Reformulating Radiology Report Error Detection as Sentence-Level Vision-Language Learning

Dabin Min[1,2], Kyungmin Jeon[1,2], Dongheon Lee[3,4,5], Chang Min Park[3,4,6]

[1]Interdisciplinary Program in Bioengineering, Seoul National University Graduate School
[2]Integrated Major in Innovative Medical Science, Seoul National University Graduate School
[3]Dept. of Radiology, Seoul National University Hospital
[4]Seoul National University College of Medicine
[5]Interdisciplinary Program in Artificial Intelligence, Seoul National University
[6]Institute of Medical and Biological Engineering,
Seoul National University Medical Research Center

{reonaledo,kmjeon,dhlee13,morphius}@snu.ac.kr

Abstract

Accurate alignment between chest X-ray (CXR) images and radiology reports is essential for clinical safety, yet automatic radiology report error detection from paired CXRs remains challenging. Prior approaches typically compare the entire report with the image, limiting the model's ability to learn fine-grained vision-language correspondences at the level of individual findings. We reformulate report error detection as a sentence-level vision-language learning problem by decomposing reports into atomic finding statements, thereby reducing semantic entanglement and supervision noise. We propose **SenRED** (**Sen**tence-level **R**eport **E**rror **D**etector), which uses sentence-conditioned patch aggregation to associate each statement with localized visual evidence. To supervise learning, we generate sentence-level synthetic errors reflecting clinically relevant categories—false positives, false negatives, and mislocalization—along with consistent image-sentence pairs. Experiments on MIMIC-CXR and Open-i show that SenRED consistently outperforms task-specific baselines and zero-shot instruction-tuned VLMs in both sentence-level and exam-level evaluations. Radiologist-confirmed evaluations further show that these gains are maintained on clinically reviewed subsets. Comparative experiments support the benefit of sentence-level reformulation, while ablations demonstrate that sentence-conditioned patch aggregation improves fine-grained consistency checking.

1 Introduction

Radiology report errors can have substantial clinical consequences, including delayed diagnosis, inappropriate treatment, and avoidable patient harm (Bruno et al., 2015). Accordingly, extensive prior work has analyzed the root causes of reporting errors and proposed strategies to reduce them (Kim and Mansfield, 2014; Pinto et al., 2012; Sangwaiya et al., 2009). Nonetheless, because reporting ultimately depends on human judgment under time pressure and imperfect information, errors cannot be eliminated entirely in routine practice. This motivates automated systems that can detect report errors reliably and at scale, providing a practical safety layer to support radiologists and improve the robustness of clinical care.

Several studies have explored AI-based radiology report error detection. Min et al. (2022) introduced a deep-learning framework that generates clinically meaningful synthetic errors and detects inconsistencies between the findings and conclusion sections of CXR reports. Subsequent studies evaluating commercial and open-source LLMs have largely focused on text-internal report errors, such as omissions, insertions, typographical errors, and findings–impression inconsistencies, often using controlled datasets containing deliberately introduced errors (Gertz et al., 2024; Kim et al., 2025; Salam et al., 2025; Yan et al., 2025).

Meanwhile, radiologist-annotated resources such as ReXVal (Yu et al., 2023) and RadEvalX (Calamida et al., 2024) provide fine-grained error annotations and evaluation protocols for generated radiology reports, but primarily evaluate candidate reports against reference reports rather than directly verifying their claims against the image. In practice, however, many consequential failures arise from image–report inconsistencies reflecting per-

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 49212–49228
October 24-29, 2026 ©2026 Association for Computational Linguistics

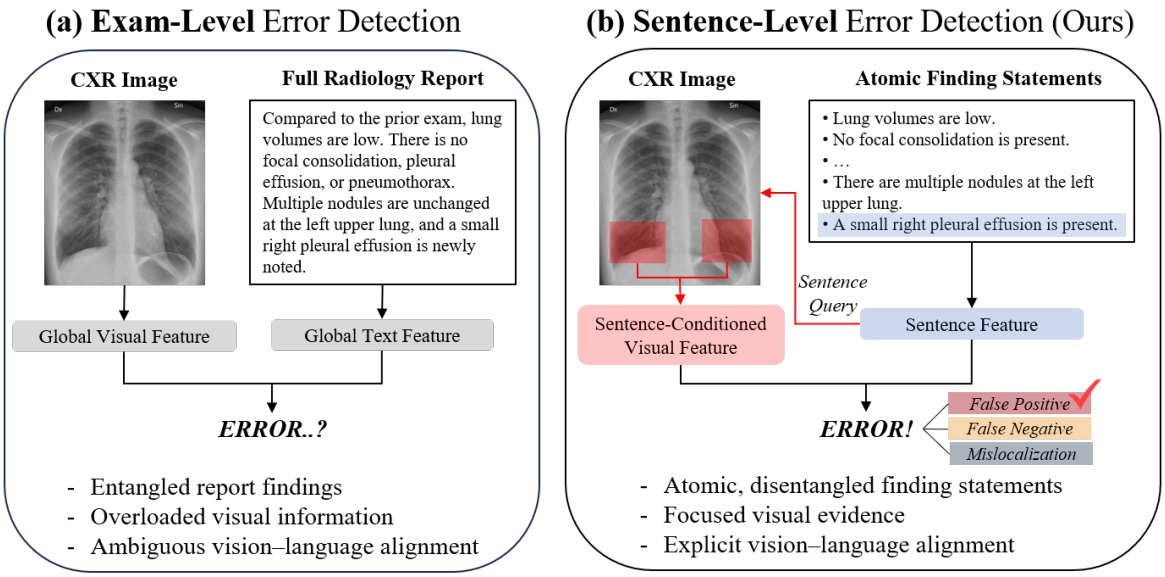


Figure 1: Exam-level versus sentence-level error detection. (a) Global report–image matching produces entangled supervision and ambiguous alignment. (b) Our method decomposes reports into atomic finding statements and aggregates sentence-conditioned visual evidence to detect errors.

ceptual or interpretive mistakes (Kim and Mansfield, 2014). More recently, CorBenchX introduced a CXR-based benchmark for image–report error detection and correction using vision-language models (Zou et al., 2025). Yet reliable exam-level detection remains challenging because clinical findings are expressed across entangled multi-sentence reports, while CXRs contain densely overlapping visual evidence, leading to ambiguous global vision–language alignment.

We reduce this ambiguity by reformulating radiology report error detection as sentence-level vision–language learning, where each atomic finding statement is independently verified against the image. To support this formulation, we leverage dual-level finding labels to synthesize erroneous image-sentence pairs spanning false positives, false negatives, and mislocalization, alongside consistent pairs. Using this supervision, we train SenRED, a sentence-conditioned vision–language model that grounds each atomic finding statement in localized visual evidence rather than relying on coarse global report-image alignment.

Our contributions are threefold. First, we formulate CXR report error detection as sentence-level vision–language consistency checking, where reports are decomposed into atomic finding statements and each statement is verified against image evidence. Second, we introduce a synthetic supervision strategy that aligns exam-level finding labels with sentence-level report claims, producing clinically meaningful error categories. Third, we propose SenRED, a sentence-conditioned patch-aggregation model that selectively aggregates visual evidence for each statement and improves both sentence-level and exam-level error detection, including on radiologist-confirmed subsets. Code, model weights, and the publicly distributable subset of the evaluation data will be released upon publication at `https://github.com/reonaledo/SenRED`.

2 Method

We formulate image-report consistency assessment as sentence-level error detection (Figure 1). Given an image I and an individual sentence s, the model predicts one of four consistency categories: false positive (FP), false negative (FN), mislocalization (MISLOC), or no error (NOERR).

2.1 Dual-Level Report Labeling

To synthesize sentence-level image–text consistency pairs, we use **exam-level** finding labels as proxies for image-level findings and **sentence-level** finding labels as representations of textual claims. We define synthetic inconsistencies as logical mismatches between these two finding label sets.

Because most existing radiology report labelers assign finding labels only at the exam level (Irvin et al., 2019; Smit et al., 2020), we train a

dedicated dual-level labeler under the CheX-GPT taxonomy (Gu et al., 2024). We use 12 finding categories plus *No Finding* after excluding *Lung Opacity* (13 labels in total): *Atelectasis*, *Consolidation*, *Effusion*, *Fracture*, *Hyperinflation*, *Nodule*, *Pleural Lesion*, *Pneumothorax*, *Pulmonary Edema*, *Subcutaneous Emphysema*, *Subdiaphragmatic Gas*, *Widened Mediastinal Silhouette*, and *No Finding*. For each category, the labeler predicts one of four assertion states: positive, negative, uncertain, and not mentioned.

Following RRIEF (Min et al., 2025), we train Llama-3-8B (Grattafiori et al., 2024) as a labeler using 3,000 exam-level instances, consisting of 1,000 randomly sampled studies each from MIMIC-CXR (Johnson et al., 2019, 2024), CheXpert (Irvin et al., 2019), and ReXGradient-160K (Zhang et al., 2025). To support sentence-level labeling, we additionally include the 3,000 most frequent sentences extracted from the reports of these sampled studies as sentence-level training instances, yielding 6,000 training instances in total. Initial labels are generated using Llama-3-70B-Instruct and subsequently reviewed and corrected by human annotators following guidelines provided by a board-certified radiologist.

2.2 Atomic Statement Decomposition

Radiology reports often express multiple findings with mixed assertions in a single sentence, which complicates sentence-level supervision. As preprocessing, we concatenate the findings and impression sections of each report and use Llama-3-70B-Instruct to split and rewrite the text so that each resulting sentence corresponds to a single atomic finding statement whenever possible. We also standardize temporal or comparative phrasing into statements about current-exam findings while preserving the original clinical meaning. Examples of this preprocessing are provided in Appendix A.

2.3 Sentence-Level Error Generation

Using exam-level finding labels for I and sentence-level finding labels for s, we synthesize FP/FN/MISLOC pairs in which s is logically inconsistent with I; consistent pairs are assigned to NOERR. We exclude uncertain labels from all error generation procedures and use only determinate positive, negative, or not-mentioned labels. To construct diverse FP, FN, and NOERR pairs from naturally occurring report language while reducing sampling bias and stylistic shortcuts, we sample sentences satisfying the required finding-label constraints from pools of atomic finding statements. We maintain a separate pool for each data split to prevent cross-split leakage. Under this framework, we construct samples according to the following category-specific constraints:

1. **FP:** If finding f is absent at the exam level (negative or not mentioned), we sample a sentence that positively mentions f and pair it with I, yielding an FP example. The paired NOERR example uses a sampled sentence that negatively mentions the same finding.
2. **FN:** If finding f is present at the exam level (positive), we sample either a sentence that negatively mentions f or a no-finding sentence (all findings negative or not mentioned, with *No Finding* positive) and pair it with I, yielding an FN example. The paired NOERR example uses a sampled sentence that positively mentions the same finding.
3. **MISLOC:** If a sentence positively mentions a laterality-sensitive finding f (*Atelectasis*, *Consolidation*, *Effusion*, *Fracture*, *Pneumothorax*, *Nodule*, *Pleural Lesion*, or *Subcutaneous Emphysema*) and contains a unilateral left/right cue, we swap left↔right, yielding a MISLOC example. The paired NOERR example uses the original, unswapped sentence. We exclude cases with bilateral or ambiguous laterality.
4. **NOERR:** In addition to the paired NOERR examples above, we include additional exam-consistent sentences as NOERR examples. For each exam-consistent atomic finding statement, we either use the original statement or sample another statement from the pool with the same sentence-level finding label.

2.4 Synthetic Error Dataset

We construct a sentence-level synthetic error dataset using MIMIC-CXR and Open-i (Demner-Fushman et al., 2016) (frontal-view images only) according to the generation procedure described in Section 2.3. For MIMIC-CXR, the training, development, and test splits are mutually patient-disjoint.

To assess whether the intended target error was clearly realized in the constructed image-sentence pairs, we additionally created radiologist-confirmed synthetic subsets. A board-certified radiologist with 22 years of experience reviewed a

Dataset	Split	Exams	Pairs	Error Types			
				FP	FN	MISLOC	NOERR
MIMIC-CXR	Train	194,494	2,043,098	194,494 (9.5)	375,681 (18.4)	89,056 (4.4)	1,383,867 (67.7)
	Dev	23,456	245,894	23,456 (9.5)	44,523 (18.1)	10,321 (4.2)	167,594 (68.2)
	Test	24,164	253,721	24,164 (9.5)	45,867 (18.1)	10,924 (4.3)	172,766 (68.1)
Open-i	Test	3,320	33,829	3,320 (9.8)	3,492 (10.3)	575 (1.7)	26,442 (78.2)
MIMIC-CXR Confirmed	Test	120	120	30 (25.0)	30 (25.0)	30 (25.0)	30 (25.0)
Open-i Confirmed	Test	60	60	15 (25.0)	15 (25.0)	15 (25.0)	15 (25.0)

Table 1: Dataset statistics. Each sample is an image-sentence pair. For each split, we report per-class counts and percentages.

subset of the constructed pairs and excluded ambiguous cases or samples in which the intended error type was not clearly realized. The review instructions are provided in Appendix B. Dataset statistics and class distributions for all splits are summarized in Table 1.

We also construct exam-level synthetic error datasets by incorporating each sentence-level sample into the original report context associated with its paired image. For each synthetic image–sentence pair (I, s), we take all atomic statements from the original report associated with image I and replace the corresponding target statement with s, leaving the remaining statements unchanged except for removing any surrounding statement whose sentence-level finding labels are incompatible with the target sentence. The resulting multi-sentence report is then paired with I and assigned the same report-level error type as the inserted target sentence: FP, FN, or MISLOC for erroneous samples, and NOERR when all retained statements are exam-consistent.

2.5 SenRED: Sentence-Level Report Error Detector

SenRED takes as input a CXR image I and a single atomic finding statement s, and outputs per-class error probabilities for FP, FN, MISLOC, and NOERR. We use independent sigmoid outputs instead of a softmax classifier to allow ambiguous or partially overlapping error-type supervision.

SenRED builds on pretrained encoders and trains only an error-detection head. We use a MedSigLIP (Google Health AI Developer Foundations, 2025) vision encoder to embed the image into a sequence of patch tokens $\mathbf{V} \in \mathbb{R}^{N \times d_v}$, and encode the input sentence using llm2vec4cxr (Ko et al., 2025) to obtain a sentence embedding $\mathbf{t} \in \mathbb{R}^{d_t}$. Both encoders are frozen during training to preserve their pretrained vision-language semantics.

Sentence-Conditioned Perceiver Head. Each report sentence refers to only a small subset of the anatomical and pathological information present in the image. Therefore, sentence-level error detection should selectively aggregate visual evidence relevant to the claim expressed by s, rather than relying on a global image-text similarity. SenRED adopts the Perceiver Resampler module (Alayrac et al., 2022) and conditions it on the input sentence to attend from sentence-derived latent queries to image patch tokens.

Concretely, we first project the patch tokens and the sentence embedding to a shared model dimension d. We then initialize R latent queries by repeating the sentence embedding R times and concatenating it with learnable seed latents, followed by a linear projection. We apply L Perceiver cross-attention blocks in which the latents act as text queries and the patch tokens, with positional embeddings added, provide keys and values. After the final block, we obtain a sentence-specific visual representation by mean-pooling over the R output latents, yielding a compact Perceiver summary feature $\mathbf{h}_{\text{perc}} \in \mathbb{R}^d$.

The latent bottleneck can compress away fine-grained patch-level cues that are critical for subtle findings. To retain this information, we additionally compute a high-fidelity feature by aggregating the Perceiver cross-attention weights into a patch-importance vector $\boldsymbol{\alpha} \in \mathbb{R}^N$, averaged over heads and latents and normalized, and forming a weighted sum over the original vision patch tokens. This yields $\mathbf{h}_{\text{hifi}} \in \mathbb{R}^{d_v}$, which preserves localized evidence highlighted by sentence-conditioned atten-

tion.

We concatenate the Perceiver summary $\mathbf{h}_{\text{perc}}$, the HiFi feature $\mathbf{h}_{\text{hifi}}$, and the projected sentence embedding, and apply a linear classifier followed by sigmoid to produce the four error probabilities.

3 Experiments

3.1 Experimental Setup

We train SenRED on the MIMIC-CXR training split using AdamW (Loshchilov and Hutter, 2019) with learning rate 1×10^{-4} and batch size 256 for five epochs. We use $R = 1$, $L = 1$, and $d = 512$ for the Perceiver head. We optimize multi-label binary cross-entropy over four sigmoid outputs. At inference, we use the NOERR output as a gate. If $p_{\text{NOERR}} \geq \tau$, we predict NOERR; otherwise, we select the class with the highest probability among FP, FN, and MISLOC. We use $\tau = 0.5$ for the primary results. For all trained models, we report single-run results using the checkpoint selected by the lowest development loss.

We evaluate SenRED on the synthetic error datasets described above, reporting internal results on MIMIC-CXR and external validation on Open-i. We report per-class and macro-average F1, with 95% confidence intervals based on 10,000 bootstrap resamples. Additional implementation details are reported in Appendix C.

For exam-level evaluation, we aggregate sentence-level predictions using a strict rule: a NOERR exam is correct only if all sentences are predicted as NOERR, whereas a corrupted exam is correct only if the injected target sentence alone is flagged and assigned the correct error type. We also report single-target top-1 aggregation, which, when one or more sentences are flagged, selects the sentence with the highest error probability (the maximum over FP, FN, and MISLOC).

3.2 Baselines

We compare SenRED against two groups of baselines: task-specific trained baselines and zero-shot instruction-tuned VLM baselines.

Task-specific trained baselines. We first compare against three baselines that differ in error construction and prediction granularity while keeping the underlying head architecture identical across methods to ensure a fair comparison.

Exam-Swap generates errors by randomly swapping the entire report of an exam with that of another exam with a different finding label; the model takes the image and full report as input and predicts a binary label.

Sent-Swap decomposes reports into atomic finding statements and generates errors by swapping a sentence with one from a different exam whose finding label differs from that of the original sentence; the model receives an image-sentence pair and predicts a binary label. Because Exam-Swap and Sent-Swap are binary consistency baselines, they predict only whether an image–text pair contains any error and do not classify the error type. We therefore report their NOERR F1, while FP, FN, and MISLOC entries are marked as not applicable.

Exam-1SentErr is trained on the exam-level synthetic error dataset, where each exam contains exactly one injected error sentence; the model uses a single global text embedding computed from all sentences in the exam, together with the image representation, to predict a four-class error for the exam.

Zero-shot instruction-tuned VLM baselines. To assess whether existing instruction-tuned VLMs can perform image-report error detection without task-specific supervision, we additionally evaluate Qwen3-VL-4B-Instruct (Bai et al., 2025), MedGemma-4B-Instruct (Sellergren et al., 2025), and CheXagent-2-3B (Chen et al., 2024) on the radiologist-confirmed subsets. We use zero-shot prompted inference: each model is given the CXR image and either the target atomic sentence or the exam-level report, and is asked a binary yes/no question for a specific error type. We run separate prompts for FP, FN, MISLOC, and NOERR and compute per-class F1 against the corresponding one-vs-rest label. Thus, the VLM baselines are evaluated as class-wise binary detectors rather than as constrained four-way classifiers. This protocol gives the VLMs explicit access to the target error type in each query and avoids requiring them to resolve conflicts among error categories. The exact prompt templates are provided in Appendix D.

3.3 Sentence-Level Results

Table 2 reports sentence-level results on the error-injected MIMIC-CXR and Open-i test sets. SenRED outperforms Exam-1SentErr in macro-F1 on both datasets: 0.796 versus 0.610 on MIMIC-CXR and 0.743 versus 0.538 on Open-i. The gain is largest for MISLOC (0.695 vs. 0.327 on MIMIC-CXR; 0.593 vs. 0.255 on Open-i), suggest-

Dataset	Method	Error Types				
		FP	FN	MISLOC	NOERR	Avg
MIMIC-CXR	Exam-Swap	–	–	–	0.706 [0.705, 0.708]	–
	Sent-Swap	–	–	–	0.842 [0.841, 0.843]	–
	Exam-1SentErr	0.601 [0.596, 0.606]	0.666 [0.662, 0.669]	0.327 [0.317, 0.336]	0.848 [0.846, 0.849]	0.610 [0.607, 0.613]
	SenRED	**0.771** [0.767, 0.775]	**0.812** [0.809, 0.815]	**0.695** [0.688, 0.702]	**0.906** [0.905, 0.907]	**0.796** [0.793, 0.798]
Open-i	Exam-Swap	–	–	–	0.854 [0.851, 0.858]	–
	Sent-Swap	–	–	–	0.884 [0.881, 0.886]	–
	Exam-1SentErr	0.613 [0.602, 0.625]	0.400 [0.383, 0.418]	0.255 [0.212, 0.297]	0.884 [0.881, 0.887]	0.538 [0.526, 0.550]
	SenRED	**0.792** [0.781, 0.802]	**0.660** [0.646, 0.673]	**0.593** [0.559, 0.625]	**0.929** [0.927, 0.931]	**0.743** [0.733, 0.753]

Table 2: Sentence-level error detection results on the full error-injected MIMIC-CXR and Open-i test sets. Values are per-class F1 and macro-F1 (Avg) with 95% confidence intervals from 10,000 bootstrap resamples. Because Exam-Swap and Sent-Swap are binary baselines, they do not produce error-type-specific predictions.

Dataset	Method	Error Types				
		FP	FN	MISLOC	NOERR	Avg
MIMIC-CXR Confirmed	Exam-Swap	–	–	–	0.609 [0.462, 0.736]	–
	Sent-Swap	–	–	–	0.682 [0.560, 0.788]	–
	Exam-1SentErr	0.585 [0.426, 0.714]	0.784 [0.640, 0.897]	0.378 [0.160, 0.561]	0.621 [0.488, 0.729]	0.592 [0.500, 0.672]
	Qwen3-VL-4B-Instruct	0.182 [0.044, 0.340]	0.716 [0.591, 0.818]	0.100 [0.000, 0.240]	0.646 [0.500, 0.769]	0.411 [0.346, 0.473]
	MedGemma-4B-Instruct	0.000 [0.000, 0.000]	0.520 [0.333, 0.678]	0.000 [0.000, 0.000]	0.500 [0.378, 0.605]	0.255 [0.205, 0.299]
	CheXagent-2-3B	0.079 [0.000, 0.167]	0.101 [0.024, 0.198]	0.458 [0.316, 0.584]	0.694 [0.559, 0.809]	0.333 [0.302, 0.358]
	SenRED	**0.737** [0.594, 0.853]	**0.983** [0.939, 1.000]	**0.808** [0.667, 0.913]	**0.806** [0.694, 0.897]	**0.833** [0.763, 0.895]
Open-i Confirmed	Exam-Swap	–	–	–	0.542 [0.353, 0.696]	–
	Sent-Swap	–	–	–	0.578 [0.383, 0.735]	–
	Exam-1SentErr	0.684 [0.483, 0.833]	0.400 [0.111, 0.645]	0.421 [0.118, 0.667]	0.605 [0.410, 0.760]	0.528 [0.386, 0.646]
	Qwen3-VL-4B-Instruct	0.200 [0.000, 0.435]	0.516 [0.273, 0.710]	0.000 [0.000, 0.000]	0.537 [0.333, 0.706]	0.313 [0.216, 0.402]
	MedGemma-4B-Instruct	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.483 [0.311, 0.630]	0.121 [0.078, 0.158]
	CheXagent-2-3B	0.174 [0.044, 0.320]	0.233 [0.056, 0.400]	0.340 [0.157, 0.509]	0.714 [0.533, 0.851]	0.365 [0.325, 0.399]
	SenRED	**0.800** [0.609, 0.938]	**0.815** [0.615, 0.952]	**0.815** [0.615, 0.957]	**0.778** [0.593, 0.909]	**0.802** [0.689, 0.896]

Table 3: Sentence-level results on radiologist-confirmed subsets. Values are per-class F1 and macro-F1 (Avg) with 95% confidence intervals from 10,000 bootstrap resamples. The zero-shot VLMs are evaluated as class-wise one-vs-rest detectors. Because Exam-Swap and Sent-Swap are binary baselines, they do not produce error-type-specific predictions.

ing that sentence-level supervision better supports laterality-sensitive image–text grounding. Unlike the binary baselines, SenRED identifies the specific error type while achieving the highest NOERR F1 among all compared methods on both datasets.

On the radiologist-confirmed subsets (Table 3), SenRED improves its macro-F1 from 0.796 to 0.833 on MIMIC-CXR and from 0.743 to 0.802 on Open-i, while consistently outperforming both task-specific baselines and zero-shot VLM baselines under a more clinically curated evaluation setting. The zero-shot VLMs show low and unstable F1 scores across error types, suggesting that global image–text alignment is insufficient to provide the finding-specific visual grounding required for fine-grained CXR–report error detection. Notably, MedGemma-4B-Instruct employs the same MedSigLIP vision encoder as SenRED, yet yields substantially inferior performance. This observation suggests that the challenge does not primarily arise from visual representation quality, but from the need for explicit sentence-level supervision and clinically grounded error modeling.

Additional analyses of threshold sensitivity, class-wise AUPRC, and image ablation are provided in Appendix E.

3.4 Exam-Level Results

Table 4 reports exam-level results on the radiologist-confirmed subsets. Using strict aggregation of its sentence-level predictions, SenRED outperforms both task-specific and zero-shot VLM baselines on both datasets. It achieves the best macro-F1 on both confirmed sets, with 0.658 on

Dataset	Method	Error Types				
		FP	FN	MISLOC	NOERR	Avg
MIMIC-CXR Confirmed	Exam-Swap	–	–	–	0.441 [0.333, 0.541]	–
	Sent-Swap	–	–	–	**0.677** [0.529, 0.795]	–
	Exam-1SentErr	0.565 [0.367, 0.724]	**0.800** [0.667, 0.899]	0.333 [0.125, 0.524]	0.571 [0.447, 0.680]	0.568 [0.475, 0.652]
	Qwen3-VL-4B-Instruct	0.368 [0.222, 0.500]	0.516 [0.353, 0.657]	0.405 [0.261, 0.535]	0.429 [0.222, 0.604]	0.430 [0.353, 0.496]
	MedGemma-4B-Instruct	0.000 [0.000, 0.000]	0.250 [0.061, 0.426]	0.000 [0.000, 0.000]	0.429 [0.319, 0.532]	0.170 [0.118, 0.220]
	CheXagent-2-3B	0.299 [0.194, 0.400]	0.333 [0.222, 0.436]	0.433 [0.325, 0.533]	0.450 [0.339, 0.551]	0.379 [0.362, 0.390]
	SenRED	**0.638** [0.457, 0.781]	0.746 [0.600, 0.857]	**0.591** [0.390, 0.750]	0.656 [0.508, 0.776]	**0.658** [0.565, 0.737]
	SenRED (Single-target top-1)	*0.667 [0.500, 0.795]*	*0.806 [0.689, 0.897]*	*0.720 [0.558, 0.846]*	*0.656 [0.508, 0.776]*	*0.712 [0.623, 0.788]*
Open-i Confirmed	Exam-Swap	–	–	–	0.388 [0.230, 0.528]	–
	Sent-Swap	–	–	–	0.625 [0.387, 0.800]	–
	Exam-1SentErr	0.615 [0.348, 0.813]	0.400 [0.111, 0.667]	0.333 [0.000, 0.588]	0.464 [0.291, 0.615]	0.453 [0.311, 0.578]
	Qwen3-VL-4B-Instruct	0.381 [0.182, 0.560]	0.444 [0.182, 0.667]	0.270 [0.069, 0.450]	0.500 [0.240, 0.710]	0.399 [0.291, 0.490]
	MedGemma-4B-Instruct	0.000 [0.000, 0.000]	0.100 [0.000, 0.300]	0.000 [0.000, 0.000]	0.394 [0.237, 0.535]	0.124 [0.069, 0.187]
	CheXagent-2-3B	0.203 [0.071, 0.339]	0.433 [0.269, 0.581]	0.462 [0.300, 0.603]	0.400 [0.233, 0.548]	0.375 [0.348, 0.392]
	SenRED	**0.759** [0.539, 0.909]	**0.583** [0.308, 0.800]	**0.546** [0.235, 0.769]	**0.750** [0.546, 0.895]	**0.659** [0.523, 0.768]
	SenRED (Single-target top-1)	*0.667 [0.444, 0.833]*	*0.710 [0.500, 0.870]*	*0.667 [0.400, 0.857]*	*0.750 [0.546, 0.895]*	*0.698 [0.565, 0.808]*

Table 4: Exam-level results on radiologist-confirmed subsets. Values are per-class F1 and macro-F1 (Avg) with 95% confidence intervals from 10,000 bootstrap resamples. The zero-shot VLMs are evaluated as class-wise one-vs-rest detectors. Because Exam-Swap and Sent-Swap are binary baselines, only their NOERR F1 is reported.

MIMIC-CXR and 0.659 on Open-i, yielding absolute gains of 0.090 and 0.206 over Exam-1SentErr, respectively. Among the three error types, the largest gains over Exam-1SentErr are observed for MISLOC on both datasets. Compared with Exam-1SentErr, which predicts from a single global report embedding, these results support sentence-wise verification followed by aggregation as an effective approach to exam-level error detection.

We further report SenRED with single-target top-1 aggregation, which improves macro-F1 from 0.658 to 0.712 on MIMIC-CXR and from 0.659 to 0.698 on Open-i. This gain indicates that even when multiple sentences are flagged, SenRED often ranks the injected erroneous sentence highest and correctly identifies its error type.

3.5 Ablation Study

Figure 2 summarizes macro-F1 on the full error-injected test sets and radiologist-confirmed subsets, with class-wise results provided in Appendix Table 9. Globally pooled linear and MLP heads consistently underperform sentence-conditioned cross-attention and Perceiver heads (SenRED). Notably, although SenRED uses the MedSigLIP vision encoder, replacing llm2vec4cxr with the jointly pretrained MedSigLIP text encoder substantially degrades performance, particularly on Open-i. This suggests that the MedSigLIP text encoder is less effective than llm2vec4cxr at capturing the fine-grained clinical semantics needed for error detection. Removing HiFi generally yields modestly lower macro-F1 point estimates across evaluation settings. Increasing Perceiver capacity in SenRED-Large ($R = 4$, $L = 2$) does not improve performance on the full test sets but achieves higher macro-F1 on both confirmed subsets, suggesting that increased capacity may be more beneficial for clinically adjudicated, less ambiguous cases.

4 Conclusion

We reformulated CXR report error detection as sentence-level vision-language learning by decomposing reports into atomic finding statements and verifying each statement against image evidence. We introduced SenRED, a sentence-level report error detector that uses sentence-conditioned patch aggregation to predict clinically relevant error categories, including false positives, false negatives, and mislocalization. Experiments on MIMIC-CXR and Open-i show that SenRED consistently improves over task-specific baselines and zero-shot instruction-tuned VLMs in both sentence-level and exam-level evaluations, with gains maintained on radiologist-confirmed subsets. These results support sentence-level consistency checking as an effective approach for fine-grained CXR-report error detection.

5 Limitations

This study evaluates synthetically injected CXR–sentence inconsistencies in MIMIC-CXR and

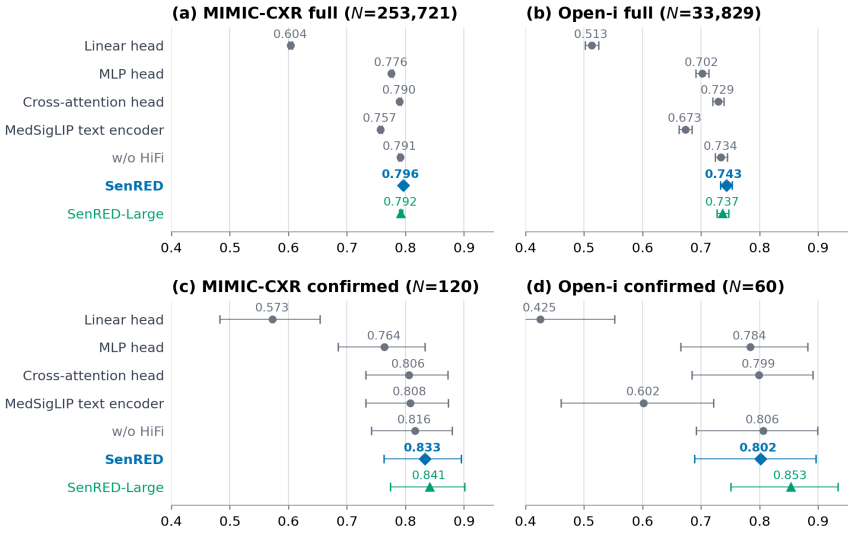


Figure 2: Macro-F1 ablation results on the MIMIC-CXR and Open-i full error-injected test sets and their radiologist-confirmed subsets. Points denote macro-F1, dotted horizontal bars indicate 95% confidence intervals from 10,000 bootstrap resamples. Complete class-wise ablation results are provided in Appendix Table 9.

Open-i rather than naturally occurring report errors. Our FP/FN/MISLOC taxonomy excludes clinically relevant errors involving severity, hedging, morphology, temporal change, and multiple interacting errors. Exam-level labels are report-derived proxies rather than direct image annotations, and the dual-level labeler lacks independent validation on a held-out benchmark; errors in source reports and label assignment may therefore propagate through training and evaluation. Although radiologist-confirmed subsets are used for more reliable evaluation, they are relatively small and may not capture the full distribution of real clinical report errors. Our exam-level aggregation assumes at most one injected target error in the confirmed setting. Because zero-shot VLMs differ in task supervision and domain-specific pretraining, they serve only as contextual references rather than supervision-matched architectural comparisons. Evaluation on natural errors across broader institutions, modalities, languages, reporting styles, and multi-error reports is needed before clinical deployment.

6 Ethical Considerations

This study was approved by our institutional review board. All experiments were conducted using de-identified, previously collected chest X-ray datasets. We used the datasets in accordance with their respective data-use agreements and licenses. In particular, restricted clinical data were not transmitted to external APIs or public online services; all preprocessing, including report decomposition and label generation, as well as model training and evaluation, was performed in a local controlled computing environment. Any derived artifacts, including synthetic or radiologist-confirmed evaluation examples, are intended for research use only and will be released only to the extent permitted by the access conditions and licenses of the source datasets.

SenRED is intended as a research prototype for studying image–report consistency in chest X-ray interpretation. It is not intended for autonomous diagnosis or direct clinical deployment. Incorrect predictions could either miss clinically relevant report errors or create unnecessary alerts, so any future clinical use would require prospective validation, careful monitoring of failure modes, and integration under qualified human oversight.

Use of AI Assistants

We used ChatGPT (OpenAI) for code debugging and manuscript editing. The authors reviewed and take responsibility for all content.

Acknowledgments

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: RS-2025-02213531, NTIS number: 2460004432), and by the National Research Foundation of Korea (NRF)

grants funded by the Ministry of Science and ICT (MSIT) (Grant No. RS-2024-00354666).

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, and 8 others. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736. Curran Associates, Inc.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Michael A. Bruno, Eric A. Walker, and Hani H. Abujudeh. 2015. Understanding and confronting our mistakes: The epidemiology of error in radiology and strategies for error reduction. *RadioGraphics*, 35(6):1668–1676.

Amos Rubin Calamida, Farhad Nooralahzadeh, Morteza Rohanian, Mizuho Nishio, Koji Fujimoto, and Michael Krauthammer. 2024. Radiology Report Generation Models Evaluation Dataset For Chest X-rays (RadEvalX). *PhysioNet*. Version 1.0.0.

Zhihong Chen, Maya Varma, Justin Xu, Magdalini Paschali, Dave Van Veen, Andrew Johnston, Alaa Youssef, Louis Blankemeier, Christian Bluethgen, Stephan Altmayer, Jeya Maria Jose Valanarasu, Mohamed Siddig Eltayeb Muneer, Eduardo Pontes Reis, Joseph Paul Cohen, Cameron Olsen, Tanishq Mathew Abraham, Emily B. Tsai, Christopher F. Beaulieu, Jenia Jitsev, and 4 others. 2024. A vision-language foundation model to enhance efficiency of chest x-ray interpretation. *Preprint*, arXiv:2401.12208.

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

R. J. Gertz, T. Dratsch, A. C. Bunck, S. Lennartz, A.-I. Iuga, M. G. Hellmich, T. Persigehl, L. Pennig, C. H. Gietzen, P. Fervers, D. Maintz, R. Hahnfeldt, and J. Kottlors. 2024. Potential of GPT-4 for detecting errors in radiology reports: Implications for reporting accuracy. *Radiology*, 311(1):e232714.

Google Health AI Developer Foundations. 2025. MedSigLIP | health ai developer foundations. Google for Developers. Last updated: 2025-07-09; accessed: 2026-02-27.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jawook Gu, Kihyun You, Han-Cheol Cho, Jiho Kim, Eun Kyoung Hong, and Byungseok Roh. 2024. CheX-GPT: Harnessing large language models for enhanced chest x-ray report labeling. *Preprint*, arXiv:2401.11505. Accessed: 2026-02-26.

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. 2019. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(1):590–597.

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

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. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317.

Songsoo Kim, Donghyun Kim, Hyun Joo Shin, Seung Hyun Lee, Yeseul Kang, Sejin Jeong, Jaewoong Kim, Miran Han, Seong-Joon Lee, Joonho Kim, Jungyon Yum, Changho Han, and Dukyong Yoon. 2025. Large-scale validation of the feasibility of GPT-4 as a proofreading tool for head CT reports. *Radiology*, 314(1):e240701.

Y. W. Kim and L. T. Mansfield. 2014. Fool me twice: Delayed diagnoses in radiology with emphasis on perpetuated errors. *American Journal of Roentgenology*, 202(3):465–470.

Hanbin Ko, Gihun Cho, Inhyeok Baek, Donguk Kim, Joonbeom Koo, Changi Kim, Dongheon Lee, and Chang Min Park. 2025. Exploring the capabilities of LLM encoders for image-text retrieval in chest x-rays. *Preprint*, arXiv:2509.15234.

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

Dabin Min, Kaeun Kim, Jong Hyuk Lee, Yisak Kim, and Chang Min Park. 2022. RRED: A radiology report error detector based on deep learning framework.

In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 41–52, Seattle, WA. Association for Computational Linguistics.

Dabin Min, Soyeon Kim, Sangheum Hwang, Kwang Nam Jin, SangHeum Bang, Won Gi Jeong, Jinwook Choi, Jung Min Chang, and Chang Min Park. 2025. Privacy-preserving information extraction framework for diverse imaging reports using large language models. Research Square preprint.

Antonio Pinto, Luca Brunese, Fabio Pinto, Riccardo Reali, Stefania Daniele, and Luigia Romano. 2012. The concept of error and malpractice in radiology. *Seminars in Ultrasound, CT and MRI*, 33(4):275–279.

Babak Salam, Claire Stüwe, Sebastian Nowak, Alois M. Sprinkart, Maike Theis, Dmitrij Kravchenko, Narine Mesropyan, Tatjana Dell, Christoph Endler, Claus C. Pieper, Daniel L. Kuetting, Julian A. Luetkens, and Alexander Isaak. 2025. Large language models for error detection in radiology reports: a comparative analysis between closed-source and privacy-compliant open-source models. *European Radiology*, 35:4549–4557.

M. J. Sangwaiya, S. Saini, M. A. Blake, K. J. Dreyer, and M. K. Kalra. 2009. Errare humanum est: Frequency of laterality errors in radiology reports. *American Journal of Roentgenology*, 192(5):W239–W244.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, and 62 others. 2025. Medgemma technical report. *Preprint*, arXiv:2507.05201.

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

Yuqi Yan, Kai Wang, Bojian Feng, Jincao Yao, Tian Jiang, Zhiyan Jin, Yin Zheng, Yahan Zhou, Chen Chen, Lin Sui, Xiayi Chen, Yanhong Du, Jie Yang, Qianmeng Pan, Lingyan Zhou, Vicky Yang Wang, Ping Liang, and Dong Xu. 2025. The use of large language models in detecting chinese ultrasound report errors. *npj Digital Medicine*, 8(1):66.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Lee, Zahra Shakeri, Andrew Ng, Curtis Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. 2023. Radiology Report Expert Evaluation (ReXVal) Dataset. *PhysioNet*. Version 1.0.0.

Xiaoman Zhang, Julián N. Acosta, Josh Miller, Ouwen Huang, and Pranav Rajpurkar. 2025. ReXGradient-160K: A large-scale publicly available dataset of chest radiographs with free-text reports. *arXiv preprint arXiv:2505.00228*.

Jing Zou, Qingqiu Li, Chenyu Lian, Lihao Liu, Xiaohan Yan, Shujun Wang, and Jing Qin. 2025. CorBenchX: Large-scale chest X-Ray error dataset and vision-language model benchmark for report error correction. *Preprint*, arXiv:2505.12057.

A Examples of Report Preprocessing

Before generating sentence-level supervision, we concatenate the findings and impression sections of each report and rewrite the combined text into atomic finding statements. The preprocessing aims to split sentences that mention multiple findings, normalize negation and uncertainty, and convert temporal or comparative descriptions into statements about the current exam while preserving the original clinical meaning. Table 5 shows representative examples.

B Radiologist Confirmation Instructions

For the radiologist-confirmed test sets, each image–sentence pair was reviewed by a board-certified radiologist. The reviewing radiologist was shown the chest X-ray image, the target report sentence, and the intended evaluation target among FP, FN, MISLOC, and NOERR. The original instructions were provided in Korean, the radiologist's native language. We provide a complete English translation of the instructions below.

Instructions. The following error-type definitions were provided for reference:

- **False Positive Error:** the report text states that a finding is present, although the finding is not present in the image.
- **False Negative Error:** the report text states that a finding is absent, although the finding is present in the image.
- **Location Mismatch Error:** the report text describes an existing finding at an incorrect location, such as the wrong laterality.

For each item, the chest X-ray image, the report sentence, and the intended error type were shown. The current evaluation target was one of FP, FN, MISLOC, and NOERR. The radiologist was asked to determine whether the given report sentence was consistent with the image findings.

The available response options were:

- Inconsistent
- Consistent
- Ambiguous / cannot determine

Items marked as ambiguous or not determinable were excluded from the radiologist-confirmed subsets.

C Implementation Details

We implemented all trainable components in PyTorch and used Hugging Face Transformers to load pretrained encoders and vision-language models. In SenRED, both the image and sentence encoders were kept frozen, and only the sentence-conditioned prediction head was trained. We used `google/medsiglip-448` as the vision encoder and `lukeingawesome/llm2vec4cxr` as the sentence encoder. Chest X-ray images were preprocessed using the default MedSigLIP image processor associated with `google/medsiglip-448`.

The frozen LLM2Vec4CXR sentence encoder contains approximately 1.0B parameters, and the frozen MedSigLIP vision encoder contains approximately 0.9B parameters. The trainable SenRED head contains 5,855,240 parameters. Overall, SenRED uses approximately 1.91B parameters, of which 5.86M are trainable.

All main training experiments were conducted on a single NVIDIA RTX A6000 GPU. Training SenRED for five epochs took approximately 132 hours, primarily due to repeated forward passes through the two frozen encoders; caching their outputs can substantially reduce prediction-head training time. The zero-shot VLM baselines were evaluated by inference only and were not fine-tuned.

For report preprocessing and initial label generation, we used Llama-3-70B with greedy decoding. Zero-shot VLM baselines were also evaluated with deterministic greedy decoding by disabling sampling. Evaluation metrics were computed using `sklearn.metrics.f1_score`.

D Prompts for Zero-Shot VLM Baselines

We evaluate Qwen3-VL-4B-Instruct, MedGemma-4B-Instruct, and CheXagent-2-3B using the same zero-shot prompting protocol. Each prompt consists of a system message and a user message. The system message is:

> You are an expert radiologist with extensive clinical experience. The user will provide a chest X-ray image and a sentence or radiology report. Your task is to classify whether the provided text contains the specified type of clinical error based on the given image.

For sentence-level evaluation, the user message contains the CXR image, the target sentence, and a class-specific query template, as shown in Table 6.

For exam-level evaluation, the user message contains the CXR image, the multi-sentence report,

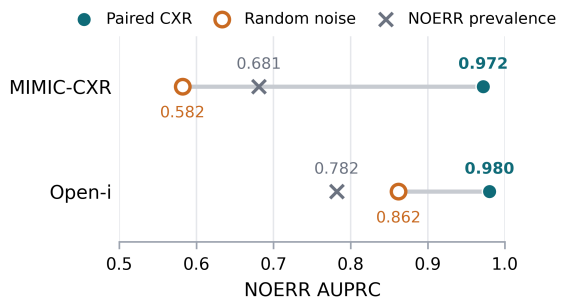


Figure 3: Image-ablated control on the full error-injected test sets. Filled circles and open circles denote NOERR AUPRC obtained with correctly paired CXRs and random-noise CXRs, respectively; gray crosses indicate the NOERR prevalence in each dataset.

and a class-specific query template, as shown in Table 7.

E Additional Analyses

E.1 Image-Ablated Control

Figure 3 compares NOERR AUPRC obtained with correctly paired CXRs and with random-noise images. Replacing the CXRs with random noise reduced AUPRC from 0.972 to 0.582 on MIMIC-CXR and from 0.980 to 0.862 on Open-i. The random-noise result fell below the NOERR prevalence on MIMIC-CXR but remained above it on Open-i, indicating that residual text-derived signal was more evident in the external dataset. Nevertheless, paired-CXR performance was consistently higher, supporting the conclusion that SenRED uses visual information and that its performance cannot be explained solely by textual construction cues.

E.2 Threshold Sensitivity and Class-Wise AUPRC

We evaluate threshold sensitivity for Exam-1SentErr and SenRED, the two four-class trained models that use the NOERR gate. The prompted VLM baselines return class-specific yes/no decisions and do not use this inference rule. Figure 4 reports macro-F1 for $\tau \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$. SenRED outperforms Exam-1SentErr at every evaluated threshold on all four evaluation sets.

Table 8 reports the corresponding threshold-free, class-wise area under the precision–recall curve (AUPRC) from the continuous output scores. SenRED achieves higher AUPRC than Exam-1SentErr for every output class on all four evaluation sets, and it also achieves the highest NOERR AUPRC among the task-specific methods.

F Class-Wise Ablation Results

Table 9 provides the per-class F1 scores and 95% confidence intervals corresponding to the macro-F1 results summarized in Figure 2. On the full test sets, the globally pooled linear head performs worst across all classes, while the largest advantage of sentence-conditioned aggregation over the stronger MLP head is observed for MISLOC: F1 increases from 0.630 to 0.686–0.695 on MIMIC-CXR and from 0.482 to 0.584–0.593 on Open-i with the cross-attention head and SenRED, respectively. This pattern is consistent with sentence-conditioned visual aggregation being particularly useful for errors that require finding-specific localization. Replacing llm2vec4cxr with the MedSigLIP text encoder lowers F1 for every class on both full test sets, with the largest reductions occurring for MISLOC, indicating that the choice of clinical sentence representation is especially important for fine-grained spatial consistency checking. The effect of HiFi is less uniform across classes: it generally benefits FN and MISLOC, whereas removing it occasionally produces slightly higher FP or NOERR point estimates, particularly on Open-i. Finally, the gains of SenRED-Large on the confirmed subsets arise from higher FP and FN F1 on MIMIC-CXR and improvements across all four classes on Open-i, most notably for MISLOC; however, the wide and substantially overlapping confidence intervals warrant cautious interpretation of these class-wise differences.

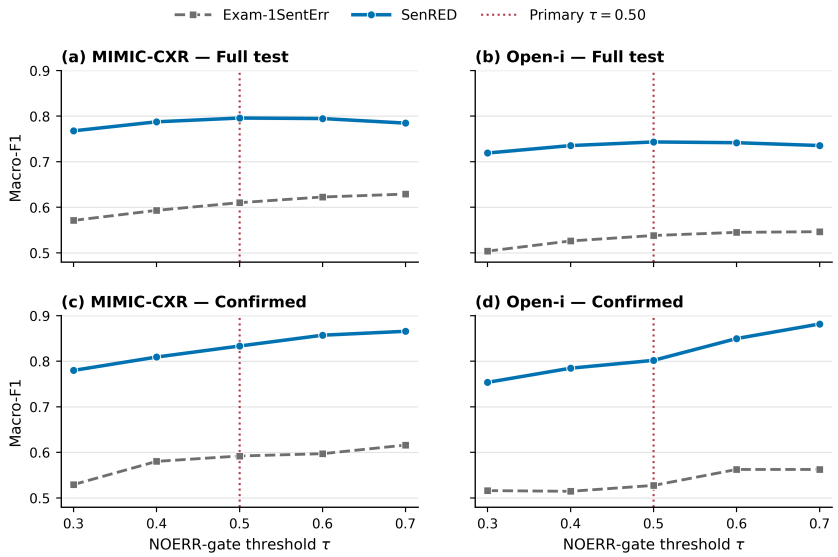


Figure 4: Threshold sensitivity of sentence-level macro-F1 on (a–b) the full error-injected test sets and (c–d) radiologist-confirmed synthetic subsets. Solid lines with circles denote SenRED, dashed lines with squares denote Exam-1SentErr, and vertical dotted lines mark the primary threshold ($\tau = 0.50$).

Original report	Preprocessed report
Example 1	
The lungs are clear of focal consolidation, pleural effusion or pneumothorax. The heart size is normal. The mediastinal contours are normal. Multiple surgical clips project over the left breast, and old left rib fractures are noted. No acute cardiopulmonary process.	The lungs are clear of focal consolidation. The lungs are clear of pleural effusion. The lungs are clear of pneumothorax. The heart size is normal. The mediastinal contours are normal. Multiple surgical clips project over the left breast. Old left rib fractures are noted. No acute cardiopulmonary process is present.
Finding labels: *Positive*: Fracture; *Negative*: Consolidation, Effusion, Pneumothorax, Widened Mediastinal Silhouette, and No Finding; *Not mentioned*: all remaining categories.	
Example 2	
Compared to chest radiographs since ___, most recently one ___. Previous mild pulmonary edema and possible concurrent pneumonia has all cleared. Heart is top-normal size, improved, and pleural effusions have resolved. Right hilar vessels are still enlarged, perhaps due to pulmonary arterial hypertension. Lateral view shows atherosclerotic coronary calcification in the left circumflex.	No pulmonary edema is present. No pneumonia is present. The heart is top-normal in size. No pleural effusions are present. Right hilar vessels are enlarged, perhaps due to pulmonary arterial hypertension. Atherosclerotic coronary calcification is seen in the left circumflex on the lateral view.
Finding labels: *Positive*: None; *Negative*: Consolidation, Effusion, Pulmonary Edema, and No Finding; *Not mentioned*: all remaining categories.	
Example 3	
As compared to ___, widespread areas of airspace consolidation have slightly improved and continue to involve the right lung to a greater degree than the left. Moderate right and small left pleural effusions are again demonstrated, and no pneumothorax detected.	There are widespread areas of airspace consolidation. The consolidation involves the right lung to a greater degree than the left. There is a moderate right pleural effusion. There is a small left pleural effusion. No pneumothorax is detected.
Finding labels: *Positive*: Consolidation and Effusion; *Negative*: Pneumothorax and No Finding; *Not mentioned*: all remaining categories.	
Example 4	
Comparison to ___. All monitoring and support devices, including the ventricular assist device, are in stable position. No substantial decrease in severity of the pre-existing pulmonary edema. Moderate cardiomegaly persists. No pneumothorax. Stable retrocardiac atelectasis.	All monitoring and support devices, including the ventricular assist device, are present. There is pulmonary edema. Moderate cardiomegaly is present. No pneumothorax. There is retrocardiac atelectasis.
Finding labels: *Positive*: Atelectasis, Pulmonary Edema, and Widened Mediastinal Silhouette; *Negative*: Pneumothorax and No Finding; *Not mentioned*: all remaining categories.	
Example 5	
The tracheostomy is in adequate position at 5.2 cm above the carina. There is a left-sided PICC line that ends in the upper portion of the superior vena cava. There is a nasogastric tube that is unchanged. There are two chest tubes without visible pneumothorax. Stability of the left subcutaneous air. There are still bilateral mild bands of atelectasis. The lung bases are better aerated. Resolution of the mild pulmonary edema. Mediastinal and cardiac contour are normal and unchanged. There are multiple rib fractures on the right axillary line that are overriding with adjacent mild hemothorax. 1. Resolution of the mild pulmonary edema. 2. Multiple, unstable, variably overriding right rib and small, stable adjacent hemothorax.	The tracheostomy is in adequate position at 5.2 cm above the carina. There is a left-sided PICC line that ends in the upper portion of the superior vena cava. There is a nasogastric tube. There are two chest tubes. No visible pneumothorax is present. There is left subcutaneous air. There are bilateral mild bands of atelectasis. The lung bases are aerated. No pulmonary edema is present. Mediastinal contour is normal. Cardiac contour is normal. There are multiple rib fractures on the right axillary line. The rib fractures are overriding. There is adjacent mild hemothorax.
Finding labels: *Positive*: Atelectasis, Fracture, and Subcutaneous Emphysema; *Negative*: Pneumothorax, Pulmonary Edema, Widened Mediastinal Silhouette, and No Finding; *Not mentioned*: all remaining categories.	

Table 5: Examples of report preprocessing and their corresponding finding labels. Original reports are formed by concatenating the findings and impression sections. Preprocessed reports split multi-finding sentences into atomic finding statements and normalize temporal or comparative descriptions into current-exam statements while preserving clinical meaning. Positive and negative finding labels are explicitly listed; all unlisted categories in the 13-label taxonomy are labeled *not mentioned*.

Target class	Sentence-level query
FP	Based on the given chest X-ray image, is this sentence describing a finding that is not actually present in the image (False Positive error)? Please answer yes or no. Sentence to evaluate: `{sentence}`
FN	Based on the given chest X-ray image, is this sentence incorrectly denying or implying the absence of a finding that is actually present in the image (False Negative error)? Please answer yes or no. Sentence to evaluate: `{sentence}`
MISLOC	Based on the given chest X-ray image, is this sentence describing a finding in the wrong anatomical location (Mislocalization error)? Please answer yes or no. Sentence to evaluate: `{sentence}`
NOERR	Based on the given chest X-ray image, does this sentence contain no radiological errors? Please answer yes if it has no error, and no if it has any radiological error. Sentence to evaluate: `{sentence}`

Table 6: Sentence-level prompt templates for zero-shot VLM baselines.

Target class	Exam-level query
FP	Based on the given chest X-ray image, is this radiology report describing any finding that is not actually present in the image (False Positive error)? Please answer yes or no without any descriptions. Radiology report to evaluate: `{report}`
FN	Based on the given chest X-ray image, is this radiology report incorrectly denying or omitting any finding that is actually present in the image (False Negative error)? Please answer yes or no without any descriptions. Radiology report to evaluate: `{report}`
MISLOC	Based on the given chest X-ray image, is this radiology report assigning any finding to the wrong anatomical location (Mislocalization error)? Please answer yes or no without any descriptions. Radiology report to evaluate: `{report}`
NOERR	Based on the given chest X-ray image, does this radiology report contain no radiological errors? Please answer yes if it has no error, and no if it has any radiological error. Do not include any descriptions. Radiology report to evaluate: `{report}`

Table 7: Exam-level prompt templates for zero-shot VLM baselines.

Dataset	Method	Class-wise AUPRC				
		FP	FN	MISLOC	NOERR	Avg
(a) Full error-injected test sets						
MIMIC-CXR	Exam-Swap	–	–	–	0.836	–
	Sent-Swap	–	–	–	0.906	–
	Exam-1SentErr	0.693	0.752	0.548	0.902	0.724
	SenRED	**0.862**	**0.869**	**0.757**	**0.972**	**0.865**
Open-i	Exam-Swap	–	–	–	0.878	–
	Sent-Swap	–	–	–	0.918	–
	Exam-1SentErr	0.749	0.521	0.400	0.928	0.649
	SenRED	**0.887**	**0.741**	**0.627**	**0.980**	**0.809**
(b) Radiologist-confirmed synthetic subsets						
MIMIC-CXR Confirmed	Exam-Swap	–	–	–	0.575	–
	Sent-Swap	–	–	–	0.769	–
	Exam-1SentErr	0.725	0.941	0.868	0.770	0.826
	SenRED	**0.866**	**1.000**	**0.964**	**0.939**	**0.942**
Open-i Confirmed	Exam-Swap	–	–	–	0.608	–
	Sent-Swap	–	–	–	0.599	–
	Exam-1SentErr	0.823	0.824	0.818	0.729	0.798
	SenRED	**0.923**	**0.976**	**0.916**	**0.979**	**0.948**

Table 8: Class-wise AUPRC on the full error-injected test sets and radiologist-confirmed synthetic subsets. Avg is the unweighted mean across FP, FN, MISLOC, and NOERR for methods that predict all four classes. Because Exam-Swap and Sent-Swap are binary consistency baselines, only their NOERR AUPRC is applicable.

Dataset	Method	Error Types				
		FP	FN	MISLOC	NOERR	Avg
(a) Full error-injected test sets						
MIMIC-CXR	Linear head	0.551 [0.545, 0.557]	0.682 [0.678, 0.685]	0.329 [0.320, 0.337]	0.854 [0.853, 0.856]	0.604 [0.601, 0.607]
	MLP head	0.760 [0.755, 0.764]	0.811 [0.808, 0.813]	0.630 [0.622, 0.637]	0.905 [0.904, 0.906]	0.776 [0.774, 0.779]
	Cross-attention head	0.767 [0.763, 0.771]	0.806 [0.803, 0.808]	0.686 [0.679, 0.693]	0.903 [0.901, 0.904]	0.790 [0.788, 0.793]
	MedSigLIP text encoder	0.736 [0.732, 0.741]	0.789 [0.786, 0.792]	0.609 [0.601, 0.617]	0.895 [0.894, 0.896]	0.757 [0.755, 0.760]
	w/o HiFi	0.767 [0.762, 0.771]	0.810 [0.807, 0.812]	0.681 [0.674, 0.688]	**0.906** [0.905, 0.907]	0.791 [0.788, 0.793]
	SenRED	**0.771** [0.767, 0.775]	**0.812** [0.809, 0.815]	**0.695** [0.688, 0.702]	**0.906** [0.905, 0.907]	**0.796** [0.793, 0.798]
	SenRED-Large	0.768 [0.763, 0.772]	0.810 [0.808, 0.813]	0.684 [0.677, 0.691]	**0.906** [0.905, 0.907]	0.792 [0.790, 0.795]
Open-i	Linear head	0.425 [0.408, 0.442]	0.494 [0.478, 0.510]	0.243 [0.204, 0.283]	0.892 [0.889, 0.894]	0.513 [0.502, 0.525]
	MLP head	0.758 [0.747, 0.769]	0.645 [0.631, 0.658]	0.482 [0.444, 0.520]	0.924 [0.921, 0.926]	0.702 [0.691, 0.713]
	Cross-attention head	0.775 [0.764, 0.786]	0.632 [0.618, 0.646]	0.584 [0.551, 0.615]	0.926 [0.924, 0.929]	0.729 [0.720, 0.739]
	MedSigLIP text encoder	0.741 [0.729, 0.752]	0.569 [0.554, 0.585]	0.463 [0.425, 0.500]	0.919 [0.917, 0.921]	0.673 [0.662, 0.684]
	w/o HiFi	**0.794** [0.783, 0.804]	0.646 [0.632, 0.660]	0.567 [0.532, 0.602]	**0.931** [0.929, 0.933]	0.734 [0.724, 0.745]
	SenRED	0.792 [0.781, 0.802]	**0.660** [0.646, 0.673]	**0.593** [0.559, 0.625]	0.929 [0.927, 0.931]	**0.743** [0.733, 0.753]
	SenRED-Large	0.784 [0.773, 0.794]	**0.660** [0.647, 0.674]	0.576 [0.541, 0.610]	0.928 [0.926, 0.931]	0.737 [0.727, 0.747]
(b) Radiologist-confirmed synthetic subsets						
MIMIC-CXR Confirmed	Linear head	0.419 [0.216, 0.595]	0.868 [0.755, 0.952]	0.468 [0.273, 0.630]	0.536 [0.409, 0.649]	0.573 [0.483, 0.654]
	MLP head	0.721 [0.577, 0.839]	0.983 [0.941, 1.000]	0.578 [0.381, 0.733]	0.773 [0.657, 0.868]	0.764 [0.685, 0.833]
	Cross-attention head	0.737 [0.591, 0.853]	0.949 [0.880, 1.000]	0.745 [0.591, 0.864]	0.795 [0.678, 0.889]	0.806 [0.732, 0.872]
	MedSigLIP text encoder	0.750 [0.604, 0.866]	0.967 [0.909, 1.000]	0.750 [0.595, 0.873]	0.763 [0.641, 0.861]	0.808 [0.732, 0.873]
	w/o HiFi	**0.772** [0.632, 0.881]	0.983 [0.941, 1.000]	0.735 [0.571, 0.857]	0.773 [0.656, 0.867]	0.816 [0.742, 0.880]
	SenRED	0.737 [0.594, 0.853]	0.983 [0.939, 1.000]	**0.808** [0.667, 0.913]	**0.806** [0.694, 0.897]	0.833 [0.763, 0.895]
	SenRED-Large	0.759 [0.619, 0.868]	**1.000** [1.000, 1.000]	0.800 [0.655, 0.909]	**0.806** [0.694, 0.897]	**0.841** [0.774, 0.901]
Open-i Confirmed	Linear head	0.300 [0.000, 0.546]	0.500 [0.182, 0.741]	0.400 [0.111, 0.647]	0.500 [0.328, 0.645]	0.425 [0.279, 0.552]
	MLP head	0.800 [0.609, 0.933]	**0.889** [0.727, 1.000]	0.696 [0.429, 0.889]	0.750 [0.571, 0.884]	0.784 [0.665, 0.882]
	Cross-attention head	0.815 [0.609, 0.957]	0.696 [0.421, 0.882]	**0.933** [0.815, 1.000]	0.750 [0.571, 0.882]	0.799 [0.684, 0.891]
	MedSigLIP text encoder	0.690 [0.455, 0.857]	0.546 [0.250, 0.774]	0.609 [0.333, 0.818]	0.565 [0.372, 0.720]	0.602 [0.461, 0.721]
	w/o HiFi	**0.828** [0.640, 0.957]	0.800 [0.571, 0.952]	0.846 [0.667, 0.968]	0.750 [0.571, 0.884]	0.806 [0.692, 0.899]
	SenRED	0.800 [0.609, 0.938]	0.815 [0.615, 0.952]	0.815 [0.615, 0.957]	0.778 [0.593, 0.909]	0.802 [0.689, 0.896]
	SenRED-Large	**0.828** [0.640, 0.957]	0.846 [0.667, 0.968]	0.929 [0.800, 1.000]	**0.811** [0.645, 0.930]	**0.853** [0.751, 0.934]

Table 9: Ablation results on (a) the full error-injected test sets and (b) radiologist-confirmed subsets of the synthetic error-injected test sets. Values are per-class F1 and macro-F1 (Avg) with 95% confidence intervals from 10,000 bootstrap resamples. SenRED denotes the default full model configuration. SenRED-Large uses increased Perceiver capacity.