RAV: RETRIEVAL-AUGMENTED TRAINING OF LLM JUDGES FOR CHEST X-RAY REPORT VERIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

With the emergence of large-scale vision language foundation models (VLM), radiology report generation has become a popular application area with realistic-looking reports being now generated. However, hallucinations are still prevalent in these reports making a verification step necessary before such reports could be viewed by clinicians. Recently, chain-of-thought (COT) reasoning vision-language judge models are emerging that can do reference-free evaluation but they are prone to self-preference bias and prefer hallucinated plausibility to clinical factuality when no objective ground truth is present. In this paper, we take a new approach to fact-checking by training such judge models to exploit the consensus present in prior radiology reports of similar patient images. Specifically, a new loss function is designed to capture the consensus of findings present in similar patient reports while allowing for patient-specific individuality and sparsity in the reporting of findings. We show through extensive studies across datasets, judge models and report generation models that such retrieval-based consensus-trained judge models improve the overall quality of automated reports by 8-17% as measured through comparison with ground truth reports on the reported findings in evaluation settings.

1 INTRODUCTION

With the rise of generative vision-language models in medicine, an important application area is automated radiology reporting providing critical preliminary reads. Such reports can be easily incorporated into structured reporting formats and presented to radiologists in their PACS systems for sign-offs, thus expediting the clinical workflow Syeda-Mahmood et al. (2020). However, for this to work reliably, the reports have to be factually correct in the identification and description of findings depicted in the image including their core finding type, laterality, location, severity, etc. This has been challenging for current radiology report generation (RRG) models with hallucinations, missed findingsBannur et al. (2024); Chen et al. (2024); Myronenko et al. (2025); Mahmood et al. (2025b) and incorrect description of findings altering presence to absence, introducing laterality errors, etc. Thus, there is a need for a verification layer that can verify the descriptions of findings at important checkpoints before presenting to clinicians in their workflows.

A common approach to do verification is to re-use superior large language models (LLMs) as judge models with chain of thought reasoning Gu. et al. (2026). The objective there is to evaluate the accuracy of the current LLM model using ground truth reports as reference. However, during report generation for a patient image in a clinical workflow, clearly no ground truth is available. Such judge models can still work in a reference free mode using reinforcement learning approaches but are known to be prone to self-preference bias preferring hallucinated plausibility over clinical factuality Ostmeier et al. (2024); Huang et al. (2024); Myronenko et al. (2025).

While a ground truth report is not available for the current patient image, PACS systems store past reports and DICOM images of other patients which could show similar pathology. If these systems could be queried with the current patient image to retrieve visually similar prior studies, then their associated radiology reports could provide an image-conditioned finding distribution to inform the reasoning of an LLM judge.

This is the key idea explored in this paper. Specifically, *we present a new retrieval-augmented approach for inference-time verification by training judge models using a combination of chain-of-*

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.

thought reasoning with consensus learning from retrieved reports of similar patients based on their imaging. We show that with newer domain-specific image encoders that capture clinical pathologies, visually similar images also point to clinically similar pathology evidence in the associated reports.

We demonstrate this idea in the domain of chest radiology through extensive studies. Specifically, we show that such a judge model can spot not only errors in described findings but also omissions, while remaining explainable throughout the process. Further, the method is agnostic to the origin of the reports and can work with various radiology report generation (RRG) models for chest X-rays.

Clinical decision support methods use image similarity to infer clinically similar patients Silva et al. (2022); T. Syeda-Mahmood & Hashmi (2007). Retrieval-augmented generation (RAG) approaches also rely on retrieved reports to improve generative models Lewis et al. (2020). However, to our knowledge, prior work has not explored retrieval-augmented consensus modeling as evidence for fine-tuning LLM judges for inference-time radiology report verification. Our primary contributions, therefore, are:

1. A retrieval-augmented verification (RAV) model trained on consensus reasoning drawn from prior radiology reports of similar patients as detected through their radiology imaging.
2. A new consensus loss function taking into account the need to balance reporting of common findings with patient-specific findings and maintaining sparsity in reporting.
3. A new chain-of-thought prompt that guides the learning by supplying similar reports associated with matching images in an image-to-image retrieval.

2 Related Work

Our work connects inference-time radiology fact-checking, vision-language judge models, patient-similarity retrieval, and retrieval-augmented generation.

Radiology fact-checking and judge models. Inference-time fact-checking models have been developed to identify errors in generated reports without requiring a ground-truth report at deployment time Mahmood et al. (2023; 2025b). More general vision-language judge models use CoT reasoning to evaluate generated text Zheng et al. (2023); Liu et al. (2023); Saha et al. (2025); Lee et al. (2024), including radiology reports Bologna et al. (2026); Ostmeier et al. (2024); Huang et al. (2024). Reference-based judges require a ground-truth report, whereas reference-free judges may exhibit self-preference and plausibility biases Myronenko et al. (2025); Chen et al. (2024). RAV addresses this limitation by supplying external evidence from prior imaging studies.

Patient retrieval and retrieval-augmented generation. Retrieval of similar patients is established in clinical decision support T. Syeda-Mahmood & Hashmi (2007). Retrieval-augmented generation (RAG) similarly grounds language-model generation in retrieved evidence Lewis et al. (2020); Shuster et al. (2021); Gao et al. (2024). Medical report-generation systems have incorporated image-to-text retrieval Jeong et al. (2023), text-to-text retrieval Sun et al. (2024), and visually similar images with associated reports Chu et al. (2025). These methods primarily use retrieval to improve generation. In contrast, RAV uses a visually coherent retrieval neighborhood to estimate an image-conditioned distribution of findings and uses that distribution as noisy evidence for verifying an existing candidate report. While RAG needs diversified document set for generation, RAV needs the retrieved reports be self-similar for consensus generation and reasoning, putting different requirements on retrieval.

3 Method

Given a chest X-ray and a candidate report produced by an RRG model, RAV determines whether reported findings are supported by the image and identifies omitted or inconsistent findings. It retrieves visually similar prior studies, aggregates their reported findings into an image-conditioned distribution, and fine-tunes a vision-language judge to reason over the query image, candidate report, and retrieved evidence. It does so while accounting for spurious, missing, and patient-specific individual findings. The framework has three components: (a) image-guided retrieval and consen-

sus construction,(b) a consensus-aware verification loss, and (c) generator-agnostic training using synthetic candidate reports and COT reasoning.

3.1 IMAGE-GUIDED CONSENSUS GENERATION

For a query chest X-ray, we use RAD-DINO Pérez-García et al. (2024) to compute an image embedding and retrieve the K nearest studies by cosine similarity. RAD-DINO is trained on chest X-ray images and captures texture patterns, nodule shapes, and subtle lung variations that are crucial for identifying localized pathologies. Findings extracted from reports associated with similar imaging are aggregated into an empirical frequency distribution. We use $K = 10$ in the main experiments. The distribution is treated as noisy evidence rather than ground truth so that recurring findings can support a candidate, while the judge can reject unsupported findings or identify findings absent from the retrieved cohort. Retrieval validation and sensitivity to K are reported in Section 4.1.

3.2 CONSENSUS-BASED VERIFICATION LOSS

The judge jointly considers the query image, candidate report, and retrieved finding distribution. Because retrieved cases may contain both relevant and spurious findings, training combines the standard language-generation loss with a consensus loss that encourages agreement with supported findings while preserving query-specific findings. Specifically, we combine two complementary sources of supervision: the retrieval-derived finding distribution, which provides an image-conditioned consensus signal, and the query-specific ground-truth findings, which allow the model to retain findings that may not be represented among the retrieved studies. The consensus loss contains three components:

1. **Consensus frequency:** encourages the prediction to reflect findings that recur among retrieved reports and discourages findings with little or no retrieval support.
2. **Sparsity:** prevents the verifier from simply reproducing all findings observed in the retrieved cohort.
3. **Individuality:** uses query-specific ground-truth supervision to preserve findings that may not be represented among the retrieved studies.

The total consensus loss is

$$L_{\text{consensus}} = \lambda_1 L_{\text{freq}} + \lambda_2 L_{\text{sparsity}} + \lambda_3 L_{\text{individuality}}. \tag{1}$$

where the coefficients $\lambda_{1,2,3}$ are learned during training to ontrol the relative contribution of the three loss components..

Frequency-weighted consensus loss. Let N denote the number of unique clinical findings, $p_i \in [0, 1]$ the frequency of finding i among the retrieved reports, and $\hat{y}_i$ the predicted probability that finding i is present in the candidate report. L_{freq} uses retrieval frequency as a soft target for the model prediction, with high-frequency findings being encouraged while absent findings are discouraged, and intermediate frequencies provide graded consensus supervision.

$$L_{\text{freq}} = -\frac{1}{N} \sum_{i=1}^{N} \left[p_i \log(\hat{y}_i) + (1 - p_i) \log(1 - \hat{y}_i) \right]. \tag{2}$$

Expected-length sparsity constraint. Because the retrieved cohort may contain findings that are irrelevant to the query, directly matching every retrieved finding could produce overly dense predictions. We therefore constrain the expected number of predicted findings to the expected finding count in the retrieved cohort. Let$K_f = \sum_{i=1}^{N} p_i$ denote this expected count. We define

$$L_{\text{sparsity}} = \left(\sum_{i=1}^{N} \hat{y}_i - K_f \right)^2. \tag{3}$$

This term encourages an appropriate finding density without requiring the verifier to reproduce every finding occurring among the retrieved studies.

Individuality constraint. The retrieval distribution cannot capture every finding specific to the query patient. A finding may therefore be important even when it is absent or uncommon among retrieved studies. To preserve such findings, we introduce a second BCE term whose target is the query-specific ground-truth finding label y_i^{gt} rather than the retrieval frequency:

$$L_{\text{individuality}} = \sum_{i=1}^{N} (1 - p_i), \text{BCE}(\hat{y}_i, y_i^{gt}). \tag{4}$$

The factor $(1 - p_i)$ gives greater weight to query-specific ground-truth supervision when retrieval provides little support, preserving patient-specific findings that may be absent from the retrieved cohort.

3.3 Training the verification model

We use Qwen3-VL-32B as the pretrained vision-language judge model. The model is fine-tuned using chain-of-thought reasoning together with the consensus objective described above.

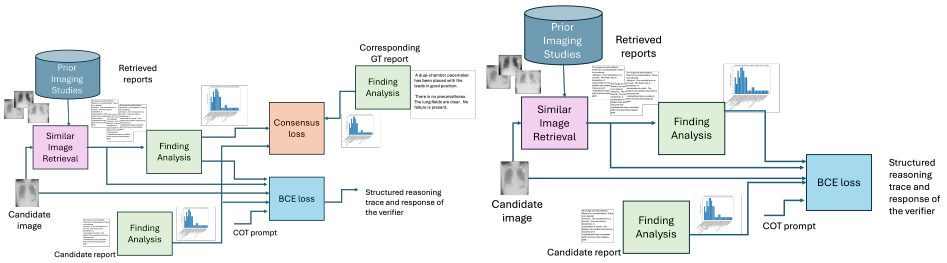
Generator-agnostic synthetic candidate reports. A central design goal is for the verifier to remain agnostic to the RRG model used by a hospital. We therefore train it using synthetic candidate reports that span a broad, model-independent space of possible factual errors rather than reproducing the error profile of a particular RRG system.

We leverage the synthetic perturbation space underlying the RADCHECK dataset Mahmood (2025). For each ground-truth report, RADCHECK enumerates compatible finding perturbations, including alternative finding configurations that can be introduced without creating explicit contradictions. Starting from the original ground truth report, we retain its report structure while selectively replacing, removing, or flipping compatible findings drawn from this perturbation space. We sample these perturbations to produce candidates with the expected finding density in automatically generated reports (12-14 findings). The resulting candidates provide supervision for missing, spurious, and inconsistent findings, while remaining independent of any particular RRG model. Further details of the perturbation procedure and its comparison with real RRG outputs are provided in Appendix A.

Pre-processing. We use FFL patterns Syeda-Mahmood et al. (2020) to represent fine-grained findings and RAD-DINO embeddings with a FAISS index for image retrieval. The constructed database of images, reports and their extracted findings are reused in training as well as inference. Additional implementation details are provided in Appendix B.

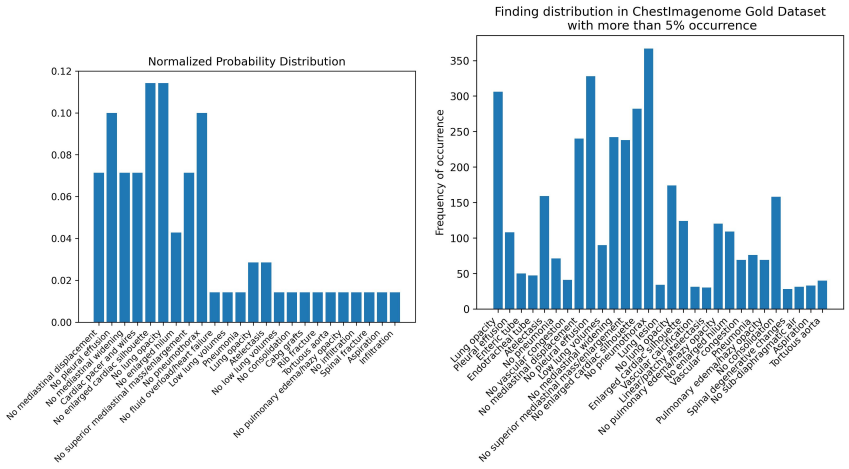
Chain-of-thought reasoning: In addition to the supervision using the verification loss, we incorporated the retrieved topK reports directly in COT reasoning. We designed a chain-of-thought prompt to elicit structured comparative reasoning. For a given image I, candidate generated report C, and retrieved report set R, and their finding probability distributions p_c and p_R, the verifier model generates a rationale Z prior to the final verdict using a prompt template T. The prompt is structured to verify in 5 phases: (1) independent image analysis to uncover findings based on the image, (2) identification of findings that appear consistently across the retrieved reports to form a retrieval consensus, (3) per-finding verification of the candidate report against Phase 1 and Phase 2 to assign verdicts (SUPPORTED, HALLUCINATION, LOCATION_ERROR, SEVERITY_ERROR, NEGATION_ERROR, or UNCERTAIN), (4) detection of omissions where findings present in Phase 1 or Phase 2 are missing from the candidate report, and (5) production of an aggregated correctness verdict (CORRECT, PARTIALLY_CORRECT, or INCORRECT) with an associated confidence. The full template structure is provided in Appendix G.

Training workflow The overall training architecture is shown in Figure 1a. For each training example, the verifier receives the query chest X-ray, a synthetic candidate report, and the retrieved reports associated with the top-K visually similar images. The retrieved reports are used to construct the query-specific finding distribution, from which the consensus-loss terms are computed. The ground-truth report provides supervision for the language-generation objective and the individuality term.



(a) (b)

Figure 1: Training and inference architecture of retrieval-augmented verification models. The core judge model can be replaced with suitable frontier models to create various versions of RAV models.



(a) (b)

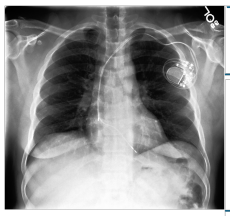
Figure 2: (a) Probability distribution of findings in the top-10 retrieved reports based on image similarity for the query image in Figure 3 . (b) Incidence distribution of the top 5% occurring findings in the Chest Imagenome Gold dataset.

Inference At inference time, the trained verifier is applied to a report generated by an RRG/VLM system as shown in Figure 1b. Given the query image and candidate report, the system retrieves the top-K visually similar prior studies and constructs their finding distribution. The verifier then jointly reasons over the image, candidate report, and retrieved evidence without access to the query's ground-truth report. The consensus used for training is no longer needed during the text generation and the usual BCE loss applies. The final RAV output is a structured list of verified and omitted findings in JSONL format. For evaluation, these outputs are post-processed into the corresponding finding-level predictions.

4 Experimental Results

We evaluate the proposed retrieval-augmented verification (RAV) framework in two stages. First, we validate the assumptions underlying image-guided consensus: (a) whether visually retrieved cases contain finding-level information beyond corpus prevalence, (b) whether the resulting consensus is query-specific, and (c) whether rare findings receive meaningful retrieval support in this process, which are frequently missed by RRG models. Second, we evaluate whether this image-conditioned verification improves the quality of automatically generated radiology reports across datasets, report generators, and judge architectures.

Datasets. We train on Chest ImaGenome Silver and evaluate on Chest ImaGenome Gold (CGold) Wu et al. (2021), MS-CXR Johnson et al. (2019), ChestXray8 Wang et al. (2017), and



Ground truth findings	RAV-COT Model	RAV-RAGCOT Model	RAV-Qwen3-VL Model
No pleural effusion No pneumothorax Cardiac pacer and wires No Pulmonary edema/hazy opacity No Pneumonia	No Pleural effusion No Pneumothorax Cardiac pacer and wires No Consolidation	No pleural effusion No pneumothorax Cardiac pacer and wires No consolidation No rib fracture	No pleural effusion No pneumothorax Cardiac pacer and wires No consolidation No Pneumonia

GT report

As compared to the previous radiograph, the patient has received a left pectoral pacemaker. The leads project over the right atrium and right ventricle respectively. No pleural effusions. No pneumonia, no pulmonary edema. No pneumothorax.

VLM report

Cardiothoracic ratio appears within normal limits, with no cardiomegaly. Costophrenic angles are sharp and well-defined, with no blunting or meniscus sign. Hilar structures are symmetric and not enlarged; mediastinum is of normal width without shift or mass effect. Lung apices and peripheral lung fields show normal visceral pleural lines without any radiolucent areas or lung edge displacement. Lung fields are clear without areas of increased opacity or air bronchograms. Pacemaker leads are clearly visible extending from the left side of the chest into the right cardiac chambers, consistent with a left-sided pacemaker system.

Figure 3: Illustration of fact-checking by the RAV model in comparison with other LLM-as-a-judge models.

Table 1: Details of the datasets used in the experiments.

Dataset	Patients	Images	Labels	Regions	Subset
RadCheck Mahmood (2025)	44,133/12,538	243,311	49	922,295	5,000
Chest ImaGenome Gold Wu et al. (2021)	288/69	461	35	5,477	325
MS-CXR Johnson et al. (2019)	478/114	925	8	2,254	325
ChestXray8 Wang et al. (2017)	457/109	880	8	1,571	325
VinDr-CXR Nguyen et al. (2022)	9,450/2,250	15,000	23	69,052	325

VinDr-CXR Nguyen et al. (2022) . CGold is additionally used for retrieval validation because it provides clinician-verified findings and image/report annotations. Dataset statistics and evaluation subsets are summarized in Table 1.

Judge architectures. We evaluate three CoT judge architectures: Qwen3-VL-32B, CheXagent-8B Chen et al. (2024), and NV-Reason-CXR-3B Myronenko et al. (2025), denoted by RAV-Qwen3-VL, RAV-CheXagent, and RAV-NVReason respectively. Qwen3-VL-32B is the primary judge. Models are fine-tuned with AdamW for three epochs using cosine learning-rate decay. Qwen3-VL uses 4-bit QLoRA, while the smaller models use BF16 full-parameter training. Full optimization and hardware details are provided in Appendix B.

Report generators. To test generator agnosticism, we evaluate candidate reports from NV-Reason-CXR Myronenko et al. (2025), XrayGPT, CheXagent Chen et al. (2024), MedGemma-27B Sellergren et al. (2025), and MAIRA-2 Bannur et al. (2024) as representative SOTA generators.

Evaluation measures. Our primary metric is FFL F1, which evaluates fine-grained finding patterns including presence, laterality, location, and severity as described in Mahmood et al. (2025a). We additionally evaluate RadGraph F1 and CheXbert F1 to test robustness to the choice of evaluation representation.

4.1 Validating Image-Guided Consensus

RAV relies on visual similarity measures to indicate clinical similarity. Further it assumes that visually similar studies provide a query-conditioned finding distribution that contains information beyond corpus-wide prevalence. We test these assumptions on CGold dataset.

Image-guided retrieval contains finding-level information. Using query images from the VinDrXr test dataset against the CGGold images, we compare findings associated with each query image against those in its RAD-DINO nearest neighbors. Table 2 compares the overlap between

Table 2: Finding overlap between the query image and its top-K retrieved cases versus a prevalence-matched null. All" includes present and absent findings, present-only" evaluates only findings annotated as present.

	All findings			Present findings only		
K	Observed	Null	Δ	Observed	Null	Δ
1	0.419	0.326	+0.093	0.189	0.095	+0.094
5	0.406	0.344	+0.062	0.222	0.127	+0.095
20	0.291	0.250	+0.041	0.188	0.118	+0.070
100	0.187	0.161	+0.026	0.127	0.082	+0.045

Table 3: Report quality before and after RAV-Qwen3-VL verification across datasets. Relative improvement is measured by the change in F1 and 95% confidence intervals are shown in the final row.

Model	**CGold**			**MS-CXR**			**ChestXray8**			**VinDrCXR**		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
(VLM, GT)	0.53	0.34	0.41	0.46	0.32	0.37	0.50	0.40	0.44	0.49	0.43	0.45
(RAV, GT)	0.56	0.42	0.48	0.47	0.45	0.46	0.51	0.48	0.49	0.51	0.50	0.50
Relative improvement	17.1 [15.1,18.4]%			24.3 [22.4,26.1]%			11.4 [9.3,12.5]%			11.1 [9.9,12.5]%		

query findings and retrieved findings with a prevalence-matched null. Image-based retrieval consistently outperforms the null for both all findings and present findings. At $K = 1$, present-finding F1 is 0.189 versus 0.095 for the null, at $K = 5$, it is 0.222 versus 0.127. Finding overlap also decreases with retrieval rank (Spearman $\rho = 0.11$, $n = 1.29$M query-neighbor pairs), indicating that the signal is associated with visual similarity rather than corpus prevalence.

These results support the weaker assumption required by RAV, namely, the retrieved neighborhood need not contain a clinically identical case. Its aggregate findings only need to provide useful image-conditioned evidence.

Query conditioning and rare findings. The retrieved finding distribution varies substantially across query images and differs from the global finding distribution, supporting the interpretation that the retrieved consensus is query-conditioned rather than a fixed corpus prior as shown in Figure 2. Rare findings also receive non-zero retrieval support, although with lower recall than common findings. Downstream stratification shows that RAV improves rare-finding F1 across all four evaluation datasets, with an average relative improvement of approximately 32%. Detailed retrieval, prevalence, rare-finding, and K-sensitivity analyses are provided in Appendix C.

4.2 Verification Performance

Having established that the retrieved neighborhood contains image-conditioned finding information, we next evaluate whether this information helps report verification and overall quality improvement of generated reports.

Performance across datasets. We first fix CheXagent-8B as the report generator and RAV-Qwen3-VL as the judge, and evaluate across CGold, MS-CXR, ChestXray8, and VinDr-CXR. Table 3 compares the findings extracted from the original generated reports with the findings validated by RAV. As can be seen, verification improves both precision and recall on all four datasets. The increase in precision reflects suppression of unsupported findings, while the increase in recall reflects recovery of findings omitted by the candidate report. The resulting relative F1 improvement ranges from 11.1% to 24.3% across the four datasets.

Performance across report generators. We next test whether the verifier remains effective when the candidate reports originate from different RRG models. Table 4 compares the generated and RAV-validated finding F1s across five report generators in comparison to the ground truth reports in various datasets. RAV improves the finding-level F1 for every evaluated report generator and dataset. The magnitude of improvement varies across generators, with the largest gains generally

Table 4: Report quality before and after RAV-Qwen3-VL verification across report generators and datasets.

Generator	CGold		MS-CXR		ChestXray8		VinDrCXR	
	(VLM,GT)	(RAV,GT)	(VLM,GT)	(RAV,GT)	(VLM,GT)	(RAV,GT)	(VLM,GT)	(RAV,GT)
NV-Reason-CXR	0.51	0.57	0.56	0.60	0.51	0.56	0.48	0.51
XrayGPT	0.39	0.46	0.47	0.52	0.42	0.48	0.35	0.40
CheXagent	0.41	0.48	0.37	0.46	0.44	0.49	0.45	0.50
MedGemma-27B	0.47	0.51	0.44	0.49	0.45	0.50	0.40	0.43
MAIRA-2	0.51	0.58	0.61	0.62	0.54	0.61	0.39	0.45
Relative improvement	13.8%		11.0%		12.0%		10.9%	

Table 5: Performance of different judge architectures after retrieval-augmented consensus training.

Judge	CGold		MS-CXR		ChestXray8		VinDrCXR	
	F1	Improvement	F1	Improvement	F1	Improvement	F1	Improvement
(VLM,GT) No judge	0.47		0.44		0.45		0.40	
RAV-NVReason	0.53	12.8%	0.48	9.1%	0.48	6.6%	0.47	17.5%
RAV-CheXagent	0.50	6.4%	0.50	11.1%	0.49	8.8%	0.44	10.0%
RAV-Qwen3-VL	0.51	8.5%	0.49	11.3%	0.50	11.1%	0.43	7.5%

occurring when the initial reports contain more omissions or unsupported findings. Across datasets, the relative improvement ranges from 10.9% to 13.8%.

Performance across judge architectures. We next evaluate whether consensus-based fine-tuning is specific to a particular judge architecture. Using MedGemma-generated reports as the candidate reports, we compare RAV-NVReason, RAV-CheXagent, and RAV-Qwen3-VL. Results are shown in Table 5. All three judge architectures improve the finding-level F1s relative to the corresponding unverified reports, indicating that the consensus-learning formulation can be transferred across model architectures and scales.

4.3 Ablation of Retrieval and Consensus Supervision

The improvements above could arise either from the retrieved reports themselves or from the proposed consensus-loss training. We therefore perform controlled ablations using Qwen3-VL as the judge, MedGemma-generated candidate reports, and CGold as the evaluation dataset.

We compare three settings: (1) direct CoT verification using only the image and candidate report (RAV-COTonly), (2) CoT verification with retrieved reports supplied directly in the prompt but without consensus-loss fine-tuning (RAV-RAGCOT), and (3) the complete RAV model with consensus-loss fine-tuning. The results are shown in Table 6.

Directly providing retrieved reports to the judge increases recall relative to the CoT-only setting, but also introduces additional spurious findings. In contrast, the complete RAV model improves both precision and recall, indicating that the consensus loss is important for learning how to use the retrieved distribution selectively rather than treating retrieved findings as ground truth.

We further test whether the improvement depends on the query-specific nature of the consensus. In a shuffle-control experiment, the trained RAV judge receives the consensus distribution associated with a randomly selected image rather than the distribution retrieved for the query image, while the query image, candidate report, and judge remain unchanged. The performance deteriorates substantially under this condition by as much as -73.2% as shown in the last row of Table 6 indicating that image-specific retrieval is necessary and that RAV is not merely using generic consensus priors.

4.4 Ablation of Consensus-Loss Components

Finally, we evaluate the individual components of the proposed consensus loss using $K = 10$. Table 7 compares consensus supervision alone, consensus with the sparsity constraint, and the complete loss including the individuality term. Consensus-only training produces relatively high recall but lower precision, consistent with the tendency to accept unsupported findings from the retrieved cohort. Adding the sparsity term increases precision by discouraging excessive findings, at the cost of some recall. Adding the individuality term subsequently recovers recall while retaining much of

Table 6: Ablation of retrieval and consensus supervision on CGold.

Judge	P	R	F1	FN rate	Spurious rate
(VLM,GT)	0.631	0.382	0.47	0.617	0.368
RAV-COTonly	0.302	0.283	0.29	0.716	0.697
RAV-RAGCOT	0.497	0.406	0.45	0.593	0.502
RAV	**0.644**	**0.426**	**0.51**	**0.574**	**0.356**
RAV-Shuffle-Qwen3-VL	0.124	0.130	0.126	0.866	0.787

Table 7: Ablation of the consensus-loss components across datasets.

Loss	**CGold**	**MS-CXR**	**ChestXray8**	**VinDrCXR**
Consensus only	0.38/0.52/0.439	0.33/0.57/0.418	0.34/0.54/0.410	0.35/0.59/0.440
Consensus + sparsity	0.58/0.39/0.460	0.49/0.41/0.440	0.53/0.44/0.480	0.52/0.46/0.488
Consensus + sparsity + individuality	0.56/0.42/0.480	0.47/0.45/0.460	0.51/0.48/0.490	0.51/0.50/0.500

the precision improvement. The complete loss consequently produces the highest F1 across all four datasets, supporting the intended division of roles among the consensus, sparsity, and individuality terms.

4.5 Qualitative Verification

Figure 3 illustrates the verification process for a MedGemma-generated report. The candidate report contains five findings, of which four are supported by the ground-truth annotations, while one is unsupported. An additional ground-truth finding is omitted. RAV identifies both the unsupported finding and the omission, demonstrating how the verifier can simultaneously suppress an incorrect candidate finding and recover a missing finding.

Additional qualitative examples and detailed K-sensitivity analyses are provided in Appendix C.

4.6 Discussion Limitations

The experiments support the central premise of RAV: visually retrieved cases provide query-conditioned finding information beyond corpus prevalence, and consensus-trained judges can use this information to improve both precision and recall. Importantly, retrieval is treated as noisy evidence rather than ground truth. The shuffle and ablation experiments show that query-specific retrieval and selective use of the retrieved evidence are important for the observed gains.

RAV nevertheless has several failure modes. Visually similar studies may differ in etiology, acquisition characteristics may dominate retrieval, and subtle or rare findings may have insufficiently similar reference cases. In multi-finding cases, retrieval may capture a dominant abnormality while missing secondary findings. The individuality term partially addresses these cases by preserving query-specific findings that lack retrieval support. Detailed failure examples and retrieval analyses are provided in Appendix C.

RAV assumes that visually similar studies contain sufficiently informative finding overlap, which may not hold for all pathologies or imaging conditions. Judge models can also introduce false positives or miss findings, and performance varies across judge architectures. Future work should improve retrieval robustness and develop stronger mechanisms for identifying when retrieved consensus is unreliable.

5 Conclusions

We presented retrieval-augmented verification (RAV), a framework for verifying generated radiology reports using findings from prior reports associated with visually similar images. RAV trains vision-language judges with a consensus loss that balances retrieval-supported findings, prediction sparsity, and query-specific findings. Across datasets, report generators, and judge architectures, RAV improves finding-level report quality while identifying both unsupported and omitted findings.

The results suggest that prior radiology reports can provide useful inference-time evidence when a ground-truth report for the current patient is unavailable. More broadly, retrieval-augmented consensus provides a way to incorporate clinical archive information into generative judge models without treating retrieved cases as ground truth.

REFERENCES

Shruthi Bannur et al. Maira-2: Grounded radiology report generation, 2024. URL https://arxiv.org/abs/2406.04449.

Federica Bologna, Jean-Philippe Corbeil, Matthew Wilkens, and Asma Ben Abacha. Vert: Reliable llm judges for radiology report evaluation. *arXiv preprint arXiv:2604.03376*, 2026. doi: 10.48550/arXiv.2604.03376. URL https://arxiv.org/abs/2604.03376.

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, et al. A vision-language foundation model to enhance efficiency of chest X-ray interpretation. *arXiv preprint arXiv:2401.12208*, 2024.

Yun-Wei Chu et al. Reducing hallucinations of medical multimodal large language models with visual retrieval-augmented generation. *https://arxiv.org/abs/2502.15040*, 2025.

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.

Jiawei Gu. et al. A survey on llm-as-a-judge. *Cell - The Innovation*, 7, 2026.

Alyssa Huang et al. FineRadScore: A radiology report line-by-line evaluation technique generating corrections with severity scores. *arXiv preprint arXiv:2405.20613*, 2024.

Jaehwan Jeong, Katherine Tian, Andrew Li, Sina Hartung, Fardad Behzadi, Juan Calle, Daniel Osayande, Michael Pohlen, Subathra Adithan, and Pranav Rajpurkar. Multimodal image-text matching improves retrieval-based chest X-ray report generation. In *Conference on Health, Inference, and Learning*, 2023.

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*, 6:317, 2019. doi: 10.1038/s41597-019-0322-0. URL https://doi.org/10.1038/s41597-019-0322-0.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*, 2017. URL https://arxiv.org/abs/1702.08734.

Seongyun Lee, Seungone Kim, Sue Hyun Park, Geewook Kim, and Minjoon Seo. Prometheus-Vision: Vision-language model as a judge for fine-grained evaluation. In *Findings of the Association for Computational Linguistics: ACL*, 2024.

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*, 2020.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. 2023.

R. Mahmood et al. Evaluating automated radiology report quality through fine-grained phrasal grounding of clinical findings. In *Proc. IEEE Int. Symp. Biomed. Imaging*, pp. 1–5, 2025a. doi: 10.1109/ISBI60581.2025.10981120.

Razi Mahmood. Radcheck: A large-scale dataset for radiology report fact-checking. https://huggingface.co/datasets/razi-mahmood/RadCheck, 2025. HuggingFace dataset.

Razi Mahmood, Ge Wang, Mannudeep Kalra, and Pingkun Yan. Fact-checking of ai-generated reports. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 14349 LNCS:214–223, 7 2023. ISSN 16113349.

Razi Mahmood, Diego Machado-Reyes, Joy Wu, Parisa Kaviani, Ken C. L. Wong, Niharika D'Souza, Mannudeep Kalra, Ge Wang, Pingkun Yan, and Tanveer Syeda-Mahmood. Phrase-grounded fact-checking for automatically generated chest X-ray reports. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2025b.

Andriy Myronenko et al. Reasoning visual language model for chest x-ray analysis. *https://arxiv.org/abs/2510.23968*, 2025.

Ha Q. Nguyen et al. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. *Scientific Data 2022 9:1*, 9:1–7, 7 2022. ISSN 2052-4463. URL https://www.nature.com/articles/s41597-022-01498-w.

Sophie Ostmeier et al. GREEN: Generative radiology report evaluation and error notation. In *Findings of EMNLP*, 2024.

Fernando Pérez-García et al. Exploring scalable medical image encoders beyond text supervision. *Nature Machine Intelligence*, 2024.

Swarnadeep Saha et al. Learning to plan and reason for evaluation with thinking-LLM-as-a-judge. *arXiv preprint arXiv:2501.18099*, 2025.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cián Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025. URL https://arxiv.org/abs/2507.05201.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. In *Findings of EMNLP*, 2021.

Wilson Silva, Tiago Gonçalves, Kirsi Härmä, Erich Schröder, Verena Carola Obmann, María Cecilia Barroso, Alexander Poellinger, Mauricio Reyes, and Jaime S. Cardoso. Computer-aided diagnosis through medical image retrieval in radiology. *Scientific Reports*, 12:20732, 2022. doi: 10.1038/s41598-022-25027-2.

Liwen Sun et al. Fact-aware multimodal retrieval augmentation for accurate medical radiology report generation. *arXiv preprint arXiv:2407.15268*, 2024.

Tanveer Syeda-Mahmood, Ken C L Wong, Yaniv Gur, Joy T Wu, Ashutosh Jadhav, Satyananda Kashyap, Alexandros Karargyris, Anup Pillai, Arjun Sharma, Ali Bin Syed, Orest Boyko, and Mehdi Moradi. Chest x-ray report generation through fine-grained label learning. In *MICCAI-2020*, 2020.

D. Beymer A. Amir M. Richmond T. Syeda-Mahmood, F. Wang and S.N. Hashmi. Aalim: Multimodal mining for cardiac decision support. In *Computers in Cardiology*, pp. 209–212, 2007. URL https://doi.org/10.1109/CIC.2007.4745458.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3462–3471, 2017.

Joy T. Wu, Nkechinyere N. Agu, Ismini Lourentzou, Arjun Sharma, Joseph A. Paguio, Jasper S. Yao, Edward C. Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, Leo A. Celi, and Mehdi Moradi. Chest imagenome dataset for clinical reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 8328–8340, 2021.

Lianmin Zheng et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

In this appendix we provide more experimental evidence to support the main paper. Some of the tables used in the main paper have been used to reproduce here for co-location purposes during the discussion.

A Synthetic Candidate Report Generation

The synthetic candidate reports provide generator-agnostic supervision for factual verification errors. Rather than modeling the error profile of a particular radiology report generation (RRG) system, we sample from a broad space of finding-level perturbations of ground-truth reports.

We build on the synthetic candidates underlying the RADCHECK dataset Mahmood (2025). For each ground-truth report, findings are converted into normalized FFL patterns, and compatible findings can be retained, removed, replaced, or flipped while avoiding contradictions between mutually exclusive states. Corresponding sentences are selected from available reports so that candidates retain natural report language and local context.

For training, we sample from this precomputed perturbation space and control the finding density to approximately 12-14 findings, representative of generated reports. The resulting candidates contain missing, spurious, and inconsistent findings without depending on a particular RRG model.

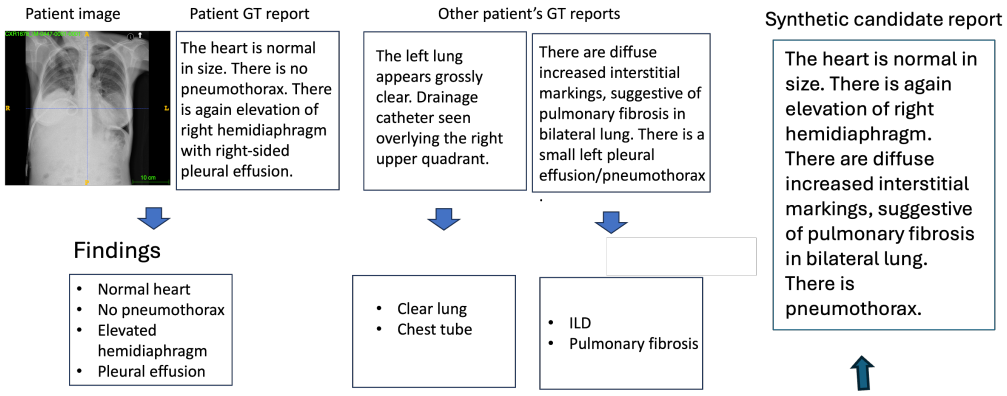


Figure 4: Synthetic candidate report generation process.

The original RADCHECK dataset contains more than 27 million report/perturbation combinations generated from the Chest ImaGenome Silver data Wu et al. (2021). We sample 5,000 candidates for fine-tuning. Increasing the synthetic training set by a factor of 10 produced only a modest additional improvement of approximately 3-5%.

To assess whether the synthetic errors resemble those in actual RRG outputs, we compared their finding-frequency distribution on CGold with those produced by five RRG models. The average Jensen-Shannon divergence between the synthetic distribution, P_{syn}, and the five RRG distributions, P_{RRG_i}, was 0.12. We additionally compared deviations from ground truth $D_{syn} = P_{syn} - P_{GT}, D_{RRGi} = P_{RRGi} - P_{GT}$. The mean canonical correlation coefficient between the synthetic deviation vector and those of the five RRG systems was 0.912. Thus, although the synthetic candidates do not reproduce the complete behavior or style of any particular RRG system, their finding-level deviations are similar to those observed in generated reports.

Figure 4 illustrates the construction of a candidate. Findings may be retained, removed, or replaced by compatible findings from the RADCHECK perturbation space, producing controlled factual errors while preserving realistic report language.

B Finding Extraction and Retrieval Implementation

Radiology reports are converted into fine-grained finding representations using the FFL representation described in Syeda-Mahmood et al. (2020). This representation normalizes findings independently of their exact wording and is used both for constructing retrieval-based finding distributions and for evaluating retrieval overlap.

Each chest X-ray is encoded using RAD-DINO Pérez-García et al. (2024). For a query image, its embedding is compared with indexed reference images using cosine similarity. We use a FAISS Johnson et al. (2017) inverted-file (IVF) index for efficient nearest-neighbor retrieval. Each indexed image is associated with its report and extracted FFL findings.

At inference time, retrieved reports provide contextual evidence for the judge, while their extracted findings are aggregated into the query-specific finding-frequency distribution. The ground-truth report of the query image is used only during training and evaluation and is unavailable to the judge at inference.

Implementation details. Qwen3-VL-32B is fine-tuned with 4-bit QLoRA, with adapters applied to all linear layers ($r = 32$, $\alpha = 64$), a micro-batch size of 1, 32 gradient-accumulation steps, and a sequence length of 4096 tokens. NV-Reason-CXR-3B and CheXagent-8B are fine-tuned in BF16 using full-parameter training with a micro-batch size of 8. All models are trained for three epochs using AdamW with cosine learning-rate decay and 5% warmup. Peak learning rates are 2×10^{-5} for Qwen3-VL-32B and 5×10^{-5} for the smaller models. Checkpoints are evaluated every 100 iterations and selected using validation loss. Experiments were conducted on an Apple M4 Max workstation with 128 GB unified memory using the Metal-accelerated MLX framework.

Table 8: Finding overlap between the query image and its top-K retrieved cases using RAD-DINO.

Top-K	Precision	Recall	F1
1	0.548	0.544	0.524
3	0.446	0.780	0.550
5	0.397	0.852	0.524
10	0.329	0.914	0.471
20	0.262	0.942	0.401

Table 9: Finding overlap between the query image and retrieved cases compared with a prevalence-matched null. All" includes present and absent findings, present only" evaluates findings annotated as present.

	All findings			Present findings only		
K	Observed	Null	Δ	Observed	Null	Δ
1	0.419	0.326	+0.093	0.189	0.095	+0.094
5	0.406	0.344	+0.062	0.222	0.127	+0.095
20	0.291	0.250	+0.041	0.188	0.118	+0.070
100	0.187	0.161	+0.026	0.127	0.082	+0.045

C Retrieval and Consensus Analysis

The main paper establishes that image-guided retrieval provides finding-level information beyond corpus prevalence and summarizes its effect on downstream verification. This appendix provides the detailed retrieval, rare-finding, and retrieval-size analyses underlying those results.

C.1 Retrieval Quality and Finding Overlap

The RAV retrieval mechanism is intended to provide an image-conditioned distribution rather than identify a clinically identical case. We therefore evaluate whether findings in the retrieved neighborhood overlap with those of the query image and whether this overlap exceeds what would be expected from corpus-wide prevalence.

Using CGold, we compare the query finding set with the pooled findings from the top-K RAD-DINO retrievals. Table 8 reports precision, recall, and F1.

Increasing K increases recall because a larger neighborhood covers more query findings, while precision decreases as unrelated findings enter the pooled distribution. At $K = 10$, the retrieved cohort covers 91.4% of query findings, motivating its use as a high-recall consensus neighborhood rather than as a set of ground-truth cases.

We reproduce the comparison of image-based retrieval with a prevalence-matched null again in Table 9. For each K, the null samples groups of K studies with the same corpus-level finding prevalence but without image-similarity information.

Image-based retrieval consistently produces greater overlap than the prevalence-matched null. The difference is particularly evident for present findings: at $K = 1$, F1 is 0.189 versus 0.095, and at $K = 5$, 0.222 versus 0.127. The advantage decreases as K grows because additional neighbors introduce increasingly unrelated findings.

We also evaluated finding overlap for individual retrieved neighbors as a function of retrieval rank. Overlap decreases as similarity rank worsens (Spearman $\rho = 0.11$, $n = 1.29$M query-neighbor pairs), providing complementary evidence that the finding signal is associated with visual similarity.

C.2 Query-Conditioned Consensus

To determine whether retrieval simply reproduces global finding prevalence, we compared distributions obtained from VinDr-CXR query images against the independent CGold retrieval corpus. The distribution of retrieved findings varies with the query and differs from the global CGold distribution, indicating that retrieval provides a query-conditioned rather than fixed population-level prior. Figure 2 illustrates the distinction between the global and query-conditioned distributions. The latter reflects the visual neighborhood of the query and can therefore emphasize findings that are not among the most prevalent findings in the retrieval corpus.

Table 10: Positive prevalence of selected rare findings in CGold.

Finding	Prevalence	Finding	Prevalence
Pneumothorax	2.01%	Mediastinal widening	2.01%
Enlarged hilum	4.22%	Fracture	7.03%
Mass/nodule	6.63%	Rib fracture	4.62%
Elevated hemidiaphragm	4.02%	Hyperaeration	3.82%
Hernia	2.81%	Lobar/segmental collapse	2.61%
Calcified nodule	1.81%	Scoliosis	1.81%
Increased reticular markings/ILD pattern	1.00%	Interstitial lung disease	0.80%

Table 11: Retrieval performance for clinician-defined rare findings at different values of K.

Top-K	Precision	Recall	F1
1	0.511	0.514	0.469
3	0.465	0.702	0.523
5	0.420	0.748	0.505
10	0.351	0.782	0.453

C.3 Retrieval Failure Modes

Visual similarity does not imply identical pathology. Retrieved studies can differ in etiology despite similar appearances, and acquisition characteristics such as portable AP imaging, ICU support devices, or positioning can influence similarity. Subtle localized abnormalities may also have weak representation among retrieved cases. In multi-finding cases, retrieval may capture the dominant abnormality while missing secondary findings. These limitations motivate treating retrieval as a noisy image-conditioned prior rather than a label source. The consensus loss is designed to prevent the judge from simply copying all findings appearing in retrieved reports.

C.4 Rare-Finding Retrieval

We evaluated whether retrieval provides useful evidence for low-prevalence findings. In CGold, selected clinician-defined rare findings occur in approximately 1-7% of studies as shown in Table 10. Table 11 reports retrieval performance for this subset.

Rare-finding recall is lower than aggregate retrieval recall but remains substantial at $K = 10$. Thus, rare findings can receive consensus support when visually similar cases contain the same finding, while patient-specific rare findings that are absent from the retrieval cohort require the query-specific evidence captured by the individuality term. We additionally stratified verification performance by the same rare-finding subset. The mean relative F1 improvement is 32.02% as shown in Table 12. This analysis considers positive rare-pathology prevalence. Negative mentions such as "no pneumothorax" are excluded from this definition.

C.5 Sensitivity to Retrieval Size

We evaluated RAV with $K \in 1, 3, 5, 10, 20$. Table 13 shows the resulting precision, recall, and F1. Small retrieval sets provide relatively precise evidence but limited coverage, whereas larger sets increase recall while introducing additional unrelated findings. $K = 10$ provides an intermediate operating point across datasets and is therefore used in the main experiments.

C.6 Qualitative Retrieval Example

Figure 5 illustrates the retrieval and consensus construction process. Column (a) shows the query chest X-ray and its ground-truth findings. Columns (b)-(e) show four highly ranked visually similar studies retrieved using RAD-DINO Pérez-García et al. (2024) and their associated findings. The retrieved cases share several findings with the query but are not identical. Some findings occur only in a subset of retrieved reports, while others are absent from the query ground truth. The example therefore illustrates why the retrieved distribution is used as a noisy prior rather than a substitute for query-specific reasoning.

D Retrieval as Noisy Evidence

The downstream results should not be interpreted as showing that retrieved findings are accepted automatically. A verifier that simply imported all retrieved findings could increase recall while introducing unsupported

Table 12: Verification performance on rare findings across evaluation datasets.

Dataset	Raw VLM P/R/F1	RAV P/R/F1	Relative F1 gain
CGold	0.43/0.21/0.28	0.46/0.38/0.42	49.9%
MS-CXR	0.32/0.24/0.27	0.38/0.39/0.38	40.7%
ChestXray8	0.41/0.29/0.34	0.45/0.34/0.39	14.7%
VinDrCXR	0.39/0.33/0.35	0.45/0.41/0.43	22.8%

Table 13: Sensitivity of RAV performance to the number of retrieved cases. Each entry reports precision/recall/F1.

Top-K	CGold	MS-CXR	ChestXray8	VinDrCXR
1	0.98/0.12/0.21	0.97/0.13/0.23	0.98/0.14/0.24	0.95/0.18/0.30
3	0.78/0.28/0.41	0.74/0.26/0.38	0.77/0.29/0.42	0.78/0.31/0.44
5	0.64/0.38/0.47	0.60/0.37/0.45	0.62/0.40/0.48	0.61/0.38/0.47
10	0.56/0.42/0.48	0.47/0.45/0.46	0.51/0.48/0.49	0.51/0.50/0.50
20	0.33/0.65/0.44	0.32/0.61/0.42	0.31/0.66/0.42	0.38/0.68/0.49

findings and reducing precision. We therefore compared three configurations: CoT verification without retrieval, CoT verification with retrieved reports supplied directly in the prompt, and the complete RAV model with consensus-loss fine-tuning. The results in Table 14 show that direct retrieval increases recall but also introduces spurious findings, whereas consensus training improves both precision and recall.

We also performed a shuffle-control experiment in which the trained RAV judge received the consensus distribution associated with a randomly selected image rather than the distribution retrieved for the query image. The query image, candidate report, prompt, and judge were unchanged. Performance deteriorated under this condition, indicating that the benefit depends on query-specific retrieval rather than generic access to finding frequencies. These results were already reported in Table 6.

Together, these experiments support interpreting retrieval as a noisy, image-conditioned source of evidence that must be selectively reconciled with the query image and candidate report.

E Ablation of Consensus-Loss Components

We isolate the roles of the three components of the consensus loss using $K = 10$ in Table 15. Each entry reports precision/recall/F1. Consensus-only training produces high recall but lower precision because unsupported findings from the retrieved cohort are more readily accepted. Adding sparsity increases precision by limiting the number of predicted findings, while adding individuality recovers recall by restoring query-specific findings that may lack retrieval support. The complete loss therefore provides the intended balance between retrieval consensus and patient-specific evidence.

F Additional Evaluation-Metric Analysis

The primary evaluation uses FFL F1 because the representation retains clinically relevant attributes including anatomy, laterality, and severity. We additionally evaluate RAV using RadGraph-F1 and CheXbert-F1. The results are shown in Table 16. Absolute scores differ because the metrics use different representations and matching criteria. The direction of improvement over the corresponding baseline remains consistent across the three evaluation schemes, indicating that the observed gains are not specific to the FFL representation.

G COT prompts

This appendix reproduces the two chain-of-thought judge prompts referenced in Section 5: **COT1** (no-retrieval mode) and **COT2** (full retrieval mode). Both share an identical system message and a five-phase reasoning structure. COT1 omits the retrieval consensus phase and the retrieved-reports input block. COT2 includes both. The output schema is identical across modes.

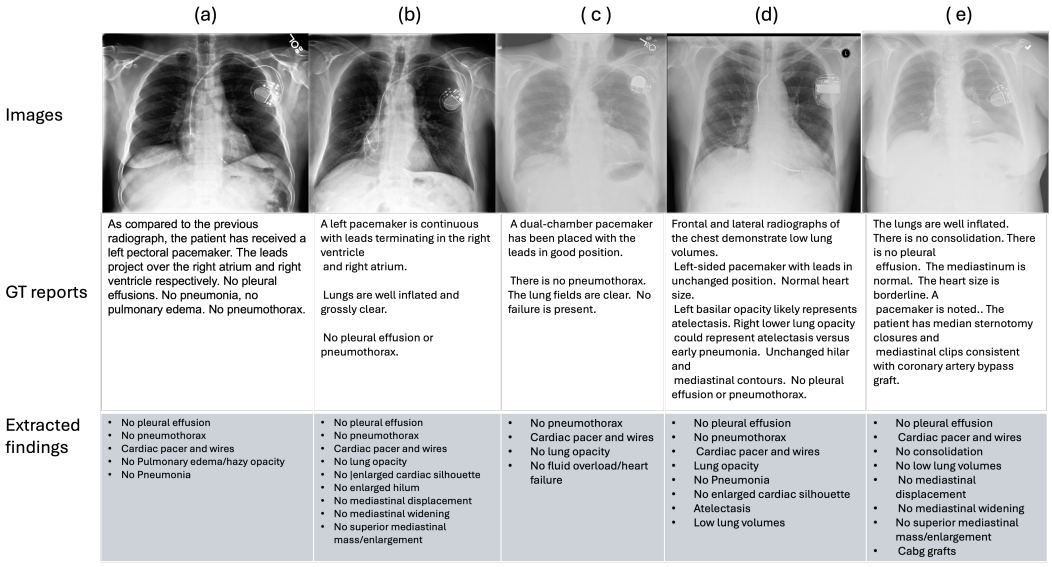


Figure 5: Illustration of consensus evident in retrieved reports based on image similarity.

Table 14: Ablation of retrieval and consensus supervision on CGold.

Judge	P	R	F1	FN rate	Spurious rate
(VLM,GT)	0.631	0.382	0.47	0.617	0.368
RAV-COTonly	0.302	0.283	0.29	0.716	0.697
RAV-RAGCOT	0.497	0.406	0.45	0.593	0.502
RAV	**0.644**	**0.426**	**0.51**	**0.574**	**0.356**

G.1 Shared System Message

You are a board-certified radiologist evaluating an automatically generated chest X-ray report. Your task is to determine whether the report accurately describes the image, identify any errors, and identify any clinically important findings the report has omitted.

You may be given up to three inputs:

1. A frontal chest X-ray image (when available).
2. A candidate report produced by an automated system. This report may contain one or more errors of the following types:
   - HALLUCINATION: a finding stated in the report that is not present in the image.
   - OMISSION: a finding present in the image that is not stated in the report.
   - LOCATION_ERROR: a finding correctly identified but assigned to the wrong anatomical location.
   - SEVERITY_ERROR: a finding correctly identified but with incorrect severity, size, or extent.
   - NEGATION_ERROR: a finding incorrectly affirmed or denied.
3. A set of K actual reports written by radiologists for chest X-ray images that are visually similar to the input image (when available). These reports belong to other patients and should be treated as evidence about what findings are commonly present in images of this visual class, not as ground truth for the input image.

Reason in phases. Complete each phase before beginning the next. Do not revise an earlier phase after seeing later inputs. Output your reasoning as a single JSON object matching the schema below. Do not include any text outside the JSON object.

G.2 COT1: No-Retrieval Prompt

Inputs. Image I and candidate report C.

User message structure.

Table 15: Ablation of the components of the consensus loss.

Loss	CGold	MS-CXR	ChestXray8	VinDrCXR
Consensus only	0.38/0.52/0.439	0.33/0.57/0.418	0.34/0.54/0.410	0.35/0.59/0.440
Consensus + sparsity	0.58/0.39/0.460	0.49/0.41/0.440	0.53/0.44/0.480	0.52/0.46/0.488
Consensus + sparsity + individuality	0.56/0.42/0.480	0.47/0.45/0.460	0.51/0.48/0.490	0.51/0.50/0.500

Table 16: RAV performance under three clinical evaluation metrics.

Dataset	RQ/FFL-F1	RadGraph-F1	CheXbert-F1
CGold	0.48	0.45	0.53
MS-CXR	0.46	0.40	0.52
ChestXray8	0.49	0.45	0.52
VinDrCXR	0.50	0.47	0.54

`[CANDIDATE REPORT]`
`<candidate report text>`
`[REASONING PHASES]`
`<phase blocks below>`

Phase 1: Independent image analysis.

Examine the chest X-ray. Without referring to the candidate report, list the radiological findings you observe. For each finding, record (a) the finding description, (b) the anatomical location, (c) the severity or extent if applicable, (d) presence or absence, and (e) your confidence on a scale of 0.0 to 1.0. Include only findings you can directly observe in the image. Record this as the field `image_findings`.

Phase 3: Per-finding verification of the candidate report.

Decompose the candidate report into atomic findings of the form (entity, location, severity, presence). For each atomic finding, determine whether it is supported by Phase 1. Assign a verdict from {SUPPORTED, HALLUCINATION, LOCATION_ERROR, SEVERITY_ERROR, NEGATION_ERROR, UNCERTAIN}. For each finding, record the verdict, a short evidence string citing Phase 1 (or "candidate-only" if Phase 1 does not contain corroborating evidence), and a confidence value between 0.0 and 1.0. Record this as the field `candidate_findings`.

Phase 4: Omission detection.

For each finding in Phase 1 that is not represented in the candidate report, record an omission with (a) the finding description, (b) the provenance source (`image` for COT1), (c) clinical severity (`minor`, `major`), and (d) confidence between 0.0 and 1.0. Major severity is reserved for findings that would change clinical management. Record this as the field `omissions`.

Phase 5: Aggregated verdict.

Produce an overall correctness label from {CORRECT, PARTIALLY_CORRECT, INCORRECT} and an overall confidence value between 0.0 and 1.0. CORRECT requires no HALLUCINATION, LOCATION_ERROR, SEVERITY_ERROR, or NEGATION_ERROR verdicts in Phase 3, and no major-severity omissions in Phase 4. INCORRECT applies when at least one major-severity error or omission is present. PARTIALLY_CORRECT applies otherwise. Record this as the field `overall`.

G.3 COT2: Full Retrieval Prompt

Inputs. Image I, candidate report C, and retrieved report set $R = \{r_1, \ldots, r_K\}$ ordered by image similarity.

System message. Identical to COT1 (Section G.1).

User message structure.

```
[CANDIDATE REPORT]
<candidate report text>
[RETRIEVED REPORTS]
The following reports describe visually similar chest
X-ray studies from different patients, ordered by visual
similarity (most similar first).
(1) <retrieved report 1>
(2) <retrieved report 2>
...
(K) <retrieved report K>
[REASONING PHASES]
<phase blocks below>
```

Phase 1: Independent image analysis. Identical to COT1.

Phase 2: Retrieval consensus.

Read the K retrieved reports. Identify findings that appear across retrieved reports, and identify the consensus expectation for images of this visual class. For each consensus finding, record (a) the finding description, (b) the number of retrieved reports that mention it, and (c) whether your independent analysis confirms this finding. Record this as the field `retrieval_consensus`.

Phase 3: Per-finding verification of the candidate report.

Decompose the candidate report into atomic findings of the form (entity, location, severity, presence). For each atomic finding, determine whether it is supported by Phase 1 and consistent with Phase 2. Assign a verdict from {SUPPORTED, HALLUCINATION, LOCATION_ERROR, SEVERITY_ERROR, NEGATION_ERROR, UNCERTAIN}. For each finding, record the verdict, a short evidence string citing Phase 1 or Phase 2, and a confidence value between 0.0 and 1.0. Record this as the field `candidate_findings`.

Phase 4: Omission detection.

For each finding in Phase 1 or Phase 2 that is not represented in the candidate report, record an omission with (a) the finding description, (b) the provenance source (`image`, `retrieval`, or `both`), (c) clinical severity (`minor`, `major`), and (d) confidence between 0.0 and 1.0. Record this as the field `omissions`.

Phase 5: Aggregated verdict. Identical to COT1.

G.4 Shared Output Schema

Both prompts require a single JSON object with the following schema. Fields not applicable to a given mode (e.g., `retrieval_consensus` under COT1) are returned as empty lists.

```
{
  "image_findings": [
    {"finding": "string", "location": "string",
     "severity": "string or null", "confidence": 0.0}
  ],
  "retrieval_consensus": [
    {"finding": "string", "support_count": 0,
     "confirmed_in_image": true}
  ],
  "candidate_findings": [
    {"finding": "string", "entity": "string",
     "location": "string", "severity": "string or null",
     "presence": "affirmed | denied",
     "verdict": "SUPPORTED | HALLUCINATION | LOCATION_ERROR
                 | SEVERITY_ERROR | NEGATION_ERROR | UNCERTAIN",
```

```
      "evidence": "string", "confidence": 0.0}
  ],
  "omissions": [
    {"finding": "string",
     "source": "image | retrieval | both",
     "severity": "minor | major",
     "confidence": 0.0}
  ],
  "overall": {
    "verdict": "CORRECT | PARTIALLY_CORRECT | INCORRECT",
    "confidence": 0.0,
    "summary": "string"
  }
}
```

H SAMPLE CHAIN-OF-THOUGHT REASONING TRACE

This appendix presents an end-to-end execution of the COT2 (full retrieval) prompt described in Appendix G.3 on the running example shown in Figures 5 and 2. The inputs are: (i) the chest X-ray identified by MIMIC-CXR `dicom_id` `3491d26b-b33be63e-f27df960-f2fd691c-2c4ce2be`, (ii) the candidate report produced by CheXagent for this image, and (iii) the top-K=5 reports retrieved by RAD-DINO over the ChestImagenome Silver corpus, ordered by cosine similarity (most similar first). The judge model is Qwen3-VL-32B operating under COT2.

Candidate report:

> Tubes, Catheters, and Support Devices: – Left-sided pacemaker with leads terminating in the right atrium and right ventricle. Lungs and Airways: – No evidence of pneumothorax. – No focal consolidation. Pleura: – No pleural effusion. Cardiovascular: – Normal heart size. Hila and Mediastinum: – Normal hilar and mediastinal contours.

Retrieved reports ($K = 5$):

> **(1)** 2. No pneumothorax. Hilar and cardiomediastinal contours are normal. Lungs are well inflated and grossly clear. No pleural effusion or pneumothorax.
>
> **(2)** The lung fields are clear. There is no pneumothorax.
>
> **(3)** Left basilar opacity likely represents atelectasis. No pleural effusion or pneumothorax. Normal heart size. Right lower lung opacity could represent atelectasis versus early pneumonia.
>
> **(4)** The mediastinum is normal. There is no consolidation. There is no pleural effusion.
>
> **(5)** Heart size is top-normal. Lungs are well aerated and essentially clear. No pleural effusion or pneumothorax is seen. Rib deformity on the right is noted, might potentially represent prior fracture. Tortuous aorta is re-demonstrated.

Judge output:

```
{
  "image_findings": [
    {"finding": "Left-sided pacemaker with leads terminating in
                 the right atrium and right ventricle",
     "location": "mediastinum", "severity": null, "confidence": 0.90},
    {"finding": "No evidence of pneumothorax",
     "location": "lungs", "severity": null, "confidence": 0.95},
    {"finding": "No focal consolidation in the lungs",
     "location": "lungs", "severity": null, "confidence": 0.90},
    {"finding": "No pleural effusion",
     "location": "pleura", "severity": null, "confidence": 0.90},
    {"finding": "Normal heart size",
     "location": "cardiomediastinum", "severity": null, \
     "confidence": 0.85},
    {"finding": "Normal hilar and mediastinal contours",
     "location": "hila and mediastinum", "severity": null, \
     "confidence": 0.90},
```

```
    {"finding": "No rib fractures or deformities",
     "location": "ribs", "severity": null, "confidence": 0.80}
  ],
  "retrieval_consensus": [
    {"finding": "No pneumothorax",
     "support_count": 5, "confirmed_in_image": true},
    {"finding": "No pleural effusion",
     "support_count": 5, "confirmed_in_image": true},
    {"finding": "Normal hilar and mediastinal contours",
     "support_count": 3, "confirmed_in_image": true},
    {"finding": "Lungs are clear or well aerated",
     "support_count": 4, "confirmed_in_image": true},
    {"finding": "Normal heart size",
     "support_count": 2, "confirmed_in_image": true}
  ],
  "candidate_findings": [
    {"finding": "Left-sided pacemaker ...",
     "entity": "pacemaker", "presence": "affirmed",
     "verdict": "SUPPORTED",
     "evidence": "Phase 1: Observed in image with high confidence",
     "confidence": 0.90},
    {"finding": "No evidence of pneumothorax",
     "entity": "pneumothorax", "presence": "denied",
     "verdict": "SUPPORTED",
     "evidence": "Phase 1: No pneumothorax observed;
                  Phase 2: Consensus across all 5 retrieved reports",
     "confidence": 0.95},
    {"finding": "No focal consolidation",
     "entity": "consolidation", "presence": "denied",
     "verdict": "SUPPORTED",
     "evidence": "Phase 1: No consolidation observed;
                  Phase 2: Consensus in 4/5 retrieved reports",
     "confidence": 0.90},
    {"finding": "No pleural effusion",
     "entity": "pleural effusion", "presence": "denied",
     "verdict": "SUPPORTED",
     "evidence": "Phase 1: No effusion observed;
                  Phase 2: Consensus across all 5 retrieved reports",
     "confidence": 0.90},
    {"finding": "Normal heart size",
     "entity": "heart size", "presence": "affirmed",
     "verdict": "SUPPORTED",
     "evidence": "Phase 1: Heart size appears normal;
                  Phase 2: Confirmed in 2/5 retrieved reports",
     "confidence": 0.85},
    {"finding": "Normal hilar and mediastinal contours",
     "entity": "hilar and mediastinal contours", "presence": "affirmed",
     "verdict": "SUPPORTED",
     "evidence": "Phase 1: Normal contours observed;
                  Phase 2: Confirmed in 3/5 retrieved reports",
     "confidence": 0.90}
  ],
  "omissions": [
    {"finding": "No rib fractures or deformities",
     "source": "image", "severity": "minor", "confidence": 0.80}
  ],
  "overall": {
    "verdict": "CORRECT",
    "confidence": 0.90,
    "summary": "All findings in the candidate report are supported
                by direct image observation and/or retrieval consensus.
                The only omission is a minor finding (no rib fractures)
                that does not impact clinical management.
                No hallucinations,
                location errors, severity errors, or negation errors
```

```
                    were detected."
        }
    }
```

I AI USE FOR THE PAPER

We use ChatGPT public domain version to edit the draft and for the purposes of shrinking to the required number of pages needed for ICLR formatting. Help was also taken from this tool for producing the latex formatting for the table of results used in appendix.