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

Anonymous ACL submission

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 approach by constructing clinically meaningful synthetic errors and training a text model to detect erroneous statements within reports. Subsequent work–including prompting-based pipelines and evaluations of commercial or open-source large language models (LLMs)–has largely focused on text-centric discrepancies such as omissions, insertions, typographical issues, and intra-report inconsistencies under controlled synthetic settings (Alzghoul et al., 2024; Gertz et al., 2024; Kim et al., 2025; Salam et al., 2025; Yan et al., 2025). Meanwhile, radiologist-annotated resources such as ReX-Val (Yu et al., 2023) and RadEvalX (Calamida et al., 2024) have refined error taxonomies and evaluation protocols, but primarily compare report-to-report pairs without verifying claims against the image itself. In practice, however, many consequential failures are image-report inconsistencies that reflect perceptual or interpretive mistakes (Kim and Mansfield, 2014). Zou et al. (2025) recently shifted attention to image-report inconsistencies for CXRs. Yet reliable exam-level detection remains chal-

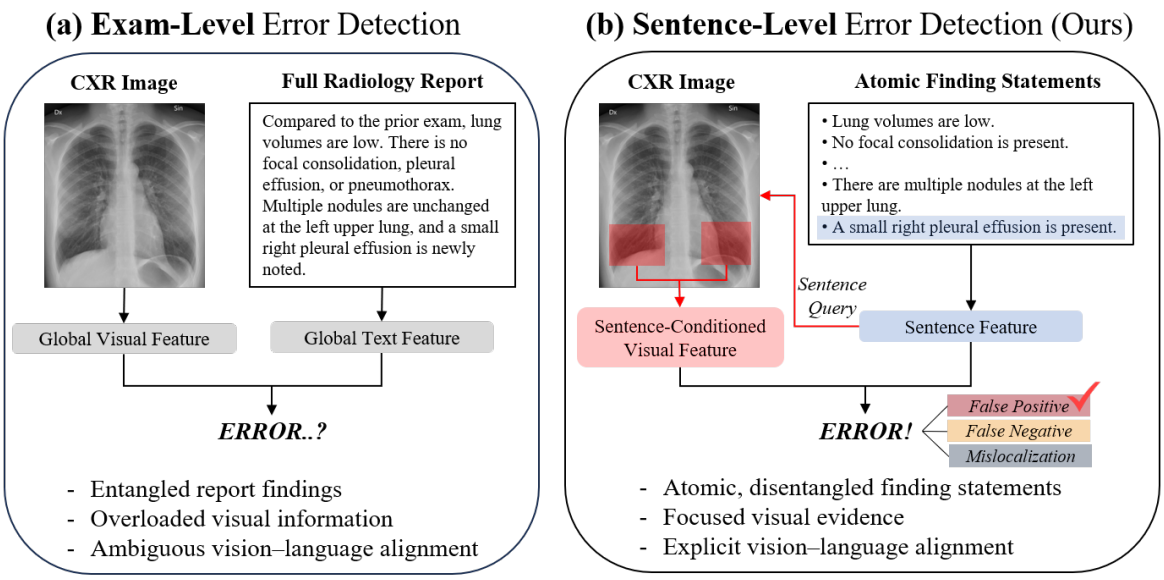

Figure 1: Exam-level versus sentence-level error detection. (a) Global report–image matching produces entangled supervision and ambiguous alignment. (b) Our method decomposes reports into atomic finding statements and aggregates sentence-conditioned visual evidence to detect errors.

lenging 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 test sets. We will also release our code, trained model, and the publicly distributable subset of our evaluation set.

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 labels: 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** labels as representations of textual claims. We define synthetic inconsistencies as logical mismatches between these two label sets.

Because most existing report labelers operate 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 13 finding cate-

gories after excluding *Lung Opacity*: *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., 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 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 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 labels for I and sentence-level labels for s, we synthesize FP/FN/MISLOC pairs by making s logically inconsistent with I, and label consistent pairs as NOERR. We exclude uncertain labels from all error generation procedures, so synthetic pairs are created only from determinate positive, negative, or not-mentioned labels. To reduce sampling bias and avoid style-based shortcuts, we build a global pool of atomic finding statements within each data split for FP, FN, and NOERR sampling. MISLOC examples are generated separately by deterministically transforming eligible original statements from the same exam. We create samples under the following label constraints:

1. **FP:** If finding f is absent at the exam level (negative or not mentioned), we sample a positive sentence for f to create an FP example. The paired NOERR example is created by sampling a negative sentence for the same finding.
2. **FN:** If finding f is present at the exam level (positive), we sample either a negative sentence for f or a no-finding sentence (all findings negative or not mentioned, with *No Finding* positive) to create an FN example. The paired NOERR example is created by sampling a positive sentence for the same finding.
3. **MISLOC:** If a laterality-sensitive finding (*Atelectasis*, *Consolidation*, *Effusion*, *Fracture*, *Pneumothorax*, *Nodule*, *Pleural Lesion*, and *Subcutaneous Emphysema*) f is positively mentioned in an original sentence from the same exam and the sentence contains a unilateral left/right cue, we swap left↔right to create a MISLOC example. The paired NOERR example is created using the original unswapped sentence. We exclude bilateral or ambiguous cases and duplicate swapped sentences.
4. **NOERR:** In addition to the paired NOERR examples above, we populate NOERR with exam-consistent sentences. These are either taken directly from the exam's own atomic finding statements or sampled from the global pool by matching the same sentence-level finding-label vector as an exam-consistent atomic finding statement.

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 more reliable evaluation, we additionally curate radiologist-confirmed test sets by having

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.

a board-certified radiologist with 22 years of experience review a subset of image-sentence pairs and remove ambiguous cases or samples where the intended error type is 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 datasets by embedding each sentence-level sample back into its original report context. 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 class 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 representations 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 global 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 informa-

tion, 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 attention.

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, if all three error-class probabilities are below 0.5, we predict NOERR; otherwise, we assign the error type with the highest probability among FP, FN, and MISLOC. For all trained models, we report single-run results using the checkpoint selected by the lowest development loss.

We evaluate SenRED on the synthetic datasets described above, reporting internal results on MIMIC-CXR and external validation on Open-i. For exam-level evaluation, we aggregate sentence-level predictions into a single exam decision. Under the strict policy, an exam is correct if the model either identifies exactly one erroneous sentence with the correct error type, or predicts NOERR throughout when no sentence is erroneous in the ground truth. We report per-class F1 and macro-average F1. Additional implementation details are reported in Appendix C.

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; the model takes the image and full report as input and predicts a binary label. **Sent-Swap** decomposes reports into atomic finding statements and creates errors by swapping a sentence with one from a different exam; 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 dataset, where each exam contains exactly one injected error sentence; the model jointly processes all sentences in an exam and predicts 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

Dataset	Method	Error Types				
		FP	FN	MISLOC	NOERR	Avg
MIMIC -CXR	Exam-Swap	–	–	–	0.706	–
	Sent-Swap	–	–	–	0.842	–
	Exam-1SentErr	0.601	0.665	0.327	0.848	0.610
	SenRED	**0.771**	**0.812**	**0.695**	**0.906**	**0.796**
Open-i	Exam-Swap	–	–	–	0.854	–
	Sent-Swap	–	–	–	**0.884**	–
	Exam-1SentErr	0.613	0.400	0.255	**0.884**	0.538
	SenRED	**0.792**	**0.660**	**0.593**	0.829	**0.743**

Table 2: Sentence-level error detection results on error-injected test sets from MIMIC-CXR and Open-i. F1 is reported for each class and macro-average (Avg). Because Exam-Swap and Sent-Swap are binary consistency baselines, they do not produce error-type-specific predictions; therefore, only their NOERR F1 is reported.

Dataset	Method	Error Types				
		FP	FN	MISLOC	NOERR	Avg
MIMIC-CXR Confirmed	Exam-Swap	–	–	–	0.609	–
	Sent-Swap	–	–	–	0.682	–
	Exam-1SentErr	0.585	0.784	0.378	0.621	0.592
	Qwen3-VL-4B-Instruct	0.182	0.716	0.100	0.646	0.411
	MedGemma-4B-Instruct	0.000	0.520	0.000	0.500	0.255
	CheXagent-2-3B	0.079	0.101	0.458	0.694	0.333
	SenRED	**0.737**	**0.983**	**0.808**	**0.806**	**0.833**
Open-i Confirmed	Exam-Swap	–	–	–	0.542	–
	Sent-Swap	–	–	–	0.578	–
	Exam-1SentErr	0.684	0.400	0.421	0.605	0.527
	Qwen3-VL-4B-Instruct	0.200	0.519	0.000	0.537	0.314
	MedGemma-4B-Instruct	0.000	0.000	0.000	0.483	0.121
	CheXagent-2-3B	0.174	0.227	0.340	0.714	0.364
	SenRED	**0.800**	**0.845**	**0.845**	**0.778**	**0.802**

Table 3: Sentence-level results on radiologist-confirmed subsets of the error-injected test sets. F1 is reported for each class and macro-average (Avg). Because Exam-Swap and Sent-Swap are binary consistency baselines, they do not produce error-type-specific predictions; therefore, only their NOERR F1 is reported.

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), suggesting that sentence-level supervision better supports laterality-sensitive image–text grounding. Although Sent-Swap and Exam-1SentErr obtain higher NOERR F1 on Open-i, they either do not predict error types or underperform on FP, FN, and MISLOC, which are the clinically actionable classes.

On the radiologist-confirmed subsets (Table 3), SenRED maintains strong performance and consistently outperforms 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, indicating that instruction following alone is insufficient for fine-grained CXR-report consistency checking. 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.

3.4 Exam-Level Results

Table 4 reports exam-level results on the radiologist-confirmed subsets. Despite being trained with sentence-level supervision, SenRED generalizes effectively to this multi-sentence setting. On both datasets, SenRED outperforms task-specific baselines and zero-shot VLM baselines, indicating that sentence-level supervision not only improves fine-grained error detection but also transfers to exam-level evaluation scenarios. It achieves the best macro-F1 on both confirmed sets, with 0.658 on MIMIC-CXR and 0.659 on Open-i, improving over Exam-1SentErr by 9.1 and 20.6 points, respectively. The largest

Dataset	Method	Error Types				
		FP	FN	MISLOC	NOERR	Avg
MIMIC-CXR Confirmed	Exam-Swap	–	–	–	0.441	–
	Sent-Swap	–	–	–	**0.677**	–
	Exam-1SentErr	0.565	**0.800**	0.333	0.571	0.567
	Qwen3-VL-4B-Instruct	0.368	0.516	0.405	0.429	0.430
	MedGemma-4B-Instruct	0.000	0.250	0.000	0.429	0.170
	CheXagent-2-3B	0.299	0.333	0.433	0.450	0.379
	SenRED	**0.638**	0.746	**0.591**	0.656	**0.658**
	SenRED (Oracle)	*0.667*	*0.806*	*0.720*	*0.656*	*0.712*
Open-i Confirmed	Exam-Swap	–	–	–	0.388	–
	Sent-Swap	–	–	–	0.625	–
	Exam-1SentErr	0.615	0.400	0.333	0.464	0.453
	Qwen3-VL-4B-Instruct	0.381	0.444	0.270	0.500	0.399
	MedGemma-4B-Instruct	0.000	0.100	0.000	0.394	0.123
	CheXagent-2-3B	0.203	0.433	0.462	0.400	0.375
	SenRED	**0.759**	**0.583**	**0.545**	**0.750**	**0.659**
	SenRED (Oracle)	*0.667*	*0.710*	*0.667*	*0.750*	*0.699*

Table 4: Exam-level results on confirmed test sets. F1 is reported per class and macro-average (Avg). *SenRED (Oracle)* assumes a single-error setting. Because Exam-Swap and Sent-Swap are binary consistency baselines, they do not produce error-type-specific predictions; therefore, only their NOERR F1 is reported.

Dataset	Method	Error Types				
		FP	FN	MISLOC	NOERR	Avg
MIMIC-CXR Confirmed	Head: Linear	0.419	0.868	0.468	0.536	0.573
	MLP	0.721	0.983	0.578	0.773	0.764
	Cross-attention	0.737	0.949	0.745	0.795	0.806
	TextEncoder: MedSigLIP	0.750	0.967	0.750	0.763	0.807
	SenRED	0.737	0.983	**0.808**	**0.806**	0.833
	SenRED-Large	**0.759**	**1.000**	0.800	**0.806**	**0.841**
Open-i Confirmed	Head: Linear	0.300	0.500	0.400	0.500	0.425
	MLP	0.800	**0.889**	0.696	0.750	0.784
	Cross-attention	0.815	0.656	**0.933**	0.750	0.798
	TextEncoder: MedSigLIP	0.690	0.545	0.609	0.565	0.602
	SenRED	0.800	0.815	0.815	0.778	0.802
	SenRED-Large	**0.828**	0.846	0.929	**0.811**	**0.853**

Table 5: Ablation study on confirmed test sets. F1 is reported per class and macro-average (Avg). SenRED denotes the full model configuration.

class-wise gains are observed for MISLOC on both datasets.

We further report SenRED (Oracle), which enforces the single-error assumption by selecting only the highest-scoring error sentence when multiple sentences are predicted as erroneous. The performance gain of the oracle variant indicates that SenRED typically ranks the true erroneous sentence highest and accurately distinguishes error types.

3.5 Ablation Study

Table 5 reports ablation results on the radiologist-confirmed subsets. Linear and MLP heads rely on globally pooled image-sentence features and consistently underperform models that operate at the patch level. Introducing a cross-attention head, which conditions image patch tokens on the sentence, yields clear performance gains. SenRED further replaces this module with a Perceiver head and achieves the best overall results. Replacing llm2vec4cxr with the original MedSigLIP text encoder substantially degrades external performance. Increasing the Perceiver capacity from the default SenRED configuration ($R = 1$, $L = 1$) to SenRED-Large ($R = 4$, $L = 2$) yields additional improvements, particularly on the external test set, suggesting that scaling latent capacity can further enhance robustness and generalization.

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 neg-

atives, 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 verification.

5 Limitations

This work focuses on chest X-ray report error detection and evaluates on MIMIC-CXR and Open-i; additional modalities, institutions, languages, and reporting styles may expose failure modes not covered here. The training data rely on synthetic sentence-level errors generated from label constraints. 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. The method depends on upstream report decomposition and dual-level labeling, so errors in these preprocessing steps can propagate to model training and evaluation. Finally, our exam-level aggregation assumes at most one injected target error in the confirmed setting; deployment in routine clinical workflows would require evaluation under reports containing multiple interacting errors.

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.

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 Menick, Sebastian Borgeaud, and 8 others. 2022. Flamingo: a visual language model for few-shot learning. *Preprint*, arXiv:2204.14198.

Renad M. Alzghoul, Abdulrahman Tabaza, Aya Abdelhaq, and Ahmad Altamimi. 2024. CLD-MEC at MEDIQA-CORR 2024 task: GPT-4 multi-stage clinical chain of thought prompting for medical errors detection and correction. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 537–556, Mexico City, Mexico. Association for Computational Linguistics.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 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, Alex Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, and 1 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. *Preprint*, arXiv:1901.07031.

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

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. *Preprint*, arXiv:1711.05101.

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, and 1 others. 2025. MedGemma technical report. *Preprint*, arXiv:2507.05201.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y. Ng, and Matthew P. Lungren. 2020. CheXbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. *Preprint*, arXiv:2004.09167.

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, Julian N. Acosta, Jamie Miller, Owen Huang, and Pranav Rajpurkar. 2025. ReXGradient-160K: A large-scale publicly available dataset of chest radiographs with free-text reports. *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 6 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 NO-ERR. 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 NO-ERR. 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. 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:

Original report	Preprocessed report
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.
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.

Table 6: Examples of report preprocessing. 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.

> 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 7.

For exam-level evaluation, the user message contains the CXR image, the multi-sentence report, and a class-specific query template, as shown in Table 8.

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 7: 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 8: Exam-level prompt templates for zero-shot VLM baselines.