WHO-CXRBench: A CXR Benchmark for Diagnostic Rule Retrieval in CLIP-style MedVLMs

Yixiang Liu[1], Pujin Cheng[2], Zhiyuan Cai[3], Yijin Huang[4], Zhonghua Wang[5], Zhongxing Xu[5], and Xiaoying Tang[1⋆]

[1] Southern University of Science and Technology, China
[2] The University of Hong Kong, Hong Kong SAR
[3] Hong Kong University of Science and Technology, Hong Kong SAR
[4] The University of British Columbia, Canada
[5] Monash University, Australia
tangxy@sustech.edu.cn

Abstract. Medical vision-language models (MedVLMs) have progressed rapidly and demonstrate strong performance across a broad range of clinical tasks, including report generation, disease classification, and cross-modal retrieval. The success can be attributed to contrastive language-image pretraining (CLIP), which holistically aligns image and text into a shared embedding space, thereby unifying visual and linguistic representations. While these models excel on coarse-grained medical tasks, where global image-level semantics are sufficient for success, such as modality recognition, disease classification, and standard image-report retrieval, it remains challenging for them to succeed in fine-grained and evidence-driven clinical scenarios. In those settings, models are supposed to identify and retrieve specific diagnostic evidence and align it to clinically applicable criteria. Otherwise, misidentifying or overlooking diagnostic evidence may lead to clinically contradictory errors and weaken the reliability and interpretability of the final diagnoses. Therefore, rigorously quantifying CLIP's fine-grained retrieval capability in realistic clinical settings is essential. In this paper, we introduce WHO-CXRBench, a benchmark grounded in WHO chest X-ray (CXR) diagnostic guidelines for fine-grained diagnostic rule retrieval. We curate an evaluation set of 6,115 patient-level samples, resulting in 29,375 patient-rule pairs. We quantify the retrieval performance by Recall@K, Precision@K, nDCG, and mAP. Experiments show that even the best-performing baseline achieves only 14.23% Recall@10 and 13.06% nDCG@10, indicating that current CLIP-style MedVLMs still have space for improvement in fine-grained diagnostic rule retrieval. We believe that WHO-CXRBench builds a valuable foundation for evaluating MedVLMs' performance under realistic clinical settings. Code and dataset are available here.

Keywords: Vision-Language Models · Fine-grained Benchmark · Clinical Reasoning · Guideline-grounded Retrieval

⋆ Corresponding author.

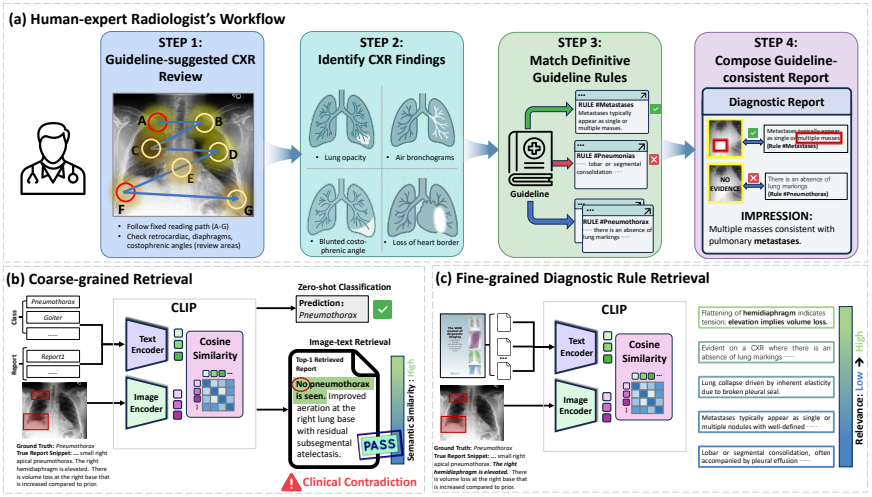

Fig. 1. Benchmark motivation. (a) Human clinician workflow; **(b)** Conventional evaluation of CLIP-style MedVLMs, where *Clinical Contradiction* means a semantically similar report that violates the actual diagnostic logic; **(c)** Fine-grained diagnostic rule retrieval task.

1 Introduction

MedVLMs have progressed rapidly and shown strong performance across a wide range of clinical tasks, including report generation, disease classification, and multimodal retrieval. Their success is largely driven by CLIP [12,20,15,16,6] that aligns image and text into a unified embedding space, enabling inference and retrieval via high-dimensional similarity. A common formulation performs predominantly global image-text matching, which can be sufficient for coarse-grained recognition such as modality identification and disease classification. Motivated by the need for finer clinical cues, recent CLIP-style MedVLMs [7,15,2] enhance contrastive pretraining with multi-granularity alignment, explicitly pairing region-level visual features with corresponding report snippets. While such designs increase cross-modal semantic similarity, they do not explicitly optimize for evidence completeness, and the retrieved evidence can still be clinically contradictory. In particular, those models remain optimized for semantic alignment rather than for verifying the presence or absence of specific diagnostic evidence. As a result, similarity-based retrieval can yield clinically contradictory outputs [8], even when the similarity scores are high (Fig. 1(b)).

Current evaluation methods for CLIP-style MedVLMs reflect a different emphasis from real-world radiology workflow [6]. Most of them emphasize multimodal recognition, abnormality or disease classification, and similar report retrieval, in which success can be achieved by learning broad correlations and superficial semantic alignments. In contrast, as shown in Fig. 1(a), real-life clinical workflows are explicitly evidence-driven and knowledge-grounded [6]. Clini-

cians systematically scan and localize findings within regions of interest, extract imaging signs that can be mapped to their prior knowledge, retrieve and align those signs to authoritative diagnostic criteria and differential diagnoses, and finally compose a medical report grounded in supportive evidence snippets plus diagnostic rules or criteria, rather than solely outputting a disease prediction. Yet existing benchmarks primarily assess the mapping from images to final reports or diagnoses, leaving a clear gap in evaluating the mapping from images to medical knowledge, namely fine-grained diagnostic rule retrieval (Fig. 1(c)). To address this gap, we introduce WHO-CXRBench, a WHO-guideline-based evaluation benchmark for fine-grained CXR diagnostic rule retrieval.

The main contributions of this paper are:

1. We introduce WHO-CXRBench, a WHO-guideline-based benchmark for fine-grained diagnostic rule retrieval in chest X-rays. The benchmark reflects an evidence-driven clinical workflow, where findings are localized, mapped to clinical concepts, and verified against diagnostic rules.
2. We introduce an automated, quality-controlled framework that converts long-form guidelines into a structured, retrieval-ready diagnostic rule corpus and generates patient-level samples paired with personalized guideline rules. To test the downstream utility of these personalized rules, we evaluate a CXR diagnostic visual question answering (VQA) task and show that personalized diagnostic rules can serve as contextual evidence to improve performance.
3. We conduct a systematic evaluation of state-of-the-art open-source CLIP-style MedVLMs (general-domain and CXR-specialist) under this fine-grained retrieval setting. Results show consistently low performance, highlighting a gap between holistic similarity matching and evidence-driven, guideline-based retrieval needed for clinical reasoning.

2 Methodology

2.1 Task Formulation

Given a patient-level CXR case $\mathcal{S} = (x, r)$ and a guideline rule corpus $\mathcal{G} = \{g_1, \ldots, g_n\}$, where x and r respectively denote the CXR image and the corresponding report, the fine-grained diagnostic rule retrieval task aims to rank rules in $\mathcal{G}$ using x as the query and return the top-k rules most relevant to the paired report r. The rule set $\mathcal{G}_{\mathcal{S}}^{+} = \{g_{\mathcal{S},1}^{+}, \ldots, g_{\mathcal{S},m}^{+}\} \subset \mathcal{G}$, where $m \ll n$, is referred to as the positive rule set for the case and is derived from r. The negative set is defined as $\mathcal{G}_{\mathcal{S}}^{-} = \mathcal{G} \setminus \mathcal{G}_{\mathcal{S}}^{+}$. Notably, the report r is never used at the retrieval time and is only used offline to derive $\mathcal{G}_{\mathcal{S}}^{+}$. Essentially, WHO-CXRBench can be used to evaluate a retriever's ability to map the image query x to its corresponding $\mathcal{G}_{\mathcal{S}}^{+}$ within $\mathcal{G}$.

2.2 WHO Rule Extraction via Prior-Guided Content RAG

We obtain an unfiltered rule corpus $\widetilde{\mathcal{G}}$ by applying retrieval-augmented generation (RAG) to summarize and extract rules from WHO guidelines [4], since

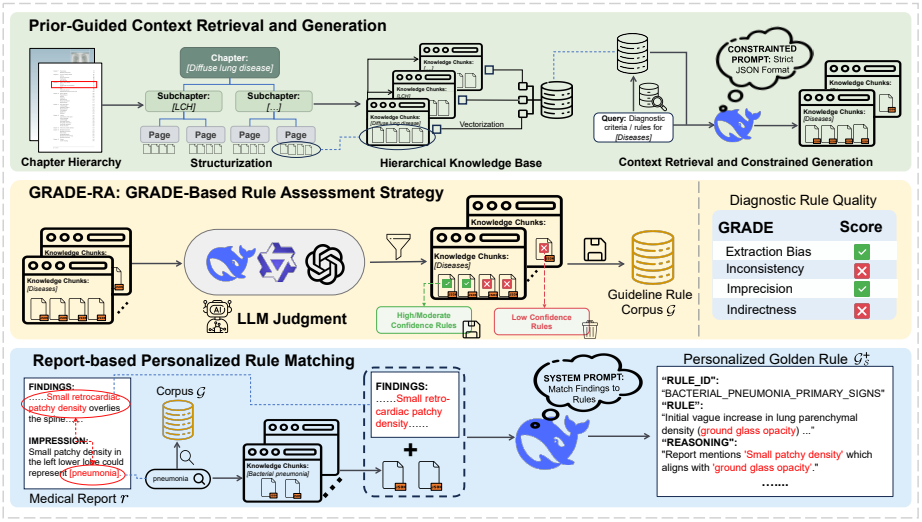

Fig. 2. Our benchmark construction pipeline. It comprises Prior-Guided Context Retrieval and Generation, GRADE-based Rule Assessment, and Report-based Personalized Rule Matching to generate patient-level personalized guideline rule samples.

the original guideline paragraphs are often too long to be processed reliably in a single pass. To improve traceability and mitigate hallucination, we adopt an outline-guided chunking strategy, named Prior-Guided Content RAG. Instead of naive fixed-length chunking, we segment guidelines according to the document outline, guided by the table of contents and heading hierarchy. The hierarchy is used purely as a structural prior derived from the guideline organization, rather than a learned or model-dependent prior. Specifically, we parse the nested chapter structure to identify disease entities and their organization, and use this hierarchy as a prior to delineate disease-level sections. We then apply fixed-size chunking within each disease-level section to obtain retrieval-ready chunks. This design enforces disease-level isolation by restricting retrieval and generation to disease-specific contexts, reducing cross-disease interference and hallucinations.

We index each disease-level chunk in a vector database and attach metadata (disease names, chapter identifiers, and line numbers). For a disease entity d, we apply a metadata filter to obtain $\mathcal{C}_d$ and run similarity search within $\mathcal{C}_d$. Then, we use a large-scale LLM to perform schema-constrained extraction to produce a traceable structured rule corpus $\widetilde{\mathcal{G}}_d$ whose extracted rules cover imaging signs, differential diagnoses, distribution patterns, and anatomical locations. We collect the extracted rules by projecting fields from each record and aggregating across diseases, defining the rule set as $\widetilde{\mathcal{G}} = \bigcup_{d \in \mathcal{D}} \widetilde{\mathcal{G}}_d$, and attach disease-specific metadata $\mathcal{C}_d$ to each rule, enabling metadata-based filtering to restrict candidate rules to the relevant disease context before similarity search.

2.3 GRADE-RA: GRADE-Based Rule Assessment

The preceding RAG step inevitably retrieves weakly relevant or even irrelevant text for CXR interpretation, which can cause the LLM to overgeneralize or hallucinate and yield noisy rules. To reduce noise in $\widetilde{\mathcal{G}}$, we propose GRADE-RA, adapting the WHO GRADE philosophy [17] to assess the operational quality of radiological rule text. We first filter the extracted candidate rules $\widetilde{g}_d$ based on their linked guideline evidence snippets. Candidates whose linked snippets lack radiologically meaningful content are directly assigned *Very Low* and removed; the remaining candidates are retained as *High*.

We then refine the certainty of these remaining candidates by applying four radiology-oriented downgrading domains. *Extraction Bias* measures faithfulness to the verbatim source snippet. *Inconsistency* measures conflicts with other rules for the same disease. *Indirectness* measures whether acquisition qualifiers, positional qualifiers, or clinical qualifiers required for CXR interpretation are missing. *Imprecision* measures hedged or vague language that weakens matchability. Each downgrading domain is scored as $s_{\mathrm{EB}}(\widetilde{g}_d, c_d)$, $s_{\mathrm{INC}}(\widetilde{g}_d, \widetilde{\mathcal{G}}_d \setminus \{\widetilde{g}_d\})$, $s_{\mathrm{IND}}(\widetilde{g}_d)$, and $s_{\mathrm{IMP}}(\widetilde{g}_d)$ in $\{-2, -1, 0\}$, where c_d denotes the set of guideline chunks that support the summarized rule $\widetilde{g}_d$ and provide traceable provenance. Notably, in GRADE-RA, the LLM only assigns per-domain scores, while the final certainty score is computed deterministically. The overall score is computed as

$$S(\widetilde{g}_d) = s_{\mathrm{EB}}(\widetilde{g}_d, c_d) + s_{\mathrm{INC}}(\widetilde{g}_d, \widetilde{\mathcal{G}}_d \setminus \{\widetilde{g}_d\}) + s_{\mathrm{IND}}(\widetilde{g}_d) + s_{\mathrm{IMP}}(\widetilde{g}_d), \tag{1}$$

and mapped to a final certainty rating

$$\mathrm{Grade}(\widetilde{g}_d) = \begin{cases} \textit{High}, & S(\widetilde{g}_d) \geq 0 \\ \textit{Moderate}, & S(\widetilde{g}_d) = -1 \\ \textit{Low}, & S(\widetilde{g}_d) = -2 \\ \textit{Very Low}, & S(\widetilde{g}_d) \leq -3. \end{cases} \tag{2}$$

This mapping follows the standard GRADE downgrading convention, where each additional concern reduces certainty by one level. We retain only *High* and *Moderate* rules to form the guideline rule corpus $\mathcal{G}$, improving the quality and interpretability for downstream alignment and retrieval evaluation.

2.4 Patient-Level Personalized Rule Alignment from Reports

We derive Personalized Golden Rules $\mathcal{G}_{\mathcal{S}}^{+}$ for each sample $\mathcal{S} = (x, r)$ by aligning the guideline rule corpus $\mathcal{G}$ with the paired report r. We extract only FINDINGS and IMPRESSION via rule-based parsing, discarding non-interpretive sections. Inspired by criteria-based LLM relevance judgments for retrieval evaluation [5], we employ an LLM judge to assess relevance between r and $\mathcal{G}$, outputting a confidence score $\mathrm{conf}(r, g) \in [0, 1]$ with a verbatim evidence snippet and a brief rationale. We treat a rule as applied if its score exceeds a fixed threshold τ (set to $\tau = 0.5$ following [13]), and define

$$\mathcal{G}_{\mathcal{S}}^{+} = \{\, g \in \mathcal{G} \mid \mathrm{conf}(r, g) \geq \tau \,\}. \tag{3}$$

Here $\mathrm{conf} : \mathcal{R} \times \mathcal{G} \to [0, 1]$ and $r \in \mathcal{R}$.

Table 1. Rule corpus statistics of WHO-CXRBench ($\widetilde{\mathcal{G}}$ vs. $\mathcal{G}$).

Corpus	#Rules	#Disease Categories	Avg. Rules per Disease	Reduction vs. $\widetilde{\mathcal{G}}$	GRADE-RA	
					High	Moderate
$\widetilde{\mathcal{G}}$	1,154	37	31.2	-	-	-
$\mathcal{G}$	300	37	8.1	−74.0% (3.85×)	153	147

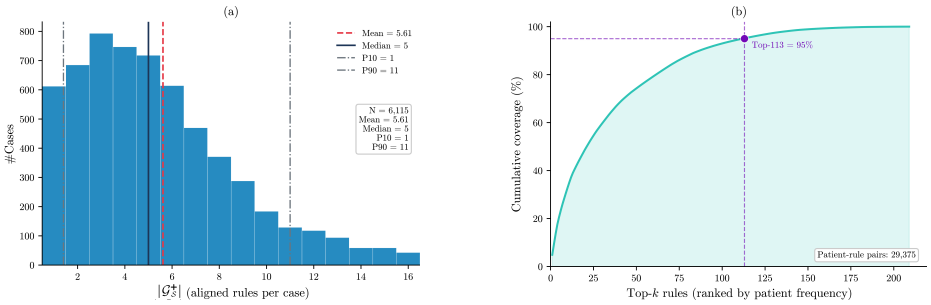

Fig. 3. Patient-level alignment statistics under $\tau = 0.5$. **(a)** Distribution of per-sample positive-rule density, measured by $|\mathcal{G}_S^+|$ (the number of aligned positive rules per sample). P10 and P90 indicate the 10th and 90th percentile values of the distribution. **(b)** Cumulative rule coverage across patients, where rules are ranked by patient frequency. Patient frequency denotes the number of cases whose aligned set $\mathcal{G}_S^+$ contains the rule.

3 Experiments and Results

3.1 Dataset and Benchmark Statistics

Dataset and Metrics. Our benchmark is built upon the WHO chest radiography manual [4] and the authoritative image-text CXR dataset MIMIC-CXR-JPG. Specifically, the rule corpus is derived exclusively from the former, while all CXR images and paired reports are sourced from the latter. Moreover, we select all entries with category "disease" in MIMIC-Ext-MIMIC-CXR-VQA, a derivative dataset extended from MIMIC-CXR-JPG, as our evaluation samples. For evaluation metrics, we report standard ranking metrics, including Recall@K, Precision@K (K=1, 5, 10), mAP, and nDCG@10. For multi-disease VQA, we report Accuracy, IoU, Micro-F1, and Macro-F1 [14].

Benchmark Statistics. In Table 1, we summarize corpus statistics before and after GRADE-RA filtering ($\widetilde{\mathcal{G}}$ vs. $\mathcal{G}$). GRADE-RA substantially reduces redundancy, shrinking the corpus from 1,154 to 300 rules (-74.0%, 3.85× smaller) while preserving coverage of all 37 disease categories. Correspondingly, the average number of rules per disease drops from 31.2 to 8.1, yielding a more concise and usable rule set for retrieval evaluation. We further report patient-level rule alignment statistics under $\tau = 0.5$. Fig. 3(a) shows the distribution of per-sample

Table 2. Performance on open-source CXR diagnosis VQA with Qwen2.5-VL-7B. BASELINE denotes zero-shot inference. No-rule SFT denotes SFT without external context. Report-ctx SFT denotes SFT with masked report context, where disease entities are hidden to prevent leakage. Rule-ctx SFT denotes SFT with patient-level personalized rules using the same entity-hiding strategy. All metrics are reported as mean ± half-width of the 95% bootstrap CI with $n_{\text{iter}} = 1000$. "$*$" indicates $p \leq 0.05$ in a paired t-test when comparing Rule-ctx SFT with Report-ctx SFT.

Method	**Metrics**		
	Acc↑	**IoU↑**	**Micro-F1↑**
BASELINE	40.05 ± 4.58	38.84 ± 4.71	51.11 ± 4.74
No-rule SFT	69.86 ± 4.34	69.53 ± 4.23	73.08 ± 3.82
Report-ctx SFT	92.23 ± 2.56	91.87 ± 2.50	92.60 ± 2.26
Rule-ctx SFT	**94.06 ± 2.16***	**93.87 ± 1.86***	**94.54 ± 1.58***

positive-rule density ($|\mathcal{G}_S^+|$), reflecting clinically realistic cases where one radiograph can align to multiple guideline-based rules. Fig. 3(b) reveals a pronounced long tail, where the Top-113 most frequent rules account for 95% of all aligned patient-rule positive pairs, indicating highly imbalanced rule occurrences. Nevertheless, retrieval models should not focus only on frequent rules, since rare but specific rules remain critical for atypical presentations and less common diseases.

3.2 Rationality Checks for Patient-Level Rule Alignment

Construction audit. We inspect 200 studies with medically trained student reviewers by checking the Top-5 highest-confidence rules in $\mathcal{G}_S^+$ and labeling each rule as supported or not based on CXR and report evidence. Overall, 92.7% of inspected rules are supported, with residual alignment errors mainly due to overly broad rule phrasing and underdocumented imaging findings, indicating high precision among high-confidence alignments and supporting WHO-CXRBench as a retrieval target under our construction protocol.

CXR Diagnosis VQA. Inspired by context learning [19] and evidence-based clinical reasoning [3], we evaluate VQA as a rationality check for the extracted patient-level personalized rules by supervised fine-tuning (SFT) through Qwen2.5-VL-7B. As shown in Table 2, Rule-ctx SFT consistently outperforms Report-ctx SFT across all metrics, indicating that personalized rules provide clinically relevant evidence and are suitable retrieval targets for downstream reasoning.

3.3 Results and Discussion

We conduct a comparison study of CLIP-style MedVLMs for fine-grained diagnostic rule retrieval, where retrieval performance is consistently low across all baselines. Table 3 reports fine-grained diagnostic rule retrieval results of representative general-purpose and CXR-specialist CLIP-style MedVLMs. The best

Table 3. Fine-grained diagnostic rule retrieval on WHO-CXRBench. MedCLIP-R50 and MedCLIP-ViT respectively denote MedCLIP with ResNet-50 and ViT visual encoders. UniMed-CLIP-B, UniMed-CLIP-L, and UniMed-CLIP-LT respectively denote UniMed-CLIP with ViT-B/16, ViT-L/14, and ViT-L/14 with a large text encoder.

Method	**Recall↑**				**Precision↑**				**mAP @10**	**nDCG @10**
	@1	**@5**	**@10**	**@20**	**@1**	**@5**	**@10**	**@20**		
General-purpose CLIP										
CLIP [12]	1.47	3.04	3.72	4.99	7.03	2.82	1.90	1.39	2.26	4.16
MedCLIP-R50 [16]	0.07	0.72	1.85	5.19	0.51	0.90	1.06	1.48	0.45	1.32
MedCLIP-ViT [16]	0.56	6.53	8.27	11.07	3.21	7.14	4.54	3.04	2.72	6.61
BioMedCLIP [18]	0.37	0.83	2.45	7.11	2.58	1.13	1.58	2.00	0.78	2.05
MedConceptCLIP [11]	0.72	3.14	6.20	12.18	3.65	3.80	3.74	3.55	2.34	5.11
UniMed-CLIP-B [9]	1.82	6.06	9.62	16.12	7.13	5.52	4.76	4.26	4.44	8.24
UniMed-CLIP-LT [9]	1.40	5.18	9.35	17.18	6.46	5.10	4.88	4.65	3.70	7.72
UniMed-CLIP-L [9]	**3.39**	**8.98**	**14.23**	**21.18**	**16.09**	**8.82**	**7.26**	**5.60**	**6.83**	**13.06**
CXR-specialist CLIP										
ConVIRT [20]	1.04	4.24	7.93	14.25	5.49	4.64	4.32	3.87	3.04	6.55
GLoRIA [7]	1.14	4.78	8.47	14.69	5.84	4.96	4.44	3.97	3.41	7.04
MGCA [15]	0.85	3.71	7.12	12.54	4.97	4.42	4.08	3.62	2.75	6.00
M-FLAG [10]	0.25	2.69	6.55	8.53	1.03	3.10	3.40	2.38	1.71	4.40
PTUnifier [1]	0.92	4.01	7.10	12.91	4.96	4.35	3.95	3.70	2.79	5.95
REFERS [21]	1.07	4.57	8.39	14.56	5.82	5.02	4.58	4.12	3.29	7.00

model attains 3.39% Recall@1 and 14.23% Recall@10, and the ground-truth rule is absent from the top-10 results for most samples, indicating that fine-grained diagnostic rule retrieval is substantially harder than conventional Med-VLM evaluation. We attribute this low retrieval performance to current CLIP-style MedVLMs' being optimized for image-report semantic alignment, whereas fine-grained diagnostic rule retrieval requires precise evidence-to-rule grounding between localized visual evidence and specific guideline rule snippets. This difficulty is further exacerbated by long, compositional diagnostic rules, a multi-label retrieval setting where each case can align with multiple applicable rules, and a large rule corpus. Notably, CXR-specialist models do not consistently outperform general-purpose models. UniMed-CLIP achieves the strongest ranking quality with 6.83% mAP@10 and 13.06% nDCG@10, surpassing the best CXR-specialist baseline GLoRIA with 3.41% mAP@10 and 7.04% nDCG@10, suggesting that broader medical-domain pretraining and stronger text encoders are advantageous for fine-grained diagnostic rule retrieval.

For architecture, visual backbone scaling has a clear impact on performance. Replacing the ResNet-50 visual encoder in MedCLIP with a ViT backbone increases Recall@10 from 1.85% to 8.27%, while scaling UniMed-CLIP's visual encoder from ViT-B/16 to ViT-L/14 increases Recall@10 from 9.62% to 14.23%. In contrast, scaling UniMed-CLIP's text encoder decreases Recall@10 to 9.35%, suggesting that visual grounding matters more than generic text-side scaling.

For failure analysis, errors are non-uniform across diseases and rules. Disease-level failures are more frequent in categories with overlapping CXR patterns, such as pneumonia and lung neoplasms, where opacity, consolidation, or infiltrates can overlap and qualifiers such as laterality, distribution, and anatomical location become decisive. At the rule level, longer and more compositional rules are harder to retrieve: rules in the upper length quartile ($\geq$23 words) achieve only 7.3% Recall@10, compared with 18.2% for shorter rules. This indicates that the benchmark difficulty comes not only from corpus size but also from fine-grained visual-rule composition

4 Conclusion

This paper introduces WHO-CXRBench, a WHO guideline-grounded benchmark for patient-level fine-grained diagnostic rule retrieval in CXRs. Extensive experiments show current CLIP-style MedVLMs achieve limited retrieval performance in this clinically realistic, evidence-driven setting, highlighting clear needs for improved fine-grained retrieval and interpretability. WHO-CXRBench further enables practical evaluation of guideline-driven, evidence-based behavior by requiring models to retrieve clinically applicable WHO rules at the patient level.

Acknowledgments. This study was supported by the National Key Research and Development Program of China (2023YFC2415400); the National Natural Science Foundation of China (T2422012); the Guangdong Basic and Applied Basic Research (2024B1515020088); the High Level of Special Funds (G030230001, G03034K003); the Guangdong Key Research and Development Program (2025B1111080001); the SUSTech Fang Keng Faculty Award.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Chen, Z., Diao, S., Wang, B., et al.: Towards unifying medical vision-and-language pre-training via soft prompts. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23403–23413 (2023)
2. Cheng, P., Lin, L., Lyu, J., et al.: Prior: Prototype representation joint learning from medical images and reports. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21361–21371 (2023)
3. Chloros, G.D., Prodromidis, A.D., Giannoudis, P.V.: Has anything changed in evidence-based medicine? Injury **54**, S20–S25 (2023)
4. Ellis, S.M., Flower, C.: The WHO manual of diagnostic imaging: radiographic anatomy and interpretation of the chest and the pulmonary system. World Health Organization (2012)
5. Farzi, N., Dietz, L.: Criteria-based llm relevance judgments. In: Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR). pp. 254–263 (2025)

6. Gu, D., Gao, Y., Zhou, M., Metaxas, D.: Anatomy-vlm: A fine-grained vision-language model for medical interpretation. arXiv preprint arXiv:2511.08402 (2025)
7. Huang, S.C., Shen, L., Lungren, M.P., et al.: GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3942–3951 (2021)
8. Irvin, J., Rajpurkar, P., Ko, M., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 590–597 (2019). https://doi.org/10.1609/aaai.v33i01.3301590
9. Khattak, M.U., Kunhimon, S., Naseer, M., et al.: UniMed-CLIP: Towards a unified image-text pretraining paradigm for diverse medical imaging modalities (2024)
10. Liu, C., Cheng, S., Chen, C., Qiao, M., Zhang, W., Shah, A., Bai, W., Arcucci, R.: M-flag: Medical vision-language pre-training with frozen language models and latent space geometry optimization. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 637–647. Springer (2023)
11. Nie, Y., He, S., Bie, Y., et al.: An explainable biomedical foundation model via large-scale concept-enhanced vision-language pre-training (2025). https://doi.org/10.48550/arXiv.2501.15579
12. Radford, A., Kim, J.W., Hallacy, C., et al.: Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (2021)
13. Ran, K., Sun, S., Anh, K.N.D., Spina, D., Zendel, O.: Rmit-adm+s at the sigir 2025 liverag challenge. arXiv preprint arXiv:2506.14516 (2025)
14. Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Information Processing & Management **45**(4), 427–437 (2009). https://doi.org/10.1016/j.ipm.2009.03.002
15. Wang, F., Zhou, Y., Wang, S., et al.: Multi-granularity cross-modal alignment for generalized medical visual representation learning. In: Advances in Neural Information Processing Systems. vol. 35, pp. 33536–33549 (2022)
16. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: MedCLIP: Contrastive learning from unpaired medical images and text. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 3876–3887. Association for Computational Linguistics (2022)
17. World Health Organization: WHO handbook for guideline development. World Health Organization (2014)
18. Zhang, S., Xu, Y., Usuyama, N., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs (2023). https://doi.org/10.48550/arXiv.2303.00915, arXiv preprint, v3 last revised Jan 2025
19. Zhang, T., Patil, S.G., Jain, N., Shen, S., Zaharia, M., Stoica, I., Gonzalez, J.E.: Raft: Adapting language model to domain specific rag. arXiv preprint arXiv:2403.10131 (2024)
20. Zhang, Y., Jiang, H., Miura, Y., et al.: Contrastive learning of medical visual representations from paired images and text. In: Proceedings of Machine Learning for Healthcare Conference. pp. 2–25. PMLR (2022)
21. Zhou, H., Chen, X., Zhang, Y., et al.: Generalized radiograph representation learning via cross-supervision between images and free-text radiology reports. Nature Machine Intelligence **4**(1), 32–40 (2022). https://doi.org/10.1038/s42256-021-00425-9