# A RELIABILITY-AWARE EVIDENCE ORGANIZATION FRAMEWORK FOR RETRIEVAL-AUGMENTED MEDICAL VISUAL QUESTION ANSWERING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Medical large vision-language models (Med-LVLMs) have advanced medical visual question answering (VQA), but they can still produce answers unsupported by the image or clinical evidence. Retrieval-augmented generation (RAG) improves factual grounding by introducing radiology reports, yet visually similar examinations do not necessarily yield question-relevant or clinically consistent evidence. Report-level retrieval therefore suffers from two fundamental mismatches: retrieved reports may fail to cover the queried finding, and the retrieved content may contain irrelevant, uncertain, or conflicting statements. We propose REO-RAG, a novel factual grounding framework for medical VQA, which converts retrieved reports into question-specific clinical evidence. Without altering the image retriever, REO-RAG decomposes reports into structured evidence units encoding clinical findings, anatomical locations, and assertion status. It then uses a question-conditioned reliability estimator to jointly assess target relevance, uncertainty, and cross-evidence inconsistency. A risk-guided organizer subsequently selects and composes lower-risk, question-aligned evidence into a compact, clinically coherent context for the Med-LVLM. By separating evidence acquisition from evidence utilization, REO-RAG enables the Med-LVLM to reason over organized question-specific evidence rather than unfiltered reports. Its key contribution is not stronger retrieval, but a representation-level bridge that transforms retrieved reports into more reliable evidence for clinical reasoning. Under a unified evaluation protocol, REO-RAG reaches F1 scores of 88.60%, 88.89%, and 56.05% on IU-Xray, MIMIC-CXR, and PadChest-GR, with larger gains when retrieved reports already contain mixed but question-relevant evidence.

## 1 INTRODUCTION

Medical visual question answering (Med-VQA) aims to answer natural-language questions grounded in medical images (Lau et al., 2018; Liu et al., 2021; Li et al., 2023). With the emergence of medical large vision-language models (Med-LVLMs), the task has moved from fixed-vocabulary classification to open-ended clinical response generation. Despite improved multimodal reasoning, these models may still produce answers that are weakly grounded in the image or inconsistent with clinical evidence. Such errors are particularly consequential in chest radiography, where the presence, absence, uncertainty, and anatomical location of a finding jointly determine the correct answer.

Fine-tuning offers a direct way to improve factual reliability, but is often limited by annotation cost and distribution shift. Retrieval-augmented generation (RAG) (Lewis et al., 2020) instead supplies radiology reports from visually similar examinations at inference time, with recent work on context filtering, domain-aware retrieval, preference alignment, and agentic search (Xia et al., 2024; 2025; Zhang et al., 2026). These methods still largely operate at the retrieval-policy or report level: once reports are retrieved, the generator must disentangle which internal statements are relevant to the question.

This design yields two distinct failure modes. A target-coverage mismatch occurs when retrieved reports omit the queried finding, while an evidence-utilization mismatch occurs when target evidence is mixed with unrelated, uncertain, generic, or opposite-polarity statements. Because retrieved re-



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.

ports usually describe different cases, opposite polarities reflect heterogeneous references rather than contradictions about the query patient. Thus, the bottleneck is not only what is retrieved, but also how retrieved statements are represented and presented.

We propose REO-RAG, which maps retrieved reports to structured clinical units and organizes them by question-target match together with auxiliary textual risk cues—without modifying the image retriever. The resulting score estimates question-conditioned textual evidence utility rather than image-level clinical correctness. On IU-Xray, MIMIC-CXR, and a constructed PadChest-GR yes/no subset, organization helps most when the Top-$M$ pool covers the queried finding but mixes distractors or heterogeneous polarity support; under sparse coverage, it cannot invent missing findings and Top-1 reports can remain competitive. Matched controls separate preference-adaptation lift from organization gains.

**Contributions.** (i) We formulate post-retrieval evidence utilization as distinct from evidence acquisition in binary chest X-ray VQA, separating missing target coverage from poor use of available evidence. (ii) We introduce structured clinical evidence units with explicit finding, anatomy, assertion, polarity, and provenance, together with a question-conditioned organizer that prioritizes target-aligned evidence and exposes heterogeneous polarity support. (iii) Through matched generator controls, representation ablations, and retrieval diagnostics, we characterize when evidence organization helps and when it is limited by missing retrieval coverage.

## 2 Related Work

### 2.1 Medical Vision-Language Models

Early Med-VQA methods typically encoded an image–question pair and selected an answer from a fixed vocabulary. VQA-RAD introduced clinician-authored questions grounded in radiology images (Lau et al., 2018), while SLAKE extended the task with bilingual questions, semantic annotations, and a medical knowledge base (Liu et al., 2021). More recently, Med-LVLMs have enabled open-ended clinical response generation. LLaVA-Med adapts a general-purpose vision-language assistant (Liu et al., 2023) to biomedical images through image–text alignment and medical instruction tuning (Li et al., 2023). These models remain vulnerable to factual errors when evidence is insufficient or noisy. We use LLaVA-Med-1.5 as the generator and improve the reliability of its external evidence context rather than modifying the backbone.

### 2.2 Multimodal Retrieval-Augmented Generation

RAG supplements model parameters with external corpus evidence (Lewis et al., 2020; Gao et al., 2024), with selective variants such as Self-RAG (Asai et al., 2024) deciding when and what to retrieve. In multimodal medical settings, contrastive encoders such as CLIP (Radford et al., 2021), MedCLIP (Wang et al., 2022), and BioViL (Bannur et al., 2023) retrieve reports from visually similar examinations. Recent medical methods improve this paradigm through calibrated context selection and preference tuning (RULE (Xia et al., 2024)), domain-aware retrieval (MMed-RAG (Xia et al., 2025)), agentic search (Zhang et al., 2026), and fine-grained knowledge reweighting to reduce interference from irrelevant retrieved medical knowledge (Liang et al., 2025). These works show that post-retrieval utilization is not itself new. REO-RAG is positioned more narrowly around structured radiology evidence units with explicit clinical attributes and question-conditioned organization of those units, rather than around fine-grained utilization in general.

### 2.3 Clinical Evidence Extraction and Organization

Clinical information extraction tools such as CheXpert (Irvin et al., 2019) and RadGraph (Jain et al., 2021) structure radiology reports, but are not designed to organize retrieved evidence for a VQA question. General evidence-aware RAG methods such as SURE-RAG and ECR verify or resolve retrieved textual evidence, but do not provide the same finding/anatomy/assertion/polarity representation for radiology statements. REO-RAG instead organizes structured report-derived evidence for binary chest X-ray VQA.

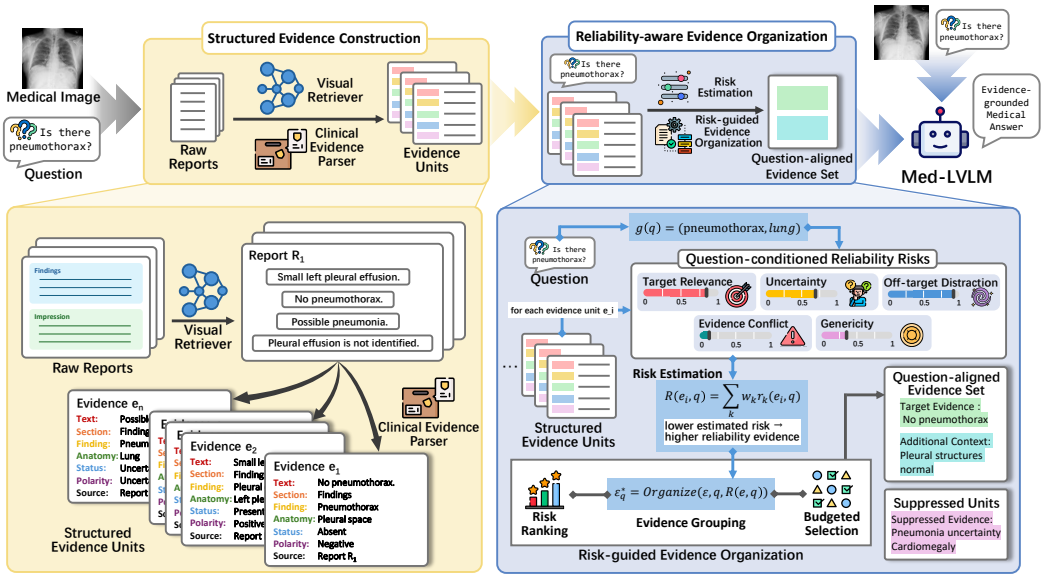


Figure 1: Overview of REO-RAG. The framework performs an evidence representation transformation: Stage I maps retrieved reports to structured clinical evidence units; Stage II organizes these units by question-target match and auxiliary textual risk cues into a compact context $\mathcal{E}_q^*$ for Med-LVLM generation (evidence utilization, not retriever redesign).

# 3 METHOD

## 3.1 OVERVIEW

REO-RAG is an interpretable evidence organization module: it exposes and reorganizes clinical statements after retrieval rather than learning a new retriever. Given a medical image $I$ and question $q$, it transforms retrieved radiology reports into a question-aligned evidence context before answer generation:

$$(I, q) \xrightarrow{\text{retrieve}} \mathcal{R}_M \xrightarrow{\text{parse}} \mathcal{E} \xrightarrow{\text{organize}} \mathcal{E}_q^* \xrightarrow{G_\theta} a. \tag{1}$$

Stage I maps retrieved reports to sentence-level clinical units with explicit finding, anatomy, assertion, and polarity. Stage II scores these units, admits lower-risk ones, and composes a compact context $\mathcal{E}_q^*$ for generation,

$$a = G_\theta\big(I,\ q,\ \text{Serialize}(\mathcal{E}_q^*)\big)\,. \tag{2}$$

Admission alone would resemble filtering by risk; Stage II additionally resolves polarity at the group level—keeping a dominant side when support is clear, or retaining both sides with an explicit `[Conflict]` cue when Yes/No evidence is near-tied.

## 3.2 STRUCTURED EVIDENCE CONSTRUCTION

### 3.2.1 VISUAL REPORT RETRIEVAL.

Let $\mathcal{D} = \{(I_j, R_j)\}_{j=1}^N$ be training-split image–report pairs, so a query never retrieves its own paired report. An image encoder $f_{\text{img}}$ ranks examinations by cosine similarity $s(I, I_j)$ and returns

$$\mathcal{R}_M = \big\{R_{\pi(j)}\big\}_{j=1}^M, \quad \pi = \text{argsort}_{j\in[N]}\, s(I, I_j) \text{ (descending)}, \tag{3}$$

with fixed coarse depth $M$. This retriever is standard: REO-RAG restructures how retrieved content is consumed, and does not update $f_{\text{img}}$ or the corpus index.

#### 3.2.2 STRUCTURED EVIDENCE REPRESENTATION.

Each clinical sentence is an evidence unit with finding, anatomy, assertion, and polarity—attributes left implicit by report concatenation. For sentence $x$ from section $c$ of report $R \in \mathcal{R}_M$, an interpretable mapper $\Phi$ yields

$$e = \Phi(x, c, R) = \big(x,\ c,\ f,\ \ell,\ z,\ p,\ R\big), \tag{4}$$

where $f$ is a finding key, $\ell$ an anatomy, $z \in \{\text{PRESENT}, \text{ABSENT}, \text{UNCERTAIN}, \text{NORMAL}\}$ the assertion status, and $p \in \{\text{POS}, \text{NEG}, \text{UNC}\}$ the polarity. As a short illustration, "small left pleural effusion" $\mapsto (f, \ell, z, p) = (\text{effusion}, \text{left}, \text{PRESENT}, \text{POS})$, while "effusion is not identified" $\mapsto (\text{effusion}, \varnothing, \text{ABSENT}, \text{NEG})$. The candidate pool is

$$\mathcal{E} = \bigcup_{R \in \mathcal{R}_M} \Phi(R) = \{e_i\}_{i=1}^{|\mathcal{E}|}. \tag{5}$$

We use a controlled clinical vocabulary and assertion rules for interpretability; the full specification of $\Phi$ is in Appendix A.1.

From the question, we extract a clinical target $g(q) = (f_q, \ell_q)$, where $f_q$ is the queried finding and $\ell_q$ an optional anatomy constraint. Target matching is

$$m_i = \mathbb{I}[f_i \sim_f f_q] \cdot \mathbb{I}[\ell_q = \varnothing \ \vee\ \ell_i \sim_\ell \ell_q]\,, \tag{6}$$

where $\sim_f$ and $\sim_\ell$ denote fixed finding-alias and anatomical compatibility relations, respectively (Appendix A.1). Units with $m_i = 1$ are *target evidence*; the remainder provide *off-target* context.

### 3.3 RELIABILITY-AWARE EVIDENCE ORGANIZATION

The structured pool may mix target, unrelated, uncertain, and opposite-polarity statements from different retrieved examinations. Stage II assigns roles, organizes polarity support, and serializes retained units:

$$\mathcal{E}_q^* = \text{Organize}(\mathcal{E}, q, R(e, q)), \tag{7}$$

where $\text{Organize}(\cdot)$ denotes the full pipeline of risk scoring, role-aware admission, polarity grouping, budgeted selection, and tagged serialization—rather than a Top-$k$ truncation or a single soft filter on similarity.

#### 3.3.1 QUESTION-AWARE TEXTUAL EVIDENCE SIGNALS.

For each $e_i \in \mathcal{E}$ we compute bounded textual risk factors $\mathbf{r}(e_i, q) \in [0, 1]^{|\mathcal{K}|}$ (lower indicates higher estimated question-conditioned utility), indexed by

$$\mathcal{K} = \{\text{tr},\ \text{od},\ \text{u},\ \text{ec},\ \text{ge}\}, \tag{8}$$

aligned with Fig. 1: *target relevance risk*, *off-target distraction risk*, *uncertainty risk*, *cross-reference polarity-disagreement risk*, and *genericity risk*. Let $\mathcal{E}_q = \{e_j \in \mathcal{E} : m_j = 1\}$ and write $\mathbb{I}[\cdot]$ for the Iverson bracket. These signals estimate the utility of retrieved text for the question; they do *not* verify whether a statement is clinically correct for the query image. Evidence polarity is used to expose heterogeneous Yes/No support across retrieved references; it is *not* used to penalize units whose polarity opposes the question's surface form (presence questions still need target-negative evidence for No answers).

**Target relevance risk and off-target distraction risk.** Target relevance risk penalizes missing coverage of the queried finding, while off-target distraction specifically penalizes irrelevant positive statements that may induce false affirmatives:

$$\begin{aligned} r_{\text{tr}}(e_i, q) &= 1 - m_i, \\ r_{\text{od}}(e_i, q) &= (1 - m_i) \cdot \mathbb{I}[z_i = \text{PRESENT}]\,. \end{aligned} \tag{9}$$

Target-matched units receive zero coverage penalty; off-target positives remain high-risk even when retrieval similarity is high.

**Algorithm 1** Reliability-aware evidence organization (REO-RAG)

**Require:** Query $(I, q)$; corpus $\mathcal{D}$; encoder $f_{\text{img}}$; evidence mapper $\Phi$; weights $\mathbf{w}$; threshold $\tau$; budget $K_{\max}$; dominance margin $\delta$
**Ensure:** Answer $a$
$\mathcal{R}_M \leftarrow$ top-$M$ reports by $s(I, I_j)$ ▷ Eq. (3)
$\mathcal{E} \leftarrow \bigcup_{R \in \mathcal{R}_M} \Phi(R)$ ▷ Eq. (5)
Extract target $g(q) = (f_q, \ell_q)$
**for** each $e_i \in \mathcal{E}$ **do**
    Compute match $m_i$ and factors $\mathbf{r}(e_i, q)$ ▷ Eqs. (6)–(12)
    $R(e_i, q) \leftarrow \langle \mathbf{w}, \mathbf{r}(e_i, q) \rangle$; set $R_{\text{adm}}$ by dropping $r_{\text{ec}}$ when $m_i$=1
**end for**
$\mathcal{A}_q \leftarrow \{e_i \in \mathcal{E} : R_{\text{adm}}(e_i, q) \le \tau\}$
**if** $\mathcal{A}_q = \varnothing$ **then**
    $\mathcal{E}_q^* \leftarrow \{\arg\min_{e_i} R(e_i, q)\}$ ▷ fallback
**else**
    $\mathcal{E}_q^* \leftarrow$ OrganizePolarity$(\mathcal{A}_q; K_{\max}, \delta)$ ▷ dominant side or conflict cue
**end if**
$a \leftarrow G_\theta\big(I, q, \text{Serialize}(\mathcal{E}_q^*)\big)$
**return** $a$

**Uncertainty.** Hedged or weakly asserted statements are elevated by

$$r_{\text{u}}(e_i, q) = \max\Big(\mathbb{I}[z_i = \text{UNCERTAIN}]\,,\; u(x_i)\Big), \tag{10}$$

where $u(x_i) \in [0, 1]$ scores hedge cues in the sentence (e.g., "possible", "may", "cannot exclude").

**Cross-reference polarity disagreement and polarity organization.** Let $\text{pol}(e_i) \in \{+1, -1, 0\}$ be the signed polarity of unit $e_i$. Because retrieved reports usually describe different examinations or patients, opposite polarities are not logical contradictions about the same case. We therefore use $r_{\text{ec}}$ only as a marker of *cross-reference polarity disagreement*, recorded by

$$r_{\text{ec}}(e_i, q) = m_i \cdot \mathbb{I}\Big[\exists\, e_j \in \mathcal{E}_q : \text{pol}(e_i) \cdot \text{pol}(e_j) < 0\Big]. \tag{11}$$

Importantly, $r_{\text{ec}}$ is *not* used as a hard discard signal at admission: target units ignore $r_{\text{ec}}$ so that both Yes and No reference support can enter the candidate set. Organization then operates at the *group* level. Target units are partitioned into $G_q^+ \cup G_q^- \cup G_q^u$ with support masses $S_+ = \sum_{e_i \in G_q^+} \alpha_i$ and $S_- = \sum_{e_i \in G_q^-} \alpha_i$, where $\alpha_i$ is a deterministic question–evidence relevance prior from normalized lexical overlap (Appendix A.1). If the margin $|S_+ - S_-|/(S_+ + S_- + \epsilon)$ exceeds a dominance threshold $\delta$, only the dominant polarity is retained; otherwise both sides are kept and serialization prepends an explicit `[Conflict]` cue. The tag denotes disagreement among retrieved references, not a contradiction within the query patient's record. The purpose of this grouping is to avoid presenting one retrieved polarity without exposing competing reference evidence.

**Genericity risk.** Vague global statements (e.g., "lungs are clear") are scored by

$$r_{\text{ge}}(e_i, q) = g_{\text{glob}}(x_i) \cdot (1 - \eta\, m_i), \tag{12}$$

where $g_{\text{glob}}(x_i) \in [0, 1]$ flags such statements and $\eta \in (0, 1]$ attenuates the penalty for otherwise target-matched units.

### 3.3.2 Risk-guided evidence organization.

The scalar textual evidence risk is the weighted aggregate

$$R(e_i, q) = \langle \mathbf{w},\, \mathbf{r}(e_i, q) \rangle = \sum_{k \in \mathcal{K}} w_k\, r_k(e_i, q), \tag{13}$$

with $\mathbf{w} \in \Delta^{|\mathcal{K}|-1}$ frozen across datasets. Lower $R(e_i, q)$ indicates lower estimated question-conditioned textual risk; it is a ranking signal rather than a calibrated confidence or an image-level correctness estimate. Weights are clinical priors; values and $\tau$ are reported in Experiments.

Construction of $\mathcal{E}_q^*$ proceeds in three steps:

1. **Admission.** Target units use $R_{\mathrm{adm}}$ (full $R$ with $r_{\mathrm{ec}}$ dropped) and off-target units use $R$, yielding $\mathcal{A}_q = \{e_i \in \mathcal{E} : R_{\mathrm{adm}}(e_i, q) \leq \tau\}$. If $\mathcal{A}_q = \varnothing$, a fallback keeps the single lowest-risk unit.

2. **Polarity organization.** OrganizePolarity applies the group-level dominant-side or conflict-cue rule under budget $K_{\max}$.

3. **Serialization.** $\mathrm{Serialize}(\mathcal{E}_q^*)$ emits tagged strings and, when both polarities remain, a leading `[Conflict]` cue.

Algorithm 1 summarizes inference.

Unless otherwise stated, $G_\theta$ is LLaVA-Med-1.5 with an IU-Xray answer-consistency DPO (AC-DPO) LoRA adapter (Hu et al., 2022; Rafailov et al., 2023). AC-DPO is preference tuning on IU-Xray yes/no pairs ($N$=1,933): chosen answers are short GT-aligned Yes/No templates, and rejected answers are the model's raw responses when they differ in format or polarity (Appendix B.1). We train once on IU-Xray and transfer the same adapter to MIMIC-CXR and PadChest-GR without target-domain retraining. It changes how the Med-LVLM consumes a given evidence string, but leaves CLIP retrieval and Stages I–II unchanged. Matched controls therefore cross the same contexts with base LLaVA-Med vs. AC-DPO.

# 4 Experiments

## 4.1 Experimental Setup

**Datasets and evaluation protocol.** We evaluate binary yes/no VQA on IU-Xray (Demner-Fushman et al., 2016), MIMIC-CXR (Johnson et al., 2019), and a constructed PadChest-GR (de Castro et al., 2025) *classic6* yes/no subset (derived from PadChest (Bustos et al., 2020); not a native Med-VQA benchmark; Appendix B.1). IU-Xray and MIMIC-CXR are the primary in-domain settings; PadChest-GR is a *target-domain calibrated stress test* of transfer (IU AC-DPO not retrained; $\tau$ selected on a PadChest-GR calibration split), rather than evidence for zero-shot or general Med-VQA transfer. The main evaluation uses 1,870 IU-Xray, 3,166 MIMIC-CXR, and 416 PadChest-GR questions. For each dataset, the retrieval corpus is built only from its training split, with a question-id-disjoint calibration set of 300 samples. PadChest-GR is patient-disjoint; IU-Xray and MIMIC-CXR follow the published question-level protocol and are not guaranteed image-/study-disjoint across calibration and holdout, a residual leakage risk discussed in Appendix B.1.

**Implementation details.** All methods use OpenCLIP RN50-quickgelu (Cherti et al., 2023; Radford et al., 2021) pretrained on CC12M for retrieval and LLaVA-Med-1.5 (Mistral-7B) for answer generation. REO-RAG retrieves $M$=8 reports and retains at most $K_{\max}$=10 evidence units. The threshold is calibrated on each dataset's own labeled split ($\tau$=0.15 / 0.35 / 0.55), while risk weights are fixed across datasets. Generation uses greedy decoding. Unless a baseline's own preference adapter is part of the reproduced system, REO-RAG uses the transferred IU AC-DPO adapter. Hyperparameters and prompts are in Appendices B.1 and A.2; efficiency is analyzed in Appendix C.3.

**Metrics.** We report accuracy, precision, recall, and positive-class F1 (primary); for PadChest-GR we also report finding-aware Macro-F1 (Appendix B.1). Retrieval diagnostics appear in Appendix C.1. A yes/no parser maps free-form generations to binary labels under a shared protocol for all methods.

**Baselines.** All methods share the same index, backbone, holdouts, decoding, and parser. We compare Report-level RAG (Top-1 full report), MedDr (He et al., 2026), FactMM-RAG (Sun et al., 2025), MR-RAG (Li et al., 2026) (MCR with Top-$K$=4 summaries and base LLaVA-Med), RULE (Xia et al., 2024), and MMed-RAG (Xia et al., 2025) (Appendix B.2). Table 1 is a system-level reference (adapters may differ); Tables 2–4 isolate organization.

Table 1: Overall comparison under the shared CLIP index, holdouts, and parser (%). F1 is primary. Adapters may differ (RULE / MMed / AC-DPO); see Table 2. $N$=1,870 / 3,166 / 416. Best in each column in bold.

| Method | IU-Xray | | | | MIMIC-CXR | | | | PadChest-GR | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acc | Pre | Rec | **F1** | Acc | Pre | Rec | **F1** | Acc | Pre | Rec | **F1** |
| Report-level RAG | 85.99 | 77.80 | 65.86 | 71.33 | 64.06 | 77.61 | 58.61 | 66.78 | 48.56 | 50.00 | 17.29 | 25.69 |
| MedDr | 85.78 | 80.05 | 61.62 | 69.63 | 71.57 | 81.16 | 70.18 | 75.27 | 51.20 | 55.91 | 24.30 | 33.88 |
| RULE | 89.09 | 74.79 | 88.69 | 81.15 | 81.27 | 85.30 | 84.12 | 84.70 | 49.52 | 50.83 | 57.01 | 53.74 |
| FactMM-RAG | 87.22 | 82.16 | 66.06 | 73.24 | 65.13 | 78.19 | 60.25 | 68.06 | 47.84 | 48.82 | 28.97 | 36.36 |
| MMed-RAG | 91.44 | 83.57 | 84.24 | 83.90 | 78.36 | 87.64 | 75.56 | 81.16 | 51.44 | **56.12** | 25.70 | 35.26 |
| MR-RAG | 86.95 | 79.95 | 67.68 | 73.30 | 75.05 | 78.12 | 82.68 | 80.34 | **52.16** | 55.10 | 37.85 | 44.88 |
| Ours | **93.90** | **87.72** | **89.49** | **88.60** | **86.67** | **91.44** | **86.48** | **88.89** | 50.24 | 51.36 | **61.68** | **56.05** |

Table 2: Controlled comparison (F1, %): same evidence JSONLs and $\tau$; only the generator changes (base LLaVA-Med vs. IU AC-DPO). Best F1 per dataset in **bold**.

| Evidence | Base LLaVA-Med | | | AC-DPO | | |
|---|---|---|---|---|---|---|
| | IU | MIMIC | PadChest | IU | MIMIC | PadChest |
| Image-only | 62.55 | 77.72 | 0.00 | 85.15 | 87.97 | 43.50 |
| Report-level | 72.93 | 67.16 | 25.61 | 86.11 | 85.41 | **59.39** |
| REO-RAG | 71.69 | 78.09 | 47.59 | **88.60** | **88.89** | 56.05 |

### 4.2 Main Results

**Overall comparison.** Table 1 reports systems under a shared CLIP index/parser. REO-RAG leads IU/MIMIC F1 (88.60% / 88.89%) and the strongest PadChest-GR baseline (56.05% vs. 53.74%; Macro-F1 22.55→36.18 vs. RULE, Appendix B.1). Gains over Report-level RAG / MedDr / FactMM-RAG are large; vs. RULE / MMed-RAG, IU/MIMIC stay competitive, while PadChest-GR is limited by sparse coverage (Appendix C.1). This table is system-level; Table 2 attributes organization vs. adaptation. Finding-wise F1 under the same protocol is reported in Appendix C.1.

**Controlled comparison of evidence and generator.** Table 2 crosses evidence context with base LLaVA-Med and the frozen IU AC-DPO adapter. AC-DPO substantially improves F1 under all evidence settings, showing that preference adaptation contributes much of the system-level gain. Without AC-DPO, REO-RAG improves over Top-1 reports on MIMIC-CXR and PadChest-GR but is roughly tied on IU-Xray. With AC-DPO, REO-RAG leads on IU/MIMIC, whereas Top-1 reports remain stronger on PadChest-GR (59.39 vs. 56.05), consistent with sparse target coverage.

### 4.3 Analysis of Evidence Organization

**Evidence transformation.** Table 3 shows that organization preserves target coverage while substantially increasing the fraction of target evidence and reducing off-target positives, especially on MIMIC-CXR. Because near-tied polarities may both be retained, disagreement among retained units can increase. Figure 2 visualizes this transformation.

**Retrieval and oracle diagnostics.** Frozen retrieval diagnostics (Appendix C.1) show that CLIP similarity is a weak sample-level trust signal: on PadChest-GR, TP/FN mean similarities differ by only 0.003 (0.421 vs. 0.418), and target-finding Evidence Recall@$K$ remains low even at $K$=20 (27.9%). Oracle-context probes (Appendix C.1) are diagnostic rather than competitive baselines: gold paired reports raise IU/MIMIC F1 to 90.75% / 95.30%, while on PadChest-GR the gold report reaches only 58.42% (+2.37% over REO-RAG), so the cross-dataset gap is not explained by over-filtering alone.

Table 3: Evidence transformation before and after organization. #Evid. is the mean number of deduplicated structured units; Cov. denotes target coverage.

| Dataset | Stage | #Evid. | Cov. (%) | Target (%) | Unc. (%) | Conf. (%) | Off+ (%) |
|---|---|---|---|---|---|---|---|
| IU-Xray | Before | 5.7 | 77.1 | 36.0 | 0.1 | 0.0 | 36.7 |
| | After | 2.6 | 77.1 | 59.7 | 0.0 | 0.0 | 21.2 |
| MIMIC-CXR | Before | 62.5 | 96.2 | 32.3 | 6.2 | 17.9 | 55.8 |
| | After | 9.7 | 96.2 | 70.8 | 4.3 | 34.8 | 0.0 |
| PadChest-GR | Before | 20.9 | 47.1 | 5.1 | 6.9 | 1.8 | 72.8 |
| | After | 9.7 | 47.1 | 9.5 | 13.3 | 2.4 | 43.1 |

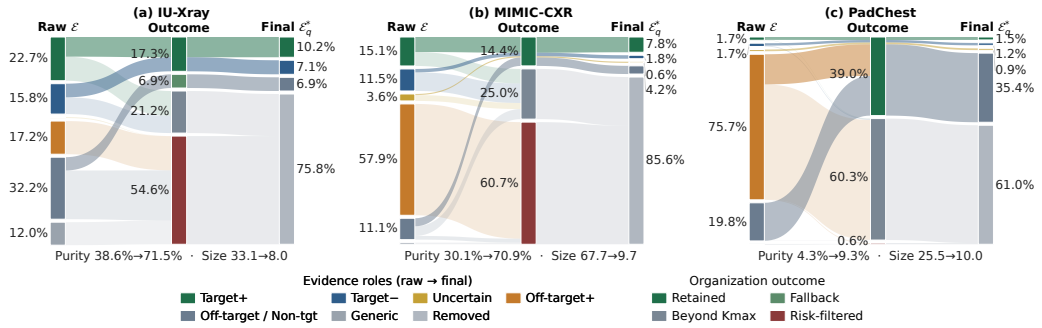


Figure 2: Evidence organization on IU-Xray, MIMIC-CXR, and PadChest-GR. Raw evidence roles flow through organization outcomes into $\mathcal{E}_q^*$. Footer statistics use per-report volumes; Table 3 reports deduplicated units.

Table 4: Evidence representation ablation under frozen IU AC-DPO and main-table $\tau$ (F1, %). Best per column in **bold**.

| Evidence representation | IU-Xray | MIMIC-CXR | PadChest-GR |
|---|---|---|---|
| Report-level context | 86.11 | 85.41 | **59.39** |
| Sentence-level evidence | 87.07 | 87.94 | 52.93 |
| Structured evidence units | 87.15 | 87.90 | 53.25 |
| REO-RAG | **88.60** | **88.89** | 56.05 |

### 4.4 Ablation Studies

**Evidence representation and organization.** Table 4 freezes IU AC-DPO and main-table $\tau$, varying only how retrieved content is represented. Moving from report-level to sentence or structured units helps only modestly; the full REO-RAG pipeline adds question-conditioned organization and yields the highest IU/MIMIC F1 in this matched-adapter comparison (88.60 / 88.89). On PadChest-GR, report-level still leads (59.39 vs. 56.05)—a boundary under sparse coverage rather than a failure of the representation cascade (Appendix C.2).

**Reliability components.** Table 5 removes selected reliability components while keeping the rest of the pipeline fixed (AC-DPO; main-table $\tau$). Zeroing target relevance yields the clearest drop ($-1.30\%$ / $-0.44\%$), whereas uncertainty, polarity-disagreement, and generic-prior removals are near-flat. These results do not support interpreting every risk factor as independently necessary. Instead, target relevance is the dominant signal under the current binary finding-oriented setting, while the remaining factors provide an interpretable organization scaffold whose marginal effects are small in aggregate F1. We therefore describe the multi-factor score as a practical heuristic organizer, not as a validated decomposition of clinical reliability. A calibration-set leave-one-out analysis, including off-target distraction, is provided in Appendix C.2.

Table 5: Reliability-component ablations (F1, %). Each row zeros one factor group; $\Delta$ vs. full model. Leave-one-out, including off-target distraction, is in Appendix C.2.

| | IU-Xray | | MIMIC-CXR | |
|---|---|---|---|---|
| Setting | F1 (%) | $\Delta$ (%) | F1 (%) | $\Delta$ (%) |
| Full REO-RAG | 88.60 | – | 88.89 | – |
| w/o target relevance | 87.30 | −1.30 | 88.45 | −0.44 |
| w/o uncertainty | 88.60 | 0.00 | 88.77 | −0.12 |
| w/o evidence conflict | 88.60 | 0.00 | 88.45 | −0.44 |
| w/o generic prior | 88.71 | +0.11 | 88.89 | 0.00 |

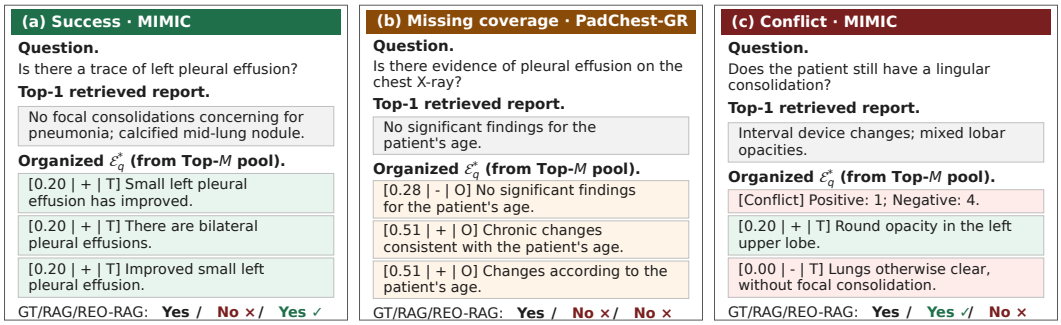


Figure 3: Qualitative cases from frozen holdout organization dumps (MIMIC $\tau$=0.35, PadChest-GR $\tau$=0.55). Gray: Top-1 report; green/orange/red: target, off-target, and polarity-disagreement units; "T/O" and "+/−": target match and polarity. Organized $\mathcal{E}_q^*$ is built from the Top-$M$ pool.

### 4.5 Qualitative Analysis

Figure 3 shows three cases. In (a), under $\tau$=0.35, the Top-1 report omits the queried finding while the Top-$M$ pool still contains target statements; REO-RAG retains low-risk target units and answers correctly. In (b), no retrieved report mentions the finding, so under $\tau$=0.55 organization admits off-target positives and cannot invent coverage—both RAG and REO-RAG fail. In (c), near-tied POS/NEG target support is kept with an explicit `[Conflict]` cue; the generator may still err, showing that organization surfaces disagreement rather than guaranteeing the answer. These cases illustrate when organization helps (noisy but covered pools) and when it cannot (missing target statements).

## 5 Conclusion

This work studied how retrieval-augmented Med-VQA uses—rather than only retrieves—radiology reports in binary, finding-oriented chest X-ray VQA. We introduced REO-RAG, which converts retrieved reports into structured clinical units and organizes them into a compact, question-aligned context. Matched controls show that AC-DPO contributes substantially to absolute F1, while evidence organization provides an additional benefit when the retrieved pool already contains the queried finding but also contains distracting or heterogeneous reference statements. Under sparse PadChest-GR coverage, report-level context can remain stronger. Ablations further show that target relevance is the dominant signal among the tested factors; the current experiments do not establish that every auxiliary risk factor is independently necessary. More broadly, the evidence supports a practical but narrower conclusion: question-target-aware structuring can improve how a Med-LVLM consumes retrieved radiology evidence, but it cannot recover findings absent from the retrieval pool and its textual risk score does not guarantee image-level clinical correctness. Stronger target-only controls, independent validation of the structured parser, direct image–evidence consistency modeling, and broader non-binary evaluation remain important next steps.

### AI USE STATEMENT

In this work, we used generative AI tools to assist with language editing, manuscript organization, and limited brainstorming and technical discussion during the development process. Generative AI tools were also used for suggestions on code debugging and implementation; all such suggestions were reviewed, adapted where appropriate, and verified by the authors through independent implementation and experiments.

We did not use generative AI tools to generate or fabricate experimental results, data, figures, or citations. The final research questions, methodological decisions, experimental design, implementation choices, result verification, and scientific conclusions were made and validated by the authors. We take full responsibility for the final content of this work, including all text, claims, results, and artifacts.

### ETHICS STATEMENT

This study uses publicly available chest X-ray datasets (IU-Xray, MIMIC-CXR, and PadChest/PadChest-GR) under their respective licenses and data use agreements. No new human subjects were recruited for this work. The method is intended to improve factual grounding of medical VQA systems and is not a clinical decision-support product; automatic Acc/F1 metrics do not replace radiologist review. Potential risks include over-reliance on retrieved report evidence and residual polarity disagreement among retrieved references after organization, which we discuss in the limitations.

### REPRODUCIBILITY STATEMENT

We provide detailed experimental protocols in Appendix B (data splits, preprocessing scripts, metrics/parser, hyperparameters, and baseline packaging) and extended analyses in Appendix C. Risk weights, thresholds, prompts, and the deterministic evidence mapper $\Phi$ are specified in Appendices A–B. Code, configuration files, and evaluation scripts will be released to support reproduction of the reported results under the shared CLIP index and parser protocol.

## REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C. Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, Sam Bond-Taylor, Atilla P. Kiraly, Ben Glocker, 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, 2023.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. PadChest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis*, 66: 101797, 2020. doi: 10.1016/j.media.2020.101797.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829, 2023.

Daniel Coelho de Castro, Aurelia Bustos, Shruthi Bannur, Stephanie L. Hyland, Kenza Bouzid, Maria Teodora Wetscherek, Maria Dolores Sánchez-Valverde, Lara Jaques-Pérez, Lourdes Pérez-Rodríguez, Kenji Takeda, José María Salinas-Serrano, Javier Alvarez-Valle, Joaquín Galant-Herrero, and Antonio Pertusa. PadChest-GR: A bilingual chest x-ray dataset for grounded radiology report generation. *NEJM AI*, 2(7), 2025. doi: 10.1056/AIdbp2401120.

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 sharing. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016. doi: 10.1093/jamia/ocv080.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2024.

Sunan He, Yuxiang Nie, Hongmei Wang, Shu Yang, Yihui Wang, Zhiyuan Cai, Zhixuan Chen, Yingxue Xu, Luyang Luo, Huiling Xiang, Xi Lin, Mingxiang Wu, Yifan Peng, George Shih, Ziyang Xu, Xian Wu, Qiong Wang, Ronald Cheong Kin Chan, Xiaohui Duan, Varut Vardhanabhuti, Winnie Chiu Wing Chu, Yefeng Zheng, Pranav Rajpurkar, Kang Zhang, and Hao Chen. Towards generalizable AI in medicine via generalist–specialist collaboration. *Nature Biomedical Engineering*, 2026. doi: 10.1038/s41551-026-01653-3.

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 (ICLR)*, 2022.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, number 01 in AAAI Conference on Artificial Intelligence, pp. 590–597, 2019. doi: 10.1609/aaai.v33i01.3301590.

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.

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.

Jason J. Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5(1):180251, 2018. doi: 10.1038/sdata.2018.251.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023.

Xuze Li, Haozhao Wang, Zhenyu Huang, Zhongxu Wang, Jinghua Zhang, and Ruixuan Li. MR-RAG: Multimodal relevance-aware retrieval-augmented generation for medical visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15010–15019, June 2026.

X. Liang, D. Wang, B. Jing, Z. Jiao, R. Li, R. Liu, Q. Miao, and Q. Wang. Fine-grained knowledge fusion for retrieval-augmented medical visual question answering. *Information Fusion*, 120: 103059, 2025.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1650–1654, 2021. doi: 10.1109/ISBI48211.2021.9434010.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

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 International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

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, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.28. URL https://aclanthology.org/2025.naacl-long.28/.

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, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.256.

Peng Xia, Kangyu Zhu, Haoran Li, Hongtu Zhu, Yun Li, Gang Li, Linjun Zhang, and Huaxiu Yao. RULE: Reliable multimodal RAG for factuality in medical vision language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1081–1093, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.62. URL https://aclanthology.org/2024.emnlp-main.62/.

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 *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=s5epFPdIW6.

Wenchuan Zhang, Jingru Guo, Hengzhe Zhang, Penghao Zhang, Jie Chen, Shuwan Zhang, Zhang Zhang, Yuhao Yi, and Hong Bu. Patho-AgenticRAG: Towards multimodal agentic retrieval-augmented generation for pathology VLMs via reinforcement learning. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence*, number 35 in AAAI Conference on Artificial Intelligence, pp. 29921–29929, 2026. doi: 10.1609/aaai.v40i35.40239.

# Appendix

This supplement is organized as follows:

- Appendix A (Method Details) expands the Method section.
- Appendix B (Experimental Setup and Details) covers data, metrics, hyperparameters, baselines, and computing infrastructure.
- Appendix C (Results Analysis) covers oracle diagnostics, retrieval diagnostics, ablations (including coverage-aware routing), and efficiency.
- Appendix D (Limitations) discusses limitations.

## A Method Details

This section expands the Method section of the main paper. Appendix A.1 details structured evidence construction; Appendix A.2 lists prompt templates used at generation. End-to-end inference matches Algorithm 1 in the main paper (implementation in the Code and Data Supplement).

### A.1 Structured Evidence Construction

This subsection expands structured evidence construction: the deterministic evidence mapper $\Phi$ used to form structured evidence units. $\Phi$ is *not* a learned NER: it applies a controlled clinical vocabulary and assertion rules so that finding, anatomy, assertion, and polarity attributes remain inspectable and reproducible. Organization (Stage II) consumes these attributes; $\Phi$ itself only *exposes* them.

#### A.1.1 Pipeline

Given a retrieved report $R$, $\Phi$ proceeds as:

1. **Segment.** Normalize whitespace and split on sentence boundaries (period/question/exclamation), capping sentences per report.
2. **Assert.** Tag each sentence as asserted / negated / uncertain using hedge and sentence-initial negation cues.
3. **Find.** Match a controlled finding vocabulary; sentences without a vocabulary hit are dropped from the structured pool.
4. **Locate.** Extract anatomy tokens with a shared laterality/region regex.
5. **Polarize.** Map tags to $\{\mathrm{POS}, \mathrm{NEG}, \mathrm{UNC}\}$ and refine with local/blanket negation around the finding span.
6. **Align.** Mark target match $m_i$ if the unit's finding intersects the question's target set (with synonyms).
7. **Serialize.** Emit a tagged string for the generator.

#### A.1.2 Assertion Cues, Lexicon, and Anatomy

**Sentence-level assertion cues.** Uncertainty matches (case-insensitive) `cannot exclude`, `possible(ly)`, `may represent/may be`, `suggest(s|ed|ing)`, `suspicious for`, `questionable`, or `indeterminate`. Sentence-initial negation fires on `no`, `without`, `free of`, `absence of`, `negative for`, `no evidence of`, `no sign of`, `no acute/no focal`, `unremarkable`, `clear`, `normal`, `within normal`, `denies`, or `rule out`. Uncertainty is checked before start-anchored negation; mixed tags are retained (e.g., `negated+uncertain`).

**Finding vocabulary.** Chest findings are a controlled vocabulary of canonical keys with surface patterns (Table 6). Matching is regex-based and case-insensitive. Question–finding aliases merge clinically related keys for target alignment only (e.g., `consolidation`↔{`pneumonia`,

Table 6: Core finding keys exposed by $\Phi$ (surface forms abbreviated). Full pattern lists are released with the code.

| Key | Example surface cues |
|---|---|
| pleural_effusion | effusion(s); pleural fluid |
| cardiomegaly | cardiomegaly; enlarged cardiac |
| atelectasis | atelectas* |
| consolidation | consolidat* |
| opacity | opacity/opacities; opacification |
| pneumothorax | pneumothora* |
| edema | (pulmonary/interstitial) edema |
| pneumonia | pneumonia; pneumonic |
| mass / nodule | mass(es); nodul* |
| infiltrate | infiltrate* |
| fracture | fracture* |
| calcification | calcificat*; calcified |
| pleural_thickening | pleural thicken*; thickening |
| emphysema | emphysem* |
| lung_clear / normal_* | lungs clear; normal cardiac/chest |

infiltrate, opacity}). PadChest-GR may optionally activate an extended ontology JSON; IU-Xray/MIMIC use the default chest vocabulary.

**Anatomy.** Tokens are extracted with left/right/bilateral, upper/lower/apical/basal/base(s), pleural/perihilar/mediastinal/hilar/costophrenic/retrocardiac, and lung(s)/cardiac/heart/hemithorax. At most four unique tokens are kept per unit; missing anatomy does not block target match unless the question states an anatomy constraint.

### A.1.3 Polarity, Targets, and Serialization

**Polarity resolution.** Base polarity follows the sentence tag: uncertainty$\rightarrow$ UNC, negation$\rightarrow$ NEG, else POS. We then refine around the finding span with local hedges, blanket-negation spans, and cues such as not seen/not identified/ruled out. For prompt status we map POS $\rightarrow$ PRESENT, NEG $\rightarrow$ ABSENT, UNC $\rightarrow$ POSSIBLE when bundle serialization is used; the main-table REO-RAG path instead serializes linearized sentences with [Negated]/[Uncertain] prefixes (structured_v2). Polarity is used for conflict scoring and Yes/No organization; it does *not* discard target-negative evidence on presence questions.

**Question targets and serialization.** Targets $f_q$ are lexicon keys (plus aliases) matched in the question text. A unit is target-matched ($m_i$=1) iff its finding set intersects $f_q$. When a normality finding (e.g. heart_size_normal) aliases an abnormality question target (e.g. cardiomegaly), target-relative polarity is inverted so that positively asserted normality supports *No*.

**Support weight $\alpha_i$.** Polarity organization aggregates admitted target units with non-negative masses $\alpha_i$, used only to form $S_+ = \sum_{e_i \in G_q^+} \alpha_i$ and $S_- = \sum_{e_i \in G_q^-} \alpha_i$ (and the dominance margin in the main paper); $\alpha_i$ is *not* a calibrated confidence and does *not* reuse coarse CLIP image–report similarity. Let $T(q)$ and $T(x_i)$ be whitespace token sets of the question and evidence sentence. A raw lexical hit count $c_i = |T(q) \cap T(x_i)|$ yields two normalized scores over the current candidate pool:

$$\rho_i = \frac{(1{+}0.1\, c_i) - \min_j(1{+}0.1\, c_j)}{\max_j(1{+}0.1\, c_j) - \min_j(1{+}0.1\, c_j) + \varepsilon}, \tag{14}$$

$$o_i = \frac{c_i}{\max_j c_j + \varepsilon}. \tag{15}$$

with the convention that a constant pool maps $\rho_i \leftarrow 1$ (and $o_i \leftarrow 0$ if all counts are zero). Writing $u_i = r_{\mathrm{u}}(e_i, q) \in [0, 1]$ for the unit uncertainty risk,

$$\alpha_i = \big(\max(\rho_i, 0.05) + 0.5\, \max(o_i, 0)\big)\,(1 - 0.6\, u_i). \tag{16}$$

Thus every unit keeps a small floor mass, higher question–sentence overlap increases support, and uncertain assertions are attenuated before Yes/No dominance is decided.

Units are linearized as report-derived evidence sentences with optional prefixes `[Negated]` / `[Uncertain]` (and a leading `[Conflict]` cue when Stage II retains both polarities under a near-tied margin). Empty admission falls back to the single lowest-risk unit.

**Reproducibility.** All patterns are frozen regular expressions in the released code (*med_report_graph.py*, *med_report_bundle.py*). No corpus-specific fine-tuning of $\Phi$ is performed for the main tables.

### A.2 PROMPTS

Prompt templates used by the generator after evidence organization (Method) and by all reproduced baselines (Experiments). All methods use conversation mode `vicuna_v1`. Below we quote the user-message templates used in the evaluation script ({`topk`} and {`evidence`} are filled at inference; the question string follows).

**REO-RAG / Structured-v2 Prompt.**

```
You are provided with a chest X-ray image, an image-related
question, and {topk} piece(s) of structured medical evidence
(report-derived sentences with negation/uncertainty tags where
applicable):

{evidence}

Please answer based on the image and the evidence.  The
evidence is for reference only and must not replace your own
visual assessment.  Answer with ``Yes'' or ``No'' first, then
briefly explain.  Your response must start with the word Yes or
the word No.  Question:  {question}
```

**Report-level / FactMM-RAG / MMed-RAG Prompt.**

```
You are provided with a chest X-ray image, an image-related
question and {topk} reference report(s):

{evidence}

Please answer the question based on the image and report and
choose from the following two options:  [yes, no].  It should
be noted that the diagnostic information in the reference
reports cannot be directly used as the basis for diagnosis, but
should only be used for reference and comparison.  Question:
{question}
```

**MedDr Prompt.**

```
You are provided with a chest X-ray image, an image-related
question, and {topk} similar case(s) retrieved from visually
similar examinations in the training corpus.  Each case below
is a radiology report from a similar patient:
```

```
{evidence}

Use these similar cases only as diagnostic reference and
comparison.  Do not copy a case diagnosis as the final answer;
always ground the decision in the query image.  Answer with
``Yes'' or ``No'' first, then briefly explain.  Your response
must start with the word Yes or the word No.  Question:
{question}
```

**MR-RAG Prompt.**

```
You are provided with a chest X-ray image, an image-related
question and {topk} reference summary(s):

{Report i.  truncated summary}

Please answer the question based on the image and report and
choose from the following two options:  [yes, no].  It should
be noted that the diagnostic information in the reference
reports cannot be directly used as the basis for diagnosis,
but should only be used for reference and comparison.  Your
response must start with the word Yes or the word No.
Question:  {question}
```

Long MIMIC/PadChest-GR summaries in the MR-RAG path are truncated (default 280 characters) so the Yes/No instruction remains in context. Serialization helpers and the answer parser are released with the code.

# B Experimental Setup and Details

This section records the experimental protocol underlying the Experiments section of the main paper: Appendix B.1 covers data splits, preprocessing, metrics, and hyperparameters; Appendix B.2 specifies the unified baseline protocol; Appendix B.3 lists computing infrastructure. Result-facing analyses appear in Appendix C.

## B.1 Experimental Setup

### B.1.1 Data Statistics

We evaluate binary yes/no chest X-ray VQA on IU-Xray, MIMIC-CXR, and PadChest-GR (de Castro et al., 2025) (derived from PadChest (Bustos et al., 2020)). IU-Xray and MIMIC-CXR are the primary in-domain settings. PadChest-GR is a *target-domain calibrated cross-dataset* evaluation: the IU AC-DPO generator adapter is not retrained on PadChest-GR, but the risk threshold $\tau$ is selected on a PadChest-GR calibration split of $N{=}300$ labeled questions (likewise, $\tau$ is calibrated separately on IU-Xray and MIMIC-CXR). For each dataset the retrieval corpus is built *only* from the training split (no calibration or test leakage). A $N{=}300$ calibration set per dataset is used solely for selecting $\tau$. Calibration is question-id disjoint from holdout; PadChest-GR is additionally patient-disjoint, while IU/MIMIC follow the published question-level split and may share images/studies across different questions (details in Sec. B.1.2).

Following a shared deterministic parser for non-binary filtering, the main VQA evaluation uses $1{,}870$ IU-Xray, $3{,}166$ MIMIC-CXR, and $416$ PadChest-GR questions. *Available* denotes the fixed holdout pool size; *Used* denotes the parser-valid subset after removing non-binary answers (IU/MIMIC). PadChest-GR-416 is a constructed cross-domain evaluation subset (classic6 yes/no questions from

Table 7: Dataset statistics and evaluation roles. Available: holdout pool size; Used: parser-valid questions in the main VQA eval ("–" for diagnostic-only pools).

| Dataset | Available | Used | Purpose |
|---|---|---|---|
| IU-Xray | 1,973 | 1,870 | Main VQA |
| MIMIC-CXR | 3,170 | 3,166 | Main VQA |
| PadChest-GR-416 | 416 | 416 | Main VQA (classic6) |
| PadChest-GR-1462 | 1,462 | – | Similarity diagnosis |
| PadChest-GR-1215 | 1,215 | – | Recall@$K$ analysis |

English finding sentences; not a native PadChest-GR VQA benchmark, and not a random subsample of a larger pool). PadChest-GR-1462 and PadChest-GR-1215 are appendix-only diagnostic pools for similarity routing and Evidence Recall@$K$, and are *not* used in the main comparison tables. Table 7 summarizes these roles.

### B.1.2 Data Preprocessing Scripts

All code required to rebuild the evaluation JSONLs from each dataset's full VQA test pool is included in the Code and Data Supplement (also summarized in *scripts/PREPROCESS.md*). Upstream pools live under *data/test/*; raw clinical images and reports are not redistributed and must be obtained under the original dataset licenses.

**IU-Xray.** Upstream pool: `data/test/iuxray_test.jsonl` (full VQA test pool). *prepare_iuxray_splits.py* shuffles with seed 42, writes the first 600 rows to *iuxray_calib.jsonl* and the remainder to `iuxray_test_holdout_clean.jsonl` (paper Available=1,973). From the 600-row pool it then reshuffles with seed 42, keeps only binary ground-truth answers (yes/no prefix), and takes the first $N$=300 as the $\tau$-calibration set (released path used in our runs: `results/iuxray/frc_structured_calib/calib_question_size300.jsonl`).

**MIMIC-CXR.** Upstream pool: *data/test/mimic_test.jsonl* (full VQA test pool). *prepare_mimic_splits.py* performs a *positional* split with *no* shuffle: the first 300 lines become *mimic_calib_n300.jsonl*; the remainder become *mimic_test_holdout.jsonl* (paper Available=3,170).

**PadChest-GR.** *build_padchest_vqa.py* constructs classic6 yes/no questions and patient-disjoint calib/holdout splits (Protocol A), including the main-table holdout ($N$=416) and the $N$=300 $\tau$-calibration set used for PadChest-GR.

IU/MIMIC calib and holdout sets are question-id disjoint. They are *not* guaranteed image-/study-disjoint: the same study may contribute different questions to calibration and holdout (IU shows substantial image overlap under the published question-level split; MIMIC shows small study-level overlap). PadChest-GR is patient-disjoint. Calibration is used only to select $\tau$; still, image-/study-level leakage is a residual risk of the question-level split. The paper's Used counts (1,870 / 3,166) are obtained at evaluation time by the shared deterministic ground-truth parser (yes/no prefix match; non-binary labels are skipped), *not* by rewriting the holdout JSONLs—Available sizes therefore remain 1,973 / 3,170 on disk.

### B.1.3 Metrics and Answer Parser

We report accuracy, precision, recall, and positive-class F1 (primary) on all three datasets. For PadChest-GR we additionally report finding-aware Macro-F1 over the classic6 findings (unweighted mean of per-finding positive-class F1). Table 8 lists PadChest-GR overall F1 and Macro-F1 for the main-table systems under the shared parser. All methods share a deterministic answer parser: predictions that do not begin with a yes/no token are marked unparsed and counted as incorrect under the extended evaluation mode used in the main tables. Ground-truth answers that are not binary are excluded from the "Used" counts in Table 7.

Table 8: PadChest-GR finding-aware Macro-F1 (%) under the shared CLIP protocol ($N$=416, classic6). Overall F1 matches the main-paper comparison table; Macro-F1 is the unweighted mean of per-finding positive-class F1. Best in each column is bold.

| Method | Overall F1 | Macro-F1 |
|---|---|---|
| Report-level RAG | 25.69 | 23.42 |
| MedDr He et al. (2026) | 33.88 | 31.53 |
| RULE Xia et al. (2024) | 53.74 | 22.55 |
| FactMM-RAG Sun et al. (2025) | 36.36 | 33.77 |
| MMed-RAG Xia et al. (2025) | 35.26 | 32.80 |
| MR-RAG Li et al. (2026) | 44.88 | **41.62** |
| REO-RAG | **56.05** | 36.18 |

Table 9: Key hyperparameters for REO-RAG (main tables).

| Component | Setting |
|---|---|
| Image encoder / index | OpenCLIP RN50-quickgelu (CC12M); train-split corpus |
| Coarse pool size $M$ | 8 |
| Evidence budget $K_{\max}$ | 10 |
| Risk threshold $\tau$ | 0.15 / 0.35 / 0.55 (IU / MIMIC / PadChest-GR) |
| Polarity dominance $\delta$ | 0.25 |
| Mixture weight $\alpha$ / CLIP margin | 0.35 / 0.10 (when mixing structured ranks) |
| Generator | LLaVA-Med-1.5 (Mistral-7B), vicuna_v1 |
| Decoding | Temperature 0, beam size 1 (greedy) |
| AC-DPO LoRA | $r$=128, $\alpha$=256; 3 epochs; LR $1\times10^{-5}$ |
| AC-DPO data / length | IU answer-consistency pairs; max length 768; bf16; ZeRO-3 |
| Adapter transfer | IU adapter transferred without target-domain training (MIMIC / PadChest-GR) |

### B.1.4 Hyperparameter Settings

Table 9 lists the frozen settings used for the main tables. Risk weights follow clinical priors and are shared across datasets:

$$\mathbf{w} = \frac{1}{88}\big(w_{\mathrm{tr}}, w_{\mathrm{od}}, w_{\mathrm{u}}, w_{\mathrm{ec}}, w_{\mathrm{ge}}\big) = \frac{1}{88}(20, 25, 15, 18, 10).$$

**Selection protocol (development).** We distinguish calibrated knobs from fixed design choices; no hyperparameter is retuned on the frozen holdouts.

- **Risk threshold $\tau$ (calibrated).** On each dataset we evaluate an 8-point grid on a disjoint calibration split ($N$=300): IU-Xray $\tau \in \{0.15, 0.20, \ldots, 0.50\}$, MIMIC-CXR / PadChest-GR $\tau \in \{0.35, 0.40, \ldots, 0.70\}$. The operating point maximizes positive-class F1 subject to fallback rate $\leq 25\%$ (if the admissible set is empty, keep the lowest-risk unit); full grids are in Appendix C.2 (Table 14).
- **Evidence budget $K_{\max}$.** Fixed to $K_{\max}$=10 as a soft cap on admitted units. Sensitivity under frozen $\tau$ and AC-DPO compares Top-$K \in \{1, 3, 5\}$, the unfiltered pool (*All*), and $K_{\max}$=10 (Table 15); gains track risk admission rather than raw length.
- **Risk weights w.** A single clinical prior (Eq. above), shared across datasets and *not* optimized on calibration or holdout. Sensitivity vs. a uniform prior ($w_k$=1/5) and leave-one-out factor ablations appear in Appendix C.2 (Table 16 and Sec. C.2.4); the prior is useful but not brittle.
- **Coarse pool $M$, polarity margin $\delta$, rank mixture $(\alpha, \mathrm{margin})$.** Fixed after small development probes on calibration/dev runs and then frozen: $M$=8; $\delta$=0.25; $(\alpha, \mathrm{margin}) = (0.35, 0.10)$ when mixing structured ranks (development grid $\alpha \in \{0.25, 0.30, 0.35\} \times \mathrm{margin} \in \{0.05, 0.10, 0.15\}$, selected by calibration F1).

Table 10: Unified protocol for main-table methods.

| Method | Evidence construction | Generator / adapter |
|---|---|---|
| Report-level RAG | CLIP Top-1 full report | Base LLaVA-Med |
| MedDr (He et al., 2026) | CLIP Top-4 similar reports + MedDr similar-case prompt | Base LLaVA-Med |
| FactMM-RAG (Sun et al., 2025) | CLIP Top-8 pool, fact-aware re-rank to Top-1 | Base LLaVA-Med |
| MR-RAG (Li et al., 2026) | Multimodal Cooperative Retrieval; Top-4 report summaries (matched budget) | Base LLaVA-Med (no IFA) |
| RULE (Xia et al., 2024) | Structured selection, $k$=4 | RULE preference adapter |
| MMed-RAG (Xia et al., 2025) | CLIP Top-1 full report | MMed preference DPO |
| REO-RAG | $M$=8 pool $\rightarrow$ $\Phi$ units $\rightarrow$ risk organization ($K_{\max}$=10) | IU AC-DPO (transferred; no target-domain training) |

- **AC-DPO / decoding.** LoRA ($r$=128, $\alpha$=256), LR $1\times10^{-5}$, 3 epochs, and max length 768 follow the released training script defaults and are not swept for the main tables. Decoding is greedy (temperature 0, beam size 1) for determinism.

### B.2 Baselines

Following the fairness practice in recent medical RAG papers (Xia et al., 2025; Li et al., 2026), all main-table rows share: (i) the same OpenCLIP RN50-quickgelu / CC12M image index over each dataset's *training* reports; (ii) the same frozen holdouts and calibration splits; (iii) LLaVA-Med-1.5 (Mistral-7B) as the generator unless a method's defining preference adapter is part of the reproduced system; (iv) greedy decoding and the same yes/no parser. We therefore reproduce prior retrieval/organization designs under a shared backbone, not proprietary generators or private indices. Official MedDr (InternVL) and FactMM's unreleased generator training are *not* swapped in as generators, because that would break the shared-backbone claim. Table 10 summarizes the evidence construction and generator/adapter for each main-table row; the paragraphs below spell out the packaging choices.

$\tau$ for each main-table dataset is selected on that dataset's own $N$=300 calibration split (target-domain calibrated), including PadChest-GR.

**Report-level RAG.** Retrieve the single most similar training examination by CLIP cosine similarity and append its full radiology report to the prompt.

**MedDr.** Retrieve Top-4 reports from the shared CLIP pool and use a MedDr/GSCo "similar cases" prompt with base LLaVA-Med (He et al., 2026).

**FactMM-RAG.** Within the CLIP Top-8 pool, re-rank reports by whether they mention the queried finding (fact-aware selection approximating FactMM's fact mining (Sun et al., 2025) without their private MARVEL index), then keep Top-1.

**MR-RAG.** Follow Multimodal Cooperative Retrieval with Top-$K$=4 report summaries under the shared CLIP/BioClinicalBERT packaging (Li et al., 2026), matched to the MedDr evidence budget; base LLaVA-Med without the official IFA adapter.

**RULE / MMed-RAG.** Reproduce structured $k$=4 selection with the RULE preference adapter (Xia et al., 2024), and Top-1 report RAG with the MMed preference DPO adapter (Xia et al., 2025), under otherwise matched inference settings.

**Controlled ablations.** The main paper reports (i) an evidence$\times$generator control (base LLaVA-Med vs. IU AC-DPO on shared JSONLs) and (ii) a representation ablation under frozen AC-DPO (report / sentence / structured / REO-RAG). Matched-adapter RULE-style ($k$=4) F1 under AC-DPO is 86.61 / 88.67 / 52.77 on IU / MIMIC / PadChest-GR (below REO-RAG on all three).

Table 11: Oracle-context diagnostics (F1, %). Gold report: paired report as raw evidence; template polarity: aligned finding-status cue (not a competitive baseline).

| Dataset | Setting | F1 (%) |
|---|---|---|
| IU-Xray | REO-RAG / gold report | 88.60 / 90.75 |
| MIMIC-CXR | REO-RAG / gold report | 88.89 / 95.30 |
| PadChest-GR-416 | Image-only / REO-RAG | 43.50 / 56.05 |
| PadChest-GR-416 | Gold report (raw) | 58.42 |
| PadChest-GR-416 | Template polarity | 100.0 |

### B.3 Computing Infrastructure

All main-table experiments were run on Linux servers equipped with NVIDIA GeForce RTX 4090 GPUs (24 GB) and NVIDIA H100 GPUs (80 GB), with CUDA 12.1 and Python 3.10. The software stack used for training and inference is: PyTorch 2.1.2, Transformers 4.36.2, PEFT 0.4.0, Accelerate 0.21.0, DeepSpeed 0.12.6 (ZeRO-3 for AC-DPO), TRL, OpenCLIP, and NumPy 1.26.4. AC-DPO training uses multi-GPU ZeRO-3; holdout inference and retrieval are typically single-GPU. Exact package pins are also listed in the Code and Data Supplement (*README.md* / *ENVIRONMENT.md*).

## C Results Analysis

This section expands the analyses in the Experiments section of the main paper: Appendix C.1 covers oracle-context diagnostics, frozen retrieval diagnostics, finding-wise F1, and the attention case study; Appendix C.2 reports threshold / $K_{\max}$ / weight sensitivity, coverage-aware routing, and leave-one-out ablations; Appendix C.3 compares evidence-context efficiency.

### C.1 Analysis of Evidence Organization

This subsection expands the analysis of evidence organization in the main paper: oracle-context diagnostics, frozen retrieval diagnostics, finding-wise F1, and a qualitative attention case study (Figure 5).

#### C.1.1 Oracle-Context Diagnostics

Table 11 localizes remaining bottlenecks rather than ranking systems. Replacing retrieved context with the query's paired gold report (raw text; the answer string is not injected) raises IU/MIMIC F1 to $90.75\%$ / $95.30\%$, indicating residual retrieval/coverage headroom in-domain. On PadChest-GR ($N$=416), the same gold report reaches only $58.42\%$ F1 ($+2.37\%$ over REO-RAG), so domain shift is not explained by missing retrieval alone. A template-polarity cue that states the ground-truth finding status without free-form report wording reaches $100\%$ F1 and mainly verifies that the generator and yes/no parser can follow an explicit assertion.

#### C.1.2 Similarity as a Sample-Level Router

We examine whether CLIP similarity alone can indicate when retrieved evidence is trustworthy. Table 12 summarizes Top-1 CLIP similarities on diagnostic pools. MIMIC-CXR has the highest mean similarity and PadChest-GR the lowest, reflecting population-level domain proximity. On PadChest-GR, however, true-positive and false-negative mean similarities differ by only $0.003$ ($0.421$ vs. $0.418$). Thresholding on CLIP similarity therefore cannot reliably decide whether retrieved evidence should be trusted for an individual question.

#### C.1.3 Target-Finding Evidence Recall@$K$

Table 13 measures Evidence Recall@$K$: the fraction of questions for which at least one of the Top-$K$ retrieved reports mentions the queried finding. On 1,215 PadChest-GR finding questions,

Table 12: Top-1 CLIP similarity on diagnostic pools. TP/FN: PadChest-GR correct/incorrect positives.

| Dataset / subset | $N$ | Similarity |
|---|---|---|
| IU-Xray | 1,973 | $0.485 \pm 0.018$ |
| MIMIC-CXR | 3,170 | $0.557 \pm 0.018$ |
| PadChest-GR | 1,462 | $0.419 \pm 0.029$ |
| TP / FN | – | 0.421 / 0.418 |

Table 13: Target-finding Evidence Recall@$K$ (%) on PadChest-GR ($N$=1,215).

| Metric | $K$=5 | $K$=10 | $K$=20 |
|---|---|---|---|
| Evidence Recall@$K$ | 11.0 | 17.5 | 27.9 |

coverage is only $11.0\%$ / $17.5\%$ / $27.9\%$ at $K$=5/10/20. Increasing retrieval depth improves coverage slowly and still leaves most questions without target evidence at $K$=20. These diagnostics motivate treating organization as an evidence-*utilization* step that cannot invent missing findings.

#### C.1.4 Finding-wise F1

Aggregate F1 in the main tables can hide finding-level heterogeneity. Figure 4 compares finding-wise positive-class F1 for REO-RAG, RULE, and MMed-RAG on the same holdouts and parsers used in the main-paper comparison table. IU-Xray and MIMIC-CXR findings are obtained by keyword matching on the question text (top-12 findings by support); PadChest-GR uses the classic6 label field. Radial length is F1 (0 at the inner ring to 1.0 at the outer ring); a short stub marks F1=0. REO-RAG is competitive on most high-support findings on IU-Xray and MIMIC-CXR, while PadChest-GR remains sparse and finding-dependent under domain shift—consistent with the coverage-limited analysis in the main paper.

#### C.1.5 Attention Map Case Study

Figure 5 visualizes last-layer self-attention on a MIMIC-CXR holdout success case under the frozen IU AC-DPO generator (same adapter for both rows). For each context we teacher-force the model's first answer token (`Yes`/`No`) and average attention over heads from that query position to the multimodal prompt. Image attentions are reshaped to the $24\times24$ CLIP patch grid and overlaid on the radiograph under a shared color scale; text tokens (retrieved report or REO units in the prompt) are shown as 1D strips with per-row peak normalization so each row has a visible peak while mid-band clutter remains comparable. Report-level RAG answers incorrectly from a Top-1 report that states "no pleural effusion," while REO-RAG answers correctly after organizing supporting evidence; image-token attention mass rises from $0.51$ to $0.67$.

### C.2 Ablation Studies

This subsection expands the ablation studies in the main paper: threshold / $K_{\max}$ / weight sensitivity and leave-one-out factor ablations.

#### C.2.1 Risk-Threshold Sensitivity

#### C.2.2 Evidence-Size / $K_{\max}$ Sensitivity

A natural concern is that REO-RAG simply feeds a longer context than fixed Top-$k$ truncations. Table 15 freezes the IU AC-DPO adapter and main-table $\tau$, and varies only the evidence budget on the *same* scored sentence pool rebuilt from coarse report indices: Top-$K$ keeps the $K$ lowest-admission-score units without the risk threshold ($K \in \{1, 3, 5\}$); *All* keeps the full scored pool (no $\tau$ filter, no $K_{\max}$); REO-RAG applies risk admission with $K_{\max}$=10. This construction enforces the per-question invariant $|\mathcal{E}_q^*| \leq |\mathcal{E}_{\text{All}}|$; the corresponding mean lengths are $\overline{K}$=2.81 vs. 5.66 on IU-Xray, 9.77 vs. 62.48 on MIMIC-CXR, and 9.68 vs. 20.89 on PadChest-GR. Thus REO-

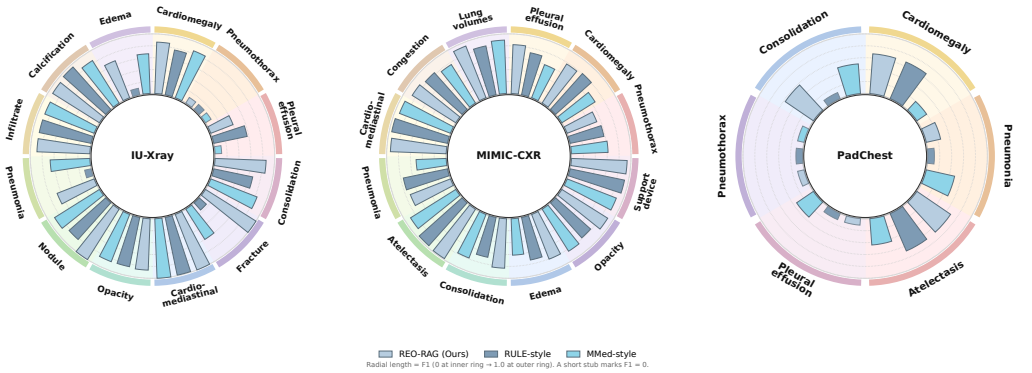


Figure 4: Circular visualization of finding-wise F1 on IU-Xray, MIMIC-CXR, and PadChest-GR. Each sector is a finding; the three bars are REO-RAG, RULE, and MMed-RAG under the shared CLIP protocol.

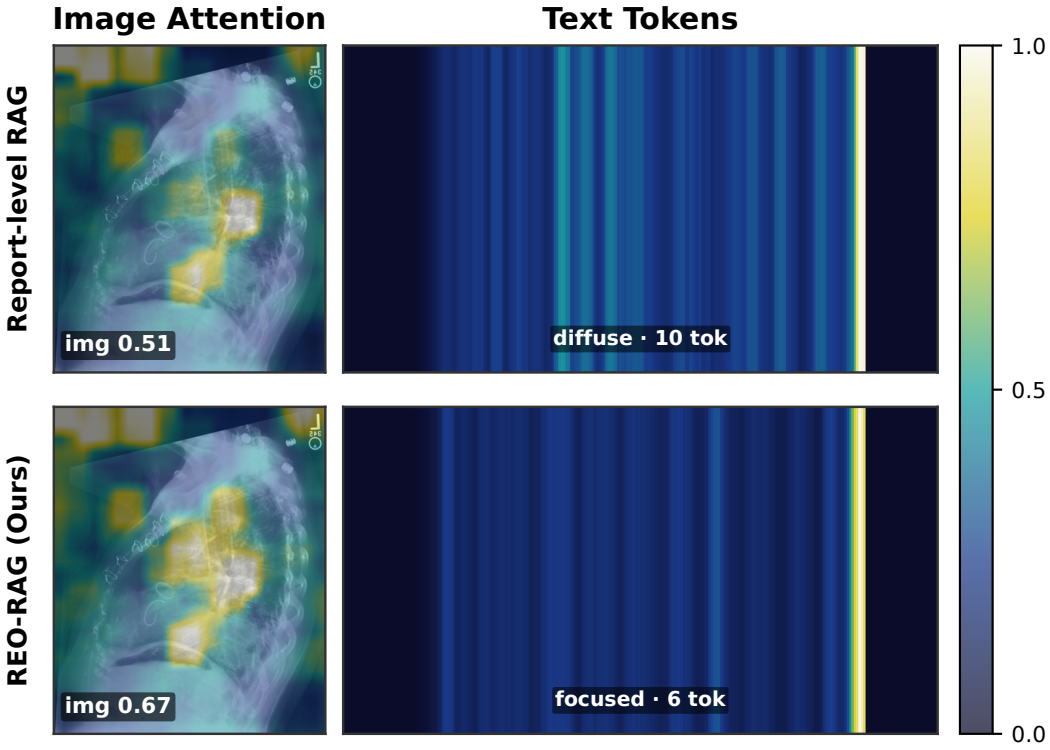


Figure 5: Last-layer answer-token attention on a screened MIMIC-CXR holdout case (shared AC-DPO generator). Left: image attention on the CXR (shared scale; mass 0.51 vs. 0.67). Right: text-token strips with per-row peak normalization (diffuse vs. focused; tokens covering 40% of text attention: 10 vs. 6). Report-level RAG is incorrect; REO-RAG is correct.

RAG is a *shorter* reliability-aware context than the unfiltered pool on every dataset. F1 values for Top-$k$/*All* under this aligned pool are re-inferred with the frozen AC-DPO adapter (Table 15); REO-RAG F1 matches the main-table organization dumps. On IU-Xray, REO-RAG attains the best F1 (88.60%) while using fewer units than Top-5/*All*, and length alone is non-monotonic (Top-3 87.27% > Top-1/5/*All*). On MIMIC-CXR, REO-RAG again leads (88.89%) over Top-5 (88.25%), while the unfiltered pool collapses to 73.77% at $\overline{K}$=62.48—longer context without risk admission

Table 14: Calibration F1 (%) vs. risk threshold $\tau$ ($N$=300 each). $\overline{K}$: mean retained units; FB: fallback rate. Bold rows are the operating points used in the main tables.

| Dataset | $\tau$ | Calib. F1 | $\overline{K}$ | FB |
|---|---|---|---|---|
| IU-Xray | **0.15** | **85.55** | **5.95** | **0.087** |
| | 0.20 | 85.06 | 6.71 | 0.087 |
| | 0.25 | 85.06 | 6.71 | 0.087 |
| | 0.30 | 84.09 | 9.34 | 0.000 |
| | 0.35 | 84.09 | 9.34 | 0.000 |
| | 0.40 | 85.06 | 9.66 | 0.000 |
| | 0.45 | 85.06 | 9.66 | 0.000 |
| | 0.50 | 85.06 | 9.66 | 0.000 |
| MIMIC-CXR | **0.35** | **91.36** | **9.77** | **0.000** |
| | 0.40 | 91.36 | 9.99 | 0.000 |
| | 0.45 | 91.36 | 9.99 | 0.000 |
| | 0.50 | 91.36 | 9.99 | 0.000 |
| | 0.55 | 91.36 | 10.0 | 0.000 |
| | 0.60 | 91.36 | 10.0 | 0.000 |
| | 0.65 | 91.36 | 10.0 | 0.000 |
| | 0.70 | 91.36 | 10.0 | 0.000 |
| PadChest-GR | 0.35 | 43.75 | 4.68 | 0.000 |
| | 0.40 | 51.02 | 7.63 | 0.000 |
| | 0.45 | 51.02 | 7.63 | 0.000 |
| | 0.50 | 51.02 | 7.67 | 0.000 |
| | **0.55** | **54.00** | **9.18** | **0.000** |
| | 0.60 | 54.00 | 9.18 | 0.000 |
| | 0.65 | 54.00 | 9.18 | 0.000 |
| | 0.70 | 54.00 | 9.18 | 0.000 |

Table 15: Evidence-size / $K_{\max}$ sensitivity under frozen AC-DPO and main-table $\tau$ (F1, %). $\overline{K}$: mean retained units on the aligned same-pool dumps. REO-RAG F1 matches the main-table organization; Top-*k*/*All* F1 are re-inferred on that pool. Bold: best F1 per dataset.

| | IU | | MIMIC | | PadChest-GR | |
|---|---|---|---|---|---|---|
| Setting | F1 | $\overline{K}$ | F1 | $\overline{K}$ | F1 | $\overline{K}$ |
| Top-1 | 87.10 | 1.00 | 86.66 | 1.00 | 50.11 | 1.00 |
| Top-3 | 87.27 | 3.00 | 88.08 | 3.00 | 57.03 | 3.00 |
| Top-5 | 86.91 | 4.20 | 88.25 | 5.00 | **59.44** | 4.97 |
| All (no $\tau$) | 86.78 | 5.66 | 73.77 | 62.48 | 56.96 | 20.89 |
| REO-RAG ($K_{\max}$=10) | **88.60** | 2.81 | **88.89** | 9.77 | 56.05 | 9.68 |

hurts sharply. On PadChest-GR, where $\tau$=0.55 admits nearly the full $K_{\max}$ budget ($\overline{K}$=9.68), a shorter Top-5 truncation of the same admission ranking reaches $59.44\%$ and outperforms REO-RAG ($56.05\%$), while the unfiltered pool ($56.96\%$ at $\overline{K}$=20.89) does not improve further—consistent with coverage/domain-shift limits rather than under-filtering.

### C.2.3 Weight Sensitivity

Table 16 compares the clinical prior $\mathbf{w} = (20, 25, 15, 18, 10)/88$ against a uniform prior over the five active factors ($w_k$=1/5), holding $\tau$, $K_{\max}$, AC-DPO, and polarity organization fixed. On IU-Xray, the clinical prior improves F1 by $1.17\%$ ($88.60\%$ vs. $87.43\%$). On MIMIC-CXR the two settings are essentially tied ($88.89\%$ vs. $88.87\%$). On PadChest-GR, overall F1 slightly favors uniform ($57.50\%$ vs. $56.05\%$), while Macro-F1 still favors the clinical prior ($36.18\%$ vs. $34.77\%$). Thus the frozen weights are a useful but not brittle design choice.

Table 16: Risk-weight sensitivity under main-table $\tau$ and frozen AC-DPO (F1, %). Default: $(20, 25, 15, 18, 10)/88$. Uniform: equal mass on the five active factors. $\overline{K}$: mean retained units. Bold: better F1 per dataset.

| | IU | | MIMIC | | PadChest-GR | |
|---|---|---|---|---|---|---|
| Weights | F1 | $\overline{K}$ | F1 | $\overline{K}$ | F1 | $\overline{K}$ |
| Uniform $(1/5)$ | 87.43 | 2.91 | 88.87 | 9.95 | **57.50** | 9.69 |
| Default (clinical) | **88.60** | 2.81 | **88.89** | 9.77 | 56.05 | 9.68 |

Table 17: Efficiency across main-table methods. Each cell is Tok. / $\overline{K}$: mean whitespace tokens in the evidence field and mean retained reports/units. MMed-RAG matches Report-level RAG's Top-1 full-report evidence under the shared protocol.

| Method | IU Tok/$\overline{K}$ | MIMIC Tok/$\overline{K}$ | PadChest-GR Tok/$\overline{K}$ |
|---|---|---|---|
| Report-level RAG | 42.3 / 1.0 | 81.7 / 1.0 | 23.5 / 1.0 |
| MedDr | 157.0 / 4.0 | 309.6 / 4.0 | 87.9 / 4.0 |
| FactMM-RAG | 42.3 / 1.0 | 84.4 / 1.0 | 24.1 / 1.0 |
| MR-RAG | 157.0 / 4.0 | 310.1 / 4.0 | 93.4 / 4.0 |
| RULE | 37.6 / 3.2 | 62.4 / 4.0 | 40.7 / 4.0 |
| MMed-RAG | 42.3 / 1.0 | 81.7 / 1.0 | 23.5 / 1.0 |
| REO-RAG | 31.5 / 2.8 | 112.9 / 9.8 | 81.3 / 9.7 |

### C.2.4 Risk-Factor Leave-One-Out

Unlike the main-paper reliability-component ablation on the IU/MIMIC holdout, the leave-one-out below is run on the IU-Xray *calibration* split ($N$=300, $\tau$=0.15), where we set each risk weight to zero in turn (plus a full-weight control). Full F1 is $85.55\%$; removing target relevance drops to $84.09\%$ ($\Delta$=$-1.46\%$), while removing off-target distraction rises to $86.05\%$ ($\Delta$=$+0.50\%$). Zeroing uncertainty, evidence conflict, or the generic prior leaves F1 unchanged at $85.55\%$. We therefore treat the multi-factor score as a practical organizer rather than as independently validated factor importances.

## C.3 Efficiency

Table 17 compares evidence-context length for all main-table methods under the shared CLIP protocol. Tokens are a whitespace proxy over serialized reference_reports; $\overline{K}$ is the mean number of retained reports/units. MedDr and MR-RAG (Top-4) inflate context by 3–4$\times$ relative to Report-level RAG. RULE keeps a short structured budget (IU 37.6 tokens at $\overline{K}$=3.2). REO-RAG is not a longer-context method: on IU-Xray it uses *fewer* tokens than Report-level RAG (31.5 vs. 42.3) despite $\overline{K}$=2.8, because retained units are short sentences rather than whole reports. On MIMIC-CXR / PadChest-GR, REO-RAG retains more units than Report-level RAG, but still far fewer than Top-4 full-report baselines (MIMIC 112.9 vs. MedDr 309.6 / MR-RAG 310.1). MMed-RAG uses the same Top-1 full-report evidence as Report-level RAG, so their token budgets match. Organization itself is CPU-only and cheap (rebuilding the IU holdout takes 20.7 s on one CPU core); almost all runtime is Med-LVLM generation. Indicative holdout wall-clock logs for Report-level RAG vs. REO-RAG are 28.9/10.9 min (IU), 50.1/24.1 min (MIMIC), and 7.1/3.9 min (PadChest-GR); these runs were not a locked same-GPU microbenchmark, so we treat tokens as the primary cross-method efficiency signal.

# D Limitations

The study is limited to chest X-ray binary finding VQA; open-ended reasoning, multi-finding questions, and other modalities are left for future work. Evidence decomposition relies on a controlled clinical vocabulary and assertion rules, which may require institution-specific adaptation and can

miss rare phrasings. Risk weights are fixed priors with dataset-specific thresholds, not learned end-to-end; calibration leave-one-out, weight swaps, and $\tau$ sweeps (Appendix C.2) show that individual factors and nearby thresholds should not be over-interpreted as independently necessary. Organization cannot invent findings absent from the retrieval pool, and residual polarity conflicts can still mislead generation (see the qualitative cases in the main paper). Main comparisons couple evidence organization with preference adapters; matched generator controls (base vs. AC-DPO) and same-adapter representation ablations reduce—but do not eliminate—this confound across every baseline variant. We evaluate under a shared CLIP index so that stronger specialized retrievers (e.g., FactMM) are compared as report-level systems rather than as private-index competitors; FactMM+REO-RAG and similar hybrids remain future work. We evaluate a single Med-LVLM backbone (LLaVA-Med-1.5), and automatic Acc/F1 metrics do not replace radiologist review of clinical correctness.