.7	56.0	24.3	14.2	28.9
MAIRA-2	41.6	50.4	58.1	59.1	<u>34.6</u>	23.4	38.4
LLaVA-Rad	39.5	47.7	57.3	57.4	–	15.4	30.6
Libra-7B	40.4	53.8	55.9	60.1	32.9	24.5	36.7
RADAR	<u>46.0</u>	<u>56.7</u>	<u>62.7</u>	<u>65.3</u>	<u>34.6</u>	<u>26.2</u>	<u>39.7</u>
RA-RRG	41.7	–	58.5	–	26.7	8.0	24.9
AnatoRAG (BLIP-3)	**50.8**	**59.8**	**65.1**	**67.5**	**34.9**	**27.7**	**40.1**

Table S14: Report-generation comparison on MIMIC-CXR Findings across the four CheXbert columns and three complementary generation metrics. All scores are percentages and higher is better. Mi and Ma denote micro- and macro-averaged CheXbert F1, and the superscript gives the label inventory. Bold and underline mark the best and second-best reported value in each column; tied values receive the same mark. Published values follow Hou et al. (2025) and Park et al. (2026); – means unreported.

K Datasets and Evaluation Cohorts

Three datasets appear in this paper. Only MIMIC-CXR contributes training, development, and retrieval-bank data; the two targets are evaluation only and no target-domain report reaches the bank, the indexes, or model selection. MIMIC-CXR is released under credentialed access and the two targets under the applicable terms, and no dataset is redistributed here. Table S15 sets the three side by side.

Evaluation cohorts. The MIMIC-CXR Findings evaluation contains 2,461 frontal studies from 285 patients; 1,748 queries have multiple positive labels. The external evaluations contain 62 CheXpert Plus and 3,199 IU X-Ray cases with nonempty Findings text and an available current radiograph. BLIP-3 and AnatoRAG (BLIP-3) are evaluated on the identical cases within each external cohort.

Prior handling. A system that reads current and prior studies receives the most recent earlier examination of the same patient, ordered by examination time. A study cannot serve as the prior for the same examination, and no later study is reachable. The MIMIC-CXR release supplies 3.5 studies per patient and the CheXpert Plus release 2.9, so a prior exists for most cases in both; IU X-Ray holds one examination per patient, so IU X-Ray supplies none. The prior is an input channel only: the prior never enters the evidence bank, and a prior study from a transfer target never contributes text to that bank or to any index.

Label source. MIMIC-CXR provides CheXbert-14 assignments and CheXpert Plus provides CheXpert and CheXbert-14 assignments. IU X-Ray distributes MeSH and RadLex tags instead; the IU X-Ray reference CheXbert-14 labels are extracted from report text. Uncertain observations count as negative throughout. For retrieval and generation evaluation, labelers score each system's output text. Training, development, and test patients are disjoint, and the candidate pool excludes the query patient.

Dataset	Split usage			Dataset scale			Prior	Target	Label inventory	Eval. cases
	Train	Dev	Test	Images	Studies	Patients				
MIMIC-CXR	✓	✓	✓	377,110	227,835	65,379	✓	Findings	CheXbert-14, released with the dataset	2,461
CheXpert Plus	✗	✗	✓	223,228	187,711	64,725	✓	Findings	CheXpert and CheXbert-14, released with the dataset	62
IU X-Ray	✗	✗	✓	7,470	3,955	3,955	✗	Findings	MeSH and RadLex index tags only; CheXbert-14 computed here	3,199

Table S15: Dataset scale and evaluation cohorts for MIMIC-CXR (Johnson et al., 2019), CheXpert Plus (Chambon et al., 2024), and IU X-Ray (Demner-Fushman et al., 2016). Training, development, and evidence statement-bank records come from MIMIC-CXR; the other datasets are evaluation targets. Prior indicates whether an earlier study can be supplied to the generator. Eval. cases gives the cohort sizes used in the reported scores. Dataset-scale counts come from the cited releases.

L Metric Definitions and Scoring Procedure

The metrics proceed from label-based CheXbert to graph-based RadGraph and RadCliQ, followed by lexical BLEU and ROUGE-L. Let Y denote a generated report and $Y^\star$ denote the reference report. Figure S5 illustrates the entity and relation structure used by RadGraph.

CheXbert F1 (Smit et al., 2020). CheXbert labels each of 14 conditions in a report as positive (1), negative (0), uncertain (-1), or unmentioned. We count uncertain and unmentioned observations as negative, so a hedge counts as a negative rather than as a positive, and we report results both over all 14 conditions and over the five-condition subset consisting of Atelectasis, Cardiomegaly, Consolidation, Edema, and Pleural Effusion. For condition ℓ, let TP_ℓ, FP_ℓ, and FN_ℓ be the counts of true positives, false positives, and false negatives over the evaluation set. Then

$$P_\ell = \frac{\mathrm{TP}_\ell}{\mathrm{TP}_\ell + \mathrm{FP}_\ell}, \qquad R_\ell = \frac{\mathrm{TP}_\ell}{\mathrm{TP}_\ell + \mathrm{FN}_\ell}, \qquad F_{1,\ell} = \frac{2P_\ell R_\ell}{P_\ell + R_\ell}. \tag{11}$$

Macro-F1 averages $F_{1,\ell}$ over conditions, giving equal weight to rare findings:

$$\text{macro-}F_1 = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} F_{1,\ell}. \tag{12}$$

Micro-F1 pools true positives, false positives, and false negatives across labels and cases before taking the ratio, so frequent labels and decisions contribute more counts:

$$\text{micro-}F_1 = \frac{2 \sum_\ell \mathrm{TP}_\ell}{2 \sum_\ell \mathrm{TP}_\ell + \sum_\ell \mathrm{FP}_\ell + \sum_\ell \mathrm{FN}_\ell}. \tag{13}$$

Macro-F1 weights labels equally; micro-F1 pools condition counts. The macro/micro difference summarizes sensitivity to label frequency, while per-label scores identify which conditions change.

RadGraph entities and relations (Jain et al., 2021). Figure S5 is the extraction on one sentence. RadGraph reads a report as a directed, typed graph rather than as a bag of words, and both Y and $Y^\star$ are passed through the same released extractor, a PubMedBERT model fine-tuned on radiologist annotations, so a score compares two extracted graphs and not an extracted graph against a radiologist's. An entity is a continuous span of text and carries one of four types, an anatomy (ANAT-DP) or an observation that is definitely present, uncertain, or definitely absent (OBS-DP, OBS-U, OBS-DA). A relation is a directed, typed arrow: the *modify* relation runs from a modifier to the modified entity and only joins two entities of the same top-level type; *located_at* runs from an observation to the corresponding anatomy; and *suggestive_of* runs from one observation to an implied observation. Modifiers are independent entities, so a severity word such as *moderate* and a side word such as *left* are nodes rather than attributes of the modified finding, which is why the four-word sentence in Figure S5 yields four entities and three relations.

RadGraph F1 and RGER (Jain et al., 2021). The scorer compares extracted entity tuples and relation tuples separately. An entity becomes the pair (span text, type), and a relation becomes the

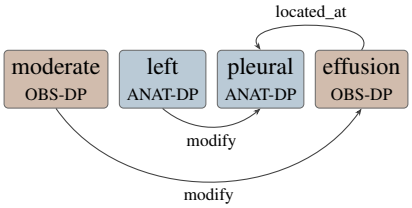


Figure S5: One sentence read as a RadGraph extraction (Jain et al., 2021). The four entity spans of *moderate left pleural effusion* sit in sentence order, each tinted by top-level entity type and carrying the fine-grained entity type beneath the span. Two relation types illustrated by this sentence appear as labelled arrows.

triple (first span and type, second span and type, relation type). Text is lowercased and duplicate tuples are removed. For either set, let $\mathcal{G}$ be the tuples extracted from Y, $\mathcal{G}^\star$ those extracted from $Y^\star$, and $M = |\mathcal{G} \cap \mathcal{G}^\star|$ the exact matches:

$$P = \frac{M}{|\mathcal{G}|}, \qquad R = \frac{M}{|\mathcal{G}^\star|}, \qquad F_1 = \frac{2PR}{P+R}. \tag{14}$$

A report in which both sides of the comparison are empty scores zero rather than one, so a blank output is not rewarded for agreeing with a blank reference. The entity F1 and the relation F1 are averaged within a report and the average is then taken across reports, which is what the RG-F1 column reports; averaging per report rather than pooling keeps a long report from carrying more weight than a short one. RGER is the recall of the entity comparison, $R = M/|\mathcal{G}^\star|$, averaged over reports in the same way. We report both because the two columns answer different questions: a candidate that omits content keeps precision at the maximum and loses only on recall, while one that adds unsupported findings keeps recall at the maximum and loses only on precision, and F1 moves under either error while recall alone separates omission from addition.

RadCliQ-v0 (Yu et al., 2023). RadCliQ-v0 combines four automated metrics: RadGraph F1; BERTScore (Zhang et al., 2020), computed with a DistilRoBERTa model and rescaled against the baseline; s_{emb}, the cosine similarity between the CheXbert embeddings of Y and $Y^\star$; and BLEU-2. The released min–max normalizer is applied before a linear regression fitted to radiologist rankings. We report the resulting v0 score on the native lower-is-better scale.

BLEU-n (Papineni et al., 2002). Let c and r be the candidate and reference lengths, and let p_i be clipped i-gram precision. For BLEU-n, uniform weights are $w_i = 1/n$ for $i = 1, \ldots, n$, with brevity penalty

$$\text{BP} = \begin{cases} 1 & c > r, \\ \exp(1 - r/c) & c \le r, \end{cases} \tag{15}$$

BLEU is

$$\text{BLEU-}n = \text{BP} \cdot \exp\left(\sum_{i=1}^{n} w_i \log p_i \right). \tag{16}$$

Of the four orders we report BLEU-1 and BLEU-4: BLEU-1 is close to unigram precision, BLEU-4 is the standard corpus-level figure, and the tables quote both.

ROUGE-L (Lin, 2004). Let L be the length of the longest common subsequence between Y and $Y^\star$. With $P_{\text{lcs}} = L/|Y|$, $R_{\text{lcs}} = L/|Y^\star|$, and $\beta = 1.2$,

$$F_{\text{lcs}} = \frac{(1+\beta^2) P_{\text{lcs}} R_{\text{lcs}}}{R_{\text{lcs}} + \beta^2 P_{\text{lcs}}}. \tag{17}$$

Reporting conventions. All CheXbert, RadGraph, RGER, BLEU, and ROUGE-L figures are reported as percentages rounded to one decimal. RadCliQ-v0 is reported to two decimals on the native scale. Every labeler scores the text a system actually emits. Retrieved text and generated text are scored in separate passes, so a score on retrieved text measures the retriever and a score on generated text measures the full system.