DCLDs-RAG: Evidence-Grounded Multimodal Retrieval-Augmented Generation for Diffuse Cystic Lung Diseases Diagnosis

Haoqing Li[1][0000−0001−9126−5079], Jun Shi[1⋆], Xianmeng Chen[2], Qiwei Jia[3], Rui Wang[4], Yating Peng[5], Li Gao[5], Qianyi Chen[5], Hong An[1], Ruoyun Ouyang[5], and Xiaowen Hu[2]

[1] School of Computer Science and Technology, University of Science and Technology of China, Hefei, China
li_haoqing@mail.ustc.edu.cn, shijun18@ustc.edu.cn
[2] Department of Pulmonary and Critical Care Medicine; Center for Diagnosis and Management of Rare Diseases, the First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, USTC, Hefei, China
[3] School of the Gifted Young, USTC, Hefei, China
[4] Department of Respiratory Medicine, No. 901 Hospital of the Chinese People's Liberation Army Logistic Support Force, Hefei, China
[5] Department of Pulmonary and Critical Care Medicine, The Second Xiangya Hospital, Central South University, Changsha, China

Abstract. Deep learning methods face dual challenges of limited clinical samples and subtle inter-class differences among Diffuse Cystic Lung Diseases (DCLDs) when performing multi-class diagnosis from Computed Tomography (CT) imaging. Although Multimodal Large Language Models (MLLMs) exhibit promising potential for rare disease diagnosis, their performance is often compromised by insufficient domain-specific knowledge and the absence of reliable multimodal evidence, leading to increased hallucination risks. To address these limitations, we propose DCLDs-RAG, a multimodal retrieval-augmented generation framework for DCLDs diagnosis. Specifically, DCLDs-RAG integrates: (1) a disease-specific multimodal corpus of DCLDs cases and a DCLDs knowledge graph corpus; (2) a visual evidence retrieval learns angular margin-aware representations with high diagnostic discriminability; and (3) a knowledge–clinical evidence retrieval extracts disease-specific entities and relational patterns from the DCLDs knowledge graph as conclusive evidence to support diagnostic reasoning. An MLLM synthesizes retrieved evidence with a query for the final diagnosis. DCLDs-RAG is validated on the multi-center cohorts that encompass four types of DCLDs, achieving superior accuracy and generating evidence-based descriptions closely aligned with expert insights. Code is available at: https://github.com/Li-Haoqing/DCLDs.

Keywords: Diffuse Cystic Lung Diseases · Retrieval-Augmented Generation · Multimodal Large Model · Computed Tomography.

⋆ Corresponding author.

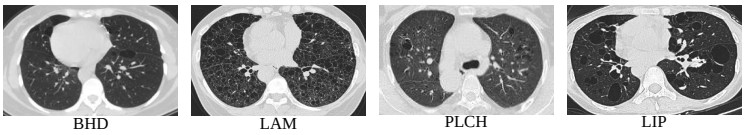

Fig. 1. Representative CT slices examples of four challenging-to-differentiate DCLDs.

1 Introduction

Diffuse Cystic Lung Diseases (DCLDs) comprise a group of rare pulmonary disorders characterized by diffuse cystic on Computed Tomography (CT) [23,11]. Despite distinct etiologies and clinical implications, DCLDs, including Birt-Hogg-Dubé syndrome (BHD), Lymphangioleiomyomatosis (LAM), Pulmonary Langerhans Cell Histiocytosis (PLCH), and Lymphocytic Interstitial Pneumonia (LIP), often exhibit highly overlapping radiological appearances, resulting in subtle inter-class differences and frequent diagnostic ambiguity [11,9].

While deep learning paradigms have reached expert-level diagnostic performance [21,25,8], most task-specific models remain heavily predicated on large-scale annotations, resulting in poor generalization to rare pathologies like DCLDs. This challenge is compounded by the substantial inter-class visual similarity in DCLD manifestations, which hinders discriminative models from establishing robust decision boundaries. Concurrently, the advent of In-Context Learning (ICL) [5,20,27,6] has empowered Large Language Models (LLMs) to excel in data-limited medical regimes. Notably, general-purpose models like GPT-4 have demonstrated the capacity to outperform domain-specific systems (e.g., Med-PaLM 2) even without task-specific fine-tuning [24,26,1]. These breakthroughs underscore the potential of Multimodal Large Language Models (MLLMs) as a promising frontier for diagnosing low-resource DCLDs. However, MLLMs are susceptible to hallucinations, yielding fluent yet incorrect content [15]. This limitation is amplified in the context of DCLDs due to the lack of specialized medical knowledge and referential radiological evidence, as illustrated in Fig. 6.

Retrieval-Augmented Generation (RAG) integrates information retrieval and text generation, leveraging external corpus to enhance MLLMs in knowledge-intensive tasks [7]. As a typically knowledge-intensive field, RAG is capable of providing domain-specific knowledge and evidence on DCLDs, addressing the hallucination inherent in MLLMs [29,12]. However, the rarity of DCLDs and the limited availability of publicly accessible resources severely constrain the construction of multimodal corpus. Under such conditions, existing RAG approaches predominantly rely on generic textual retrieval, which often introduces substantial irrelevant or noisy information, rather than providing logically grounded diagnostic evidence capable of supporting reliable medical reasoning [22,19,32].

To address these challenges, we propose DCLDs-RAG, an evidence-grounded multimodal retrieval-augmented generation framework tailored for DCLDs diagnosis. As shown in Fig. 2, in response to the scarcity of external knowledge to construct the RAG system, we sourced clinical cases of DCLDs from the respiratory

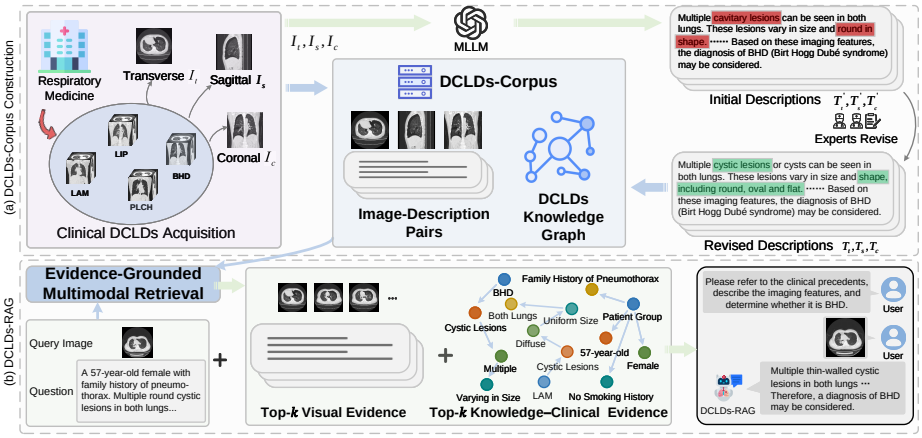

Fig. 2. An overview of the proposed DCLDs-RAG framework.

department to establish a dedicated DCLD-Corpus. Unlike prior RAG methods that passively inject retrieved text, DCLDs-RAG explicitly models diagnostic evidence from two complementary perspectives. First, to overcome the challenge of optimizing the decision boundary among DCLDs, a visual evidence retriever expands the angular margin and extract pertinent pairs from the corpus. Second, a knowledge–clinical evidence retriever systematically mines disease-specific entities and their relational patterns from a DCLDs knowledge graph, transforming them into structured clinical evidence that explicitly constrains and informs the diagnostic reasoning process. DCLDs-RAG is validated on the multi-center cohorts encompassing four types of DCLDs, demonstrating superior accuracy and generating evidence-based descriptions consistent with expert assessments.

2 Methodology

2.1 DCLDs-Corpus and Knowledge Graph Construction

To support DCLDs-RAG, we construct a multimodal DCLDs-corpus and a disease-specific knowledge graph, serving as the retrieval source for both visual, knowledge and clinical evidence.

DCLDs Image–Description Corpus. It is essential to gather extensive image-description paired examples. We extract representative CT slices from three views, i.e., coronal, sagittal, and transverse, denoted as $\{I_t, I_s, I_c\}$. Each slice is required to contain diagnostically relevant cystic lesions. Moreover, GPT-4-turbo [1] is employed to generate initial radiological descriptions, which are subsequently reviewed and refined by respiratory medicine experts to ensure diagnostic accuracy and logical consistency, as shown in Fig. 2 (a). The descriptions $\{T_t, T_s, T_c\}$ generated by experts-assisted MLLM are paired with the corresponding images to construct the corpus $\{\mathcal{I}_{corpus}, \mathcal{T}_{corpus}\} = \{(I_t, T_t), (I_s, T_s), (I_c, T_c)\}$.

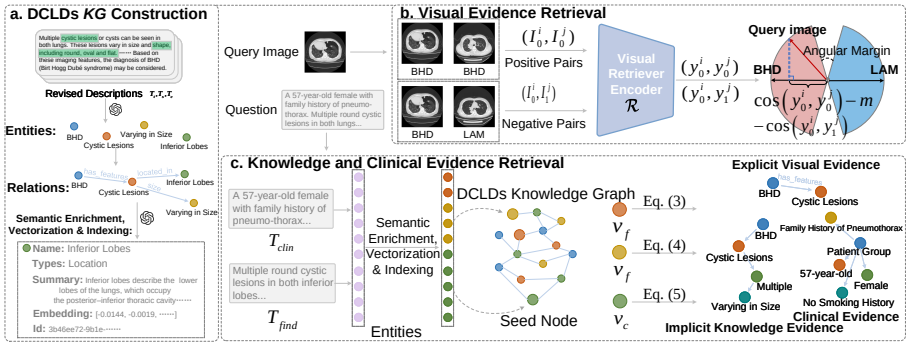

Fig. 3. The proposed Evidence-Grounded Multimodal Retriever. DCLDs-RAG grounds MLLM reasoning on explicit diagnostic evidence from two complementary sources. DCLDs-RAG retrieves visual evidence from the disease-specific multimodal corpus and knowledge-clinical evidence from the DCLDs knowledge graph.

DCLDs-corpus is designed to emphasize distinctive DCLDs features, providing fine-grained comparative analyses to reduce domain knowledge ambiguity.

DCLDs Knowledge Graph Corpus. We construct a DCLDs knowledge graph corpus to provide structured knowledge and clinical evidence for reasoning. Built from expert-refined descriptions, the graph ensures all elements are clinically grounded. Disease-related concepts are normalized into entities $e \in E$, covering DCLDs subtypes, lesion characteristics, anatomical location and clinical information, etc. To support semantic retrieval, each entity is encoded into a dense embedding $\mathbf{v}$. Based on diagnostic semantics within the same description context, typed relations $r \in R$ are established between entity pairs, where each relation is represented as a triple (e_h, r, e_t), where e_h and e_t denote the head and tail entities. Formally, the DCLDs knowledge graph is defined as:

$$\mathcal{G} = \{(e_j, \mathbf{v}_j)_{j=1}^{|E|}\} \cup \{(e_h, r, e_t) | e_h, e_t \in E, r \in R\}, \tag{1}$$

2.2 Evidence-Grounded Multimodal RAG

Visual Evidence Retrieval with Angular Margin Learning Visual evidence retrieval aims to identify the minimal corpus subset that exhibits imaging phenotypes consistent with the query [4]. However, due to the substantial visual overlap among different DCLDs subtypes, vanilla cosine similarity strategies fail to reliably disentangle fine-grained patterns in the embedding space.

We formulate visual retrieval as a metric learning with angular margin [16,18], as illustrated in Fig. 3 (b). Both positive and negative samples are matched within the same anatomical view. Positive pairs, e.g. (I_c^i, I_c^j), are constructed by randomly sampling different CT slices from the same class, while negative pairs, e.g. $(I_c^i, I_{c'}^j)$, are generated from different classes. The embedding y_c^i is the output of the visual retriever encoder $\mathcal{R}$.

Due to the substantial overlap of imaging features among the DCLDs subtypes, conventional approaches based on cosine similarity and Softmax struggle to achieve precise discrimination, as evidenced in Table 1. To explicitly enhance inter-class separability in the angular embedding space, we adopt an angular margin learning objective that enforces stricter decision boundaries between visually similar diseases:

$$\mathcal{L} = \frac{1}{N} \sum_{i} -\log \frac{\exp\left(s\left(\cos(\mathbf{y}_c^i, \mathbf{y}_c^j) - m\right)\right)}{\exp\left(s\left(\cos(\mathbf{y}_c^i, \mathbf{y}_c^j) - m\right)\right) + \sum_{c' \neq c} \exp\left(s \cos(\mathbf{y}_c^i, \mathbf{y}_{c'}^j)\right)} \quad (2)$$

where N denotes the number of pairs, $cos(.)$ is the cosine similarity, s is the scaling factor, and m is the angular margin. This objective enforces compact intra-class visual evidence while explicitly increasing the angular separation between DCLDs embeddings, resulting in discriminative visual retrieval space [28].

Knowledge-Clinical Evidence Retrieval We propose a knowledge–clinical evidence retriever that decomposes diagnostic support into explicit visual evidence, implicit knowledge evidence, and clinical evidence, enabling fine-grained and interpretable reasoning. We initialize our retrieval by aligning the entities from imaging T_{find} and clinical information T_{clin} into a set of unified seed nodes $v_{seed} = \{v_f, v_c\}$. This serves as the anchor for distilling diagnostic evidence through three mutually reinforcing scoring functions.

A key challenge in medical synthesis is the prevalence of non-informative commonalities. To suppress non-specific findings and emphasize disease-discriminative nodes, we introduce specific paths score for prioritizing high-specificity imaging phenotypes. **Explicit visual evidence** is quantified by:

$$S_v(v_f) = \frac{\alpha}{\sqrt{\max\left(|\mathcal{N}_{\text{dis}}(v_f)|,\, 1\right)}}, \quad (3)$$

where $\mathcal{N}_{\text{dis}}(v_f)$ denotes the set of diseases directly adjacent to v_f, α is the weight. By assigning higher scores S to paths conveying rare but highly diagnostic information, we identify a subset of paths p_f that represent explicit visual evidence $\mathcal{E}_f(p_f, \mathcal{C}_f)$, where $\mathcal{C}$ parameterizes the confidence.

Implicit diagnostic knowledge is encoded in the multi-hop relational structure of the DCLDs knowledge graph, capturing latent medical associations that are not explicitly stated in imaging. **Implicit knowledge evidence** $\mathcal{E}_k(p_k, \mathcal{C}_k)$ is obtained via constrained graph traversal. Starting from the seeded entity set V_{seed}, we enumerate candidate paths that connect imaging-derived entities to disease nodes within a limited hop depth. Each candidate path p_k is evaluated by a topology-aware scoring function to:

$$S_k(p_k) = \left(w_1 \bar{C}_E + w_2 \bar{C}_D + w_3 \bar{C}_B + w_4 \bar{C}_T\right) \cdot \frac{1}{1 + \gamma \cdot \text{len}(p_k)} \quad (4)$$

where $\bar{C}_E, \bar{C}_D, \bar{C}_B, \bar{C}_T$ denote the average eigenvector, degree, betweenness, and triangle-based centrality of nodes along path p_k, respectively. The coefficients

w control the contribution of different topological properties, while γ is the coefficient for penalizing long inference chains.

Besides structural knowledge, **clinical evidence** $\mathcal{E}_c(p_c, \mathcal{C}_c)$ (medical regularities and patient-specific clinical attributes) is essential for phenotypically similar DCLDs diagnosis. We leverage *PatientGroup* nodes v_c, which anchor canonical clinical profiles H_c (encompassing genetic, smoking, and medical histories, etc.) within the graph. The clinical matching score quantifies the fitness:

$$S_c(v_c) = \lambda_1 \mathbb{I}(\text{sex}) + \lambda_2 \mathbb{I}(\text{age}) + \lambda_3 \left| H_{\text{input}} \cap H_c \right|, \tag{5}$$

where $\mathbb{I}(\text{sex})$ and $\mathbb{I}(\text{age})$ are indicator functions evaluating sex and age compatibility, and the H term emphasizes high-specificity clinical cues. This transforms unstructured clinical descriptions into explicit and computable constraints.

Retrieval-Augmented Generation for DCLDs Diagnosis The top-k image-description pairs $\{\mathcal{I}_{topk}, \mathcal{T}_{topk}\}$ retrieved from the corpus are combined with evidence $\mathcal{E} = (\mathcal{E}_f, \mathcal{E}_k, \mathcal{E}_c)$. The objective is to diagnose DCLDs from the query $\mathcal{Q} = (I_q, T_q)$ and generate an accurate response $\tilde{A}$:

$$\{\mathcal{I}_{topk}, \mathcal{T}_{topk}\} = \underset{\substack{\mathcal{I}_{topk} \in \mathcal{I}_{corpus} \\ |\mathcal{I}_{topk}| = k}}{\arg\max} \left\{ \cos\left(\mathcal{R}(I_q, \mathcal{I}^i_{corpus})\right) \right\}, \; i = 0, \ldots, n-1, \tag{6}$$

$$\tilde{A} = f(\mathcal{Q}; \mathcal{M}, \mathcal{I}_{corpus}, \mathcal{T}_{corpus}) = \arg\max_{A} \mathbb{P}_{\mathcal{M}}(A \mid \mathcal{Q}, \mathcal{I}_{topk}, \mathcal{T}_{topk}, \mathcal{E}), \tag{7}$$

where $f(.)$ is the proposed DCLDs-RAG, $\mathcal{M}$ denotes the MLLM, $\tilde{A}$ and A are the predicted response and ground truth, respectively.

3 Experiments and Results

Data Curation and Implementation Details CT scans with confirmed DCLDs diagnoses were collected from The First Affiliated Hospital of USTC, following histopathological or guideline-based criteria, comprising 141 cases (74 BHD, 25 LAM, 8 PLCH, and 34 LIP; median slice thickness 1.25 mm). These data are split at patient-level into train / corpus (78 cases) and internal test set (63 cases). An external test cohort from The Second Xiangya Hospital includes 80 cases (14 BHD, 33 LAM, 15 PLCH, and 18 LIP). Representative slices containing typical lesions were selected. The corpus consists of 753, 255, 107, and 329 slices for BHD, LAM, PLCH, and LIP, respectively, each paired with a corresponding textual description.

To ensure the response efficiency, ResNet [10] retriever is trained for 500 epochs using AdamW (learning rate 1.0×10^{-4}, batch size 32). Images are lung-window normalized and resized to 256×256. All experiments were implemented in PyTorch and conducted on 8 NVIDIA Tesla A100 80 GB GPUs.

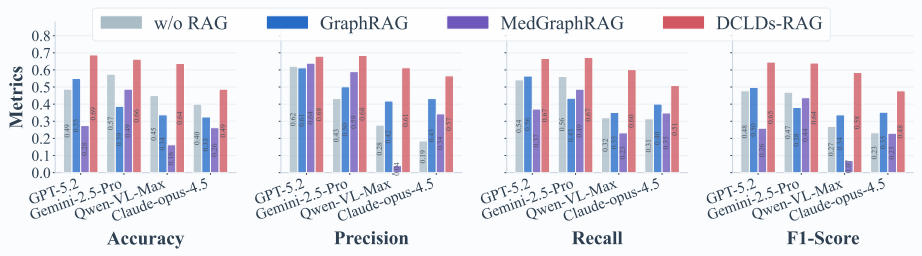

Fig. 4. Performance Comparison of three RAG methods across four MLLMs on the external test set (four-Class classification).

Table 1. Comparison of four-class DCLDs classification performance (%) on internal and external test sets. **Bold** text indicates the best results.

Method	Internal Test				External Test			
	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1
ResNet-50 3D [10]	42.86	32.64	26.35	25.63	28.75	38.81	32.55	22.95
DenseNet-121 3D [13]	49.21	24.70	26.70	24.40	28.75	24.41	29.27	23.79
ResNeXt-50 3D [30]	53.97	25.84	32.69	28.77	37.50	27.61	34.51	28.82
M3T [14]	47.62	14.71	22.06	17.65	20.00	17.83	27.70	16.41
MedicalNet-50 [2]	52.38	37.36	37.63	36.65	22.50	21.37	24.22	20.10
LLaVA-Med [17]	53.97	13.49	25.00	17.53	17.50	4.38	25.00	7.45
Ours (Qwen-VL-Max [31])	85.71	87.41	80.08	80.68	63.75	61.25	60.17	58.48
Ours (Claude-opus-4.5)	90.48	91.15	83.82	84.82	48.75	56.53	50.85	47.77
Ours (GPT-5.2)	92.06	95.29	92.19	92.89	**68.75**	68.00	66.72	**64.57**
Ours (Gemini-2.5-Pro [3])	**95.24**	**96.53**	**95.31**	**95.61**	66.25	**68.44**	**67.31**	64.02

Qualitative and Quantitative Results As summarized in Table 1 and Fig. 4, DCLDs-RAG consistently outperforms all the baselines on both internal and external test sets. On the internal cohort, DCLDs-RAG (Gemini-2.5-Pro) achieves the best performance (95.24 % accuracy, 95.61 % F1-score), substantially surpassing the strongest discriminative model MedicalNet-50. On the external test set, where domain shift is present, DCLDs-RAG maintains clear superiority over all baselines. While standard 3D models exhibit severe performance degradation, DCLDs-RAG preserves robust accuracy and F1-score, demonstrating strong generalization enabled by explicit visual and knowledge–clinical evidence grounding.

Fig. 5 (a) illustrates the one-vs-rest binary classification performance for BHD, LAM, PLCH, and LIP. Across four tasks, DCLDs-RAG consistently achieves more balanced performance than baselines. These results indicate that evidence-grounded DCLDs-RAG integration improves class-wise robustness and mitigates prediction bias induced by overlapping cystic patterns. Fig. 6 illustrates that DCLDs-RAG overcomes baseline hallucinations to achieve accurate diagnosis via robust evidence grounding. This qualitative performance corroborates the improvements in both accuracy and reliability.

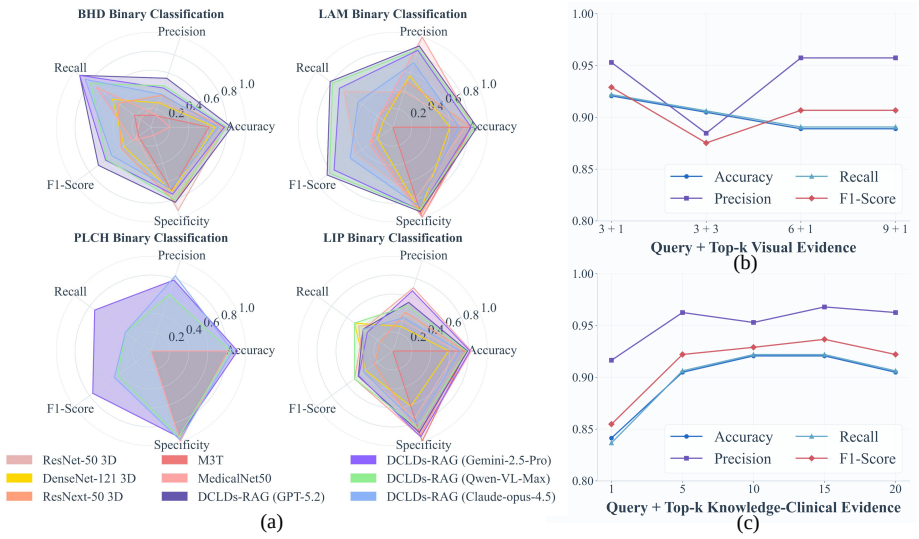

Fig. 5. (a) Comparison of four binary classification on the external test set, (b) influence of top-k visual evidence and (c) knowledge-clinical on DCLDs-RAG.

Table 2. Ablation study of retrievals on four-class DCLDs classification.

Visual Evidence	K-C Evidence	Acc. (%)	Prec. (%)	Rec. (%)	F1 (%)
✘ (Disabled)	✘	63.49	42.28	53.48	44.69
✓ (Enabled)	✘	68.25	53.37	57.22	54.19
✘	✓	90.48	**96.25**	90.62	92.20
✓	✓	**92.06**	95.29	**92.19**	**92.89**

Ablation Study Table 2 presents an ablation study analyzing the contributions of visual evidence and knowledge–clinical (K-C) evidence. Removing either component leads to a noticeable performance drop, while discarding both causes a substantial degradation (from 92.06 % to 63.49 % accuracy). Notably, using K-C evidence alone yields a significantly larger improvement than visual evidence alone, underscoring the importance of implicit knowledge reasoning and clinical population priors for DCLDs differentiation. The full approach integrates both evidence sources, achieves the best performance, confirming that visual cues and knowledge–clinical reasoning are complementary rather than redundant. These results highlight that robust diagnosis of rare lung diseases with subtle visual differences necessitates evidence-grounded reasoning that integrates visual similarity with structured medical knowledge and clinical context.

4 Conclusion

DCLDs-RAG supports evidence-grounded MLLM diagnosis of DCLDs by integrating visual evidence, clinical knowledge. Multi-center evaluations show im-

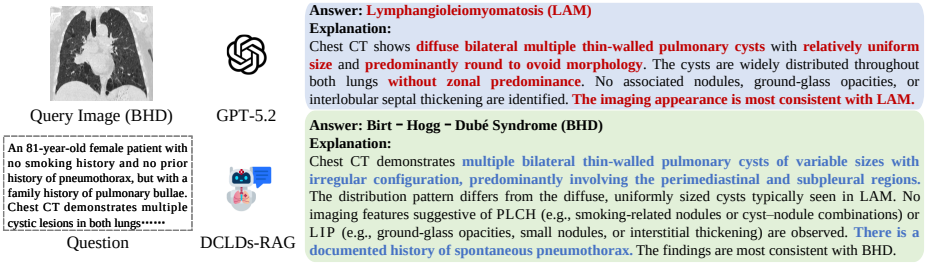

Fig. 6. Qualitative comparison of DCLDs-RAG and GPT-5.2 diagnoses. Errors and hallucinations are highlighted in red, while evidence-grounded diagnosis are in blue.

proved accuracy, interpretability, and hallucination control. Future work will focus on disease-aware retrieval policies that adaptively balance visual and clinical evidence under diagnostic ambiguity.

Acknowledgments. This work was supported in part by the Xiaomi Young Scholars Program.

Disclosure of Interests. The authors declare no competing interests for this paper.

References

1. Achiam, J., Adler, S., Agarwal, et al.: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Chen, S., Ma, K., Zheng, Y.: Med3D: Transfer learning for 3d medical image analysis. arXiv preprint arXiv:1904.00625 (2019)
3. Comanici, G., Bieber, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
4. Cuconasu, F., Trappolini, G., Siciliano, F., Filice, S., Campagnano, C., Maarek, Y., Tonellotto, N., Silvestri, F.: The power of noise: Redefining retrieval for RAG systems. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 719–729 (2024)
5. Dai, D., Sun, Y., Dong, L., Hao, Y., Ma, S., Sui, Z., Wei, F.: Why can GPT learn in-context? language models secretly perform gradient descent as meta-optimizers. In: Findings of the Association for Computational Linguistics: ACL 2023. pp. 4005–4019 (2023)
6. Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., Xia, H., Xu, J., Wu, Z., Liu, T., et al.: A survey on in-context learning. arXiv preprint arXiv:2301.00234 (2022)
7. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, H.: Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997 (2023)
8. Goh, E., Gallo, R., et al.: Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Network Open **7**(10), e2440969–e2440969 (2024)

9. Group, E.C.: Expert consensus on the diagnosis and management of birt-hogg-dubé syndrome. Chinese Journal of Tuberculosis and Respiratory Diseases **46**(9), 897–908 (2023)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
11. Hu, D., Wang, R., Liu, J., Chen, X., Jiang, X., Xiao, J., Ryu, J.H., Hu, X.: Clinical and genetic characteristics of 100 consecutive patients with birt-hogg-dubé syndrome in eastern chinese region. Orphanet Journal of Rare Diseases **19**(1), 348 (2024)
12. Hu, Y., Lei, Z., Zhang, Z., Pan, B., Ling, C., Zhao, L.: GRAG: Graph retrieval-augmented generation. In: Findings of the Association for Computational Linguistics: NAACL 2025. pp. 4145–4157 (2025)
13. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4700–4708 (2017)
14. Jang, J., Hwang, D.: M3T: three-dimensional medical image classifier using multi-plane and multi-slice transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20718–20729 (2022)
15. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A., Fung, P.: Survey of hallucination in natural language generation. ACM Computing Surveys **55**(12), 1–38 (2023)
16. Kaya, M., Bilge, H.Ş.: Deep metric learning: A survey. Symmetry **11**(9), 1066 (2019)
17. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36**, 28541–28564 (2023)
18. Li, X., Yang, X., Ma, Z., Xue, J.H.: Deep metric learning for few-shot image classification: A review of recent developments. Pattern Recognition **138**, 109381 (2023)
19. Liu, N.F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., Liang, P.: Lost in the middle: How language models use long contexts. Transactions of the Association for Computational Linguistics **12**, 157–173 (2024)
20. Liu, Y., Liu, J., Shi, X., Cheng, Q., Huang, Y., Lu, W.: Let's learn step by step: Enhancing in-context learning ability with curriculum learning. arXiv preprint arXiv:2402.10738 (2024)
21. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
22. Mao, K., Liu, Z., Qian, H., Mo, F., Deng, C., Dou, Z.: RAG-studio: Towards in-domain adaptation of retrieval augmented generation through self-alignment. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 725–735 (2024)
23. Menko, F.H., Van Steensel, M.A., Giraud, S., Friis-Hansen, L., Richard, S., Ungari, S., Nordenskjöld, M., vO Hansen, T., Solly, J., Maher, E.R.: Birt-hogg-dubé syndrome: diagnosis and management. The Lancet Oncology **10**(12), 1199–1206 (2009)
24. Nori, H., Lee, Y.T., et al.: Can generalist foundation models outcompete special-purpose tuning? case study in medicine. arXiv preprint arXiv:2311.16452 (2023)
25. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. pp. 234–241 (2015)

26. Singhal, K., Tu, T., et al.: Toward expert-level medical question answering with large language models. Nature Medicine **31**(3), 943–950 (2025)
27. Wang, F., Shao, P., Zhang, Y., Yu, B., Liu, S., Ding, N., Cao, Y., Kang, Y., Wang, H.: OmniRL: In-context reinforcement learning by large-scale meta-training in randomized worlds. arXiv preprint arXiv:2502.02869 (2025)
28. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5265–5274 (2018)
29. Wu, J., Zhu, J., Qi, Y., Chen, J., Xu, M., Menolascina, F., Jin, Y., Grau, V.: Medical graph RAG: Evidence-based medical large language model via graph retrieval-augmented generation. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics. pp. 28443–28467 (2025)
30. Xie, S., Girshick, R., Dollar, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (July 2017)
31. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
32. Yoran, O., Wolfson, T., Ram, O., Berant, J.: Making retrieval-augmented language models robust to irrelevant context. In: The Twelfth International Conference on Learning Representations (2024)