Multimodal Large Language Model-Driven Self-Verification Reasoning for Interpretable Diagnosis of Diffuse Cystic Lung Diseases

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

[1] School of the Gifted Young, University of Science and Technology of China, Hefei, China

[2] School of Computer Science and Technology, University of Science and Technology of China, Hefei, China
shijun18@ustc.edu.cn

[3] 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

[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. Diffuse Cystic Lung Diseases (DCLDs) comprise a group of rare pulmonary disorders with heterogeneous morphological patterns and overlapping imaging manifestations, posing a significant challenge for differential diagnosis. Although Multimodal Large Language Models (MLLMs) have shown promising zero-shot capabilities for DCLDs recognition, their limited context length prevents them from processing the entire CT scan, and the resulting incomplete spatial perception often leads to prediction hallucinations, characterized by poor accuracy and interpretability. To address these problems, we propose an interpretable MLLM-driven diagnostic reasoning framework for DCLDs inspired by the clinical workflow. To enable effective utilization of CT images, we design a **Global Spatial Perception (GSP)** strategy that projects critical 3D lesion characteristics into compact 2D representations, preserving essential spatial distribution and morphology information while substantially reducing input context. Furthermore, we introduce a **Multistep Self-Verification Reasoning (MSVR)** mechanism that guides the model to identify and validate discriminative visual evidence from representative axial slices, iteratively refining diagnostic probabilities and enhancing interpretability. We construct a large-scale, multi-center dataset comprising diverse DCLDs subtypes and conduct extensive experimental evaluations. Our method achieves a diagnostic accuracy of **82.5%**, significantly outperforming state-of-the-art approaches, while ablation studies confirm the contribution of each key component. These results demonstrate the feasibility and clinical

⋆ Corresponding author.

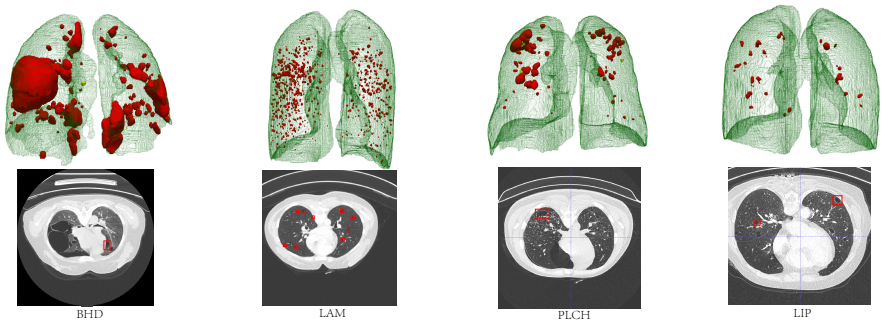

Fig. 1. Representative 3D renderings and axial slices of four DCLDs subtypes.

potential of MLLMs for reliable DCLDs diagnosis. Code available at `https://github.com/JQW2006/dcld-gsp-msvr`.

Keywords: Diffuse Cystic Lung Disease · Multimodal Large Language Model · Self-Verification Reasoning.

1 Introduction

Diffuse cystic lung diseases (DCLDs) constitute a spectrum of rare pulmonary disorders characterized by multiple cystic lesions within the lung parenchyma, primarily encompassing Birt-Hogg-Dubé Syndrome (BHD), Lymphangioleiomyomatosis (LAM), Pulmonary Langerhans Cell Histiocytosis (PLCH), and Lymphocytic Interstitial Pneumonia (LIP) [1–3]. Despite distinct pathogenesis, these subtypes exhibit highly similar cystic appearances on chest CT, rendering differential diagnosis exceedingly difficult [4]. Accurate diagnosis critically depends on cross-scale imaging analysis: macroscopic three-dimensional spatial distribution of cysts throughout the lung, and microscopic morphological features on individual slices, including wall thickness, shape, and relationships with bronchovascular bundles [5, 1]. This simultaneous demand for global spatial reasoning and local morphological discrimination renders DCLDs one of the most challenging differential diagnostic tasks in radiology (Fig. 1).

Recent years have witnessed significant advances in deep learning for automated pulmonary disease diagnosis, yet notable bottlenecks persist in the DCLDs setting. As a rare disease spectrum, DCLDs suffer from extremely limited high-quality annotated data, causing end-to-end supervised models to overfit severely and generalize poorly [6]. Meanwhile, the absence of publicly available cyst segmentation datasets and mature pretrained models impedes the construction of robust diagnostic systems based on lesion-level feature modeling. Multimodal Large Language Models (MLLMs) [7–11], empowered by large-scale medical knowledge and vision-language alignment training, offer a promising alternative through their strong zero-shot and few-shot reasoning capabilities [12]. Recent agent-based approaches have further extended MLLMs with tool-use

and iterative reasoning, achieving strong performance on chest X-ray interpretation [13, 14] and 3D CT question answering [15]. Nevertheless, existing MLLMs face a fundamental challenge when processing volumetric CT data: **limited context length**. Directly inputting an entire CT volume (typically 300∼500 slices) incurs severe context explosion, whereas naïve sparse sampling disrupts the overall spatial distribution of lesions, preventing the model from forming reliable global perception and substantially increasing reasoning instability and prediction hallucinations, which manifest as low diagnostic accuracy and poor interpretability [16–19].

Inspired by the clinical imaging diagnostic workflow, we propose an interpretable MLLM-driven diagnostic reasoning framework for DCLDs. Clinicians interpreting DCLDs typically follow a cognitive pathway of "first assessing overall spatial distribution, then locally verifying key morphological features" [5, 3]: they survey whole-lung images to identify cyst distribution patterns, scrutinize cyst morphology on representative slices, and ultimately integrate clinical information to reach a conclusion. Drawing on chain-of-thought reasoning [20, 21] and self-verification strategies [22–24], our framework is driven by two synergistic components. First, **Global Spatial Perception (GSP)** segments the lungs and DCLDs-related lesions and projects critical 3D lesion information into compact 2D representations, substantially reducing input context while preserving spatial distribution patterns and cyst morphology. Second, **Multistep Self-Verification Reasoning (MSVR)** guides the model to actively retrieve and analyze the most diagnostically informative axial slices, progressively mining and verifying key visual evidence, including wall thickness, morphological differences, and distribution regions, while dynamically refining diagnostic probabilities to achieve an interpretable reasoning process [21]. This design avoids context explosion and endows the model with hierarchical reasoning capabilities analogous to those of human radiologists. Furthermore, we construct a large-scale, multi-center DCLDs dataset with pixel-level cyst annotations to systematically evaluate the generalization of our framework under realistic clinical conditions. Extensive experiments demonstrate that our method significantly outperforms existing state-of-the-art approaches in diagnostic accuracy, stability, and interpretability, confirming the potential of MLLMs for intelligent diagnosis of complex rare diseases such as DCLDs.

2 Methodology

2.1 Framework Overview

Given a CT volume $\mathbf{X} \in \mathbb{R}^{D\times H\times W}$ and clinical metadata $\mathbf{c}$, we formulate DCLDs diagnosis as an interpretable self-verification reasoning process that emulates the radiologist's cognitive pathway: first assessing global spatial distribution, then locally verifying key morphological features. The framework obtains cyst and lung segmentation masks via dedicated models [25, 26] and fuses them into an anatomical label map $\mathbf{L}$. Two core stages follow: (1) **Global Spatial Perception (GSP)** renders $\mathbf{L}$ into a compact 3D visualization o_{3D} and queries

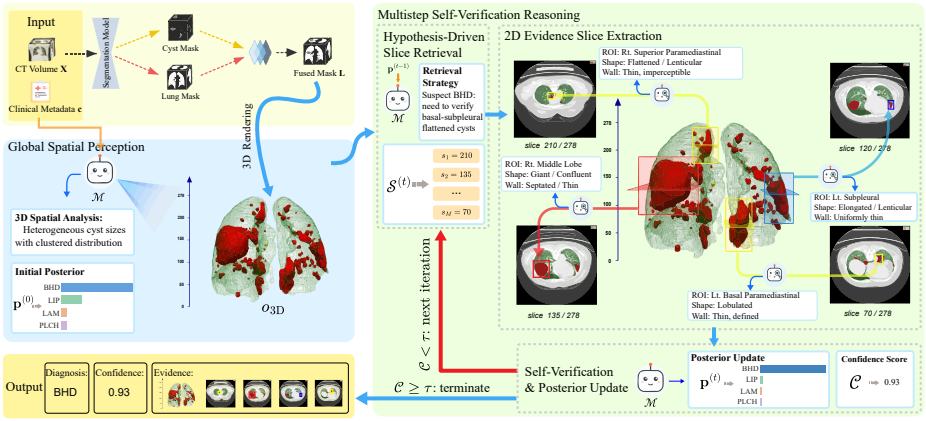

Fig. 2. Overview of the proposed framework.

the MLLM $\mathcal{M}$ to form an initial diagnostic hypothesis $\mathbf{p}^{(0)}$, addressing spatial understanding under limited context length; (2) **Multistep Self-Verification Reasoning (MSVR)** takes $\mathbf{p}^{(0)}$ as starting hypothesis and iteratively retrieves informative slices, verifies morphological evidence, and refines diagnostic probabilities until confidence exceeds τ or reports uncertainty after T rounds (Fig. 2). Algorithm 1 details the full procedure.

2.2 Global Spatial Perception (GSP)

Differential diagnosis of DCLDs relies heavily on the three-dimensional spatial distribution of cysts, yet directly inputting D ($\approx 300 \sim 500$) slices exceeds the context window of typical MLLMs. GSP overcomes this limitation by compressing 3D lesion spatial information into a single compact rendering while preserving key diagnostic cues.

GSP takes the fused label map $\mathbf{L}$ as input and applies surface rendering [27] to extract cyst and lung boundary surfaces, producing a coronal-view rendering $o_{3\mathrm{D}}$. Cysts appear as solid structures within transparent lung wireframes, compressing distribution patterns, size heterogeneity, and spatial density into a single image instead of $\mathcal{O}(D)$ slices. We append a slice-index axis aligned with the DICOM z-order to establish a spatial mapping between the 3D rendering and individual axial slices. This axis transforms the rendering into a navigation map that enables MSVR to precisely locate target slices for hypothesis-driven retrieval.

Upon obtaining $o_{3\mathrm{D}}$, the MLLM $\mathcal{M}$ receives the rendering along with clinical metadata $\mathbf{c}$, performs 3D spatial analysis, and outputs an initial diagnostic posterior $\mathbf{p}^{(0)}$ over the DCLDs subtypes. This initial posterior serves as the starting hypothesis for MSVR, providing a concrete direction for evidence retrieval.

Algorithm 1 Interpretable Self-Verification Reasoning for DCLDs Diagnosis.

Require: CT volume $\mathbf{X}$; clinical metadata $\mathbf{c}$; label set $\mathcal{Y} = \{\text{BHD}, \text{LAM}, \text{PLCH}, \text{LIP}\}$; cyst segmentation model $\mathcal{F}_{\text{cyst}}$; lung segmentation model $\mathcal{F}_{\text{lung}}$; MLLM $\mathcal{M}$; slice budget M; max iterations T; confidence threshold τ.

Ensure: diagnosis $\hat{y} \in \mathcal{Y} \cup \{\texttt{Uncertain}\}$, confidence $\mathcal{C}$, evidence $\mathcal{E} = \{o_{3\text{D}}\} \cup \bigcup_t o_{2\text{D}}^{(t)}$.

$\mathbf{M}_{\text{cyst}} \leftarrow \mathcal{F}_{\text{cyst}}(\mathbf{X})$, $\mathbf{M}_{\text{lung}} \leftarrow \mathcal{F}_{\text{lung}}(\mathbf{X})$ ▷ segmentation
$\mathbf{L} \leftarrow \mathbf{M}_{\text{cyst}} \oplus \mathbf{M}_{\text{lung}}$ ▷ mask fusion
$o_{3\text{D}} \leftarrow \textsc{GlobalSpatialPerception}(\mathbf{L})$ ▷ 3D rendering
$\mathbf{p}^{(0)} \leftarrow \mathcal{M}(o_{3\text{D}}, \mathbf{c})$ ▷ initial posterior
$\mathcal{H} \leftarrow \{o_{3\text{D}}, \mathbf{c}, \mathbf{p}^{(0)}\}$ ▷ initial history
for $t = 1$ **to** T **do** ▷ MSVR loop
  $\mathcal{S}^{(t)} \leftarrow \mathcal{M}(\mathcal{H}; M)$ ▷ hypothesis-driven slice retrieval ($\leq M$)
  $o_{2\text{D}}^{(t)} \leftarrow \{\mathbf{X}_k\}_{k \in \mathcal{S}^{(t)}}$ ▷ 2D evidence slice extraction
  $\mathbf{p}^{(t)} \leftarrow \mathcal{M}(\mathcal{H}, o_{2\text{D}}^{(t)})$ ▷ self-verification & posterior update
  $\mathcal{H} \leftarrow \mathcal{H} \cup \{\mathcal{S}^{(t)}, o_{2\text{D}}^{(t)}, \mathbf{p}^{(t)}\}$ ▷ history update
  $\mathcal{C} \leftarrow \max_{y \in \mathcal{Y}} \mathbf{p}^{(t)}(y)$ ▷ confidence score
  if $\mathcal{C} \geq \tau$ **then**
    break
  end if
end for
if $\mathcal{C} \geq \tau$ **then**
  $\hat{y} \leftarrow \arg\max_{y \in \mathcal{Y}} \mathbf{p}^{(t)}(y)$
else
  $\hat{y} \leftarrow \texttt{Uncertain}$ ▷ unresolved after T iterations
end if
return $\hat{y}, \mathcal{C}, \mathcal{E}$

2.3 Multistep Self-Verification Reasoning (MSVR)

Global spatial perception alone is insufficient for a reliable diagnosis, as DCLDs differentiation further requires verifying fine-grained morphological features on axial slices, including wall thickness, cyst shape, and their relationships with bronchovascular bundles. MSVR implements an iterative evidence-gathering and verification process: the model actively seeks evidence guided by its current hypothesis and verifies or revises that hypothesis at each step, producing transparent reasoning chains with probability updates throughout.

Hypothesis-Driven Slice Retrieval. At each iteration, the MLLM formulates a retrieval strategy based on the current posterior $\mathbf{p}^{(t-1)}$ and accumulated reasoning history $\mathcal{H}$, selecting slices according to subtype-specific diagnostic criteria. The model outputs a sparse index set $\mathcal{S}^{(t)}$ ($|\mathcal{S}^{(t)}| \leq M$), leveraging the slice-index axis established by GSP to precisely locate target positions, and extracts the corresponding 2D evidence slices $o_{2\text{D}}^{(t)}$ for morphological analysis. Unlike uninformative uniform sampling, this strategy enables targeted examination. For example, suspecting BHD triggers the retrieval of basal-subpleural slices to verify flattened, thin-walled cysts (Fig. 2).

Table 1. Quantitative comparison of 4-class DCLDs diagnosis on internal and external cohorts. **Bold** text indicates the best results.

Method	Input	Center A (Internal)		Center B (External)	
		Accuracy	mF1	Accuracy	mF1
Supervised Deep Learning Baselines					
ConvNeXt-S [28]	3D Render	76.4	69.6	59.2	53.0
ResNet-18 [29]	3D Render	71.4	56.7	54.0	48.5
ResNet-50-3D [29]	3D Vol.	56.4	34.6	26.0	19.6
DenseNet-121-3D [30]	3D Vol.	57.9	42.8	39.0	29.6
ResNeXt-50-3D [31]	3D Vol.	60.0	38.2	25.7	20.0
ResNeXt-50-3D [31]	3D Vol.+Mask	65.0	46.8	33.8	26.4
Medical MLLMs					
LLaVA-Med [7]	2D×15+Clin.	52.6	17.2	17.5	7.5
M3D-LaMed [8]	3D Vol.+Clin.	51.7	17.0	14.6	6.4
Ours					
Claude-sonnet-4.5	3D Vol.+Clin.	75.9	71.5	67.5	66.9
GPT-5.2	3D Vol.+Clin.	78.8	75.6	67.5	66.7
Gemini-3-Pro	3D Vol.+Clin.	**87.6**	**82.9**	**73.8**	**72.3**

Abbreviations: 3D Render (3D Rendering), 3D Vol. (3D Volume), 2D×15 (5 Slices/View), +Clin. (With Clinical Metadata), mF1 (Macro F1-Score). All metrics reported in %.

Self-Verification and Posterior Update. After retrieval, the MLLM performs self-verification by evaluating the observed morphological features against the current hypothesis. Confidence increases when 2D evidence supports the hypothesis (*e.g.*, thin-walled flattened cysts confirmed in basal regions). Conversely, if the evidence is contradictory (*e.g.*, unexpectedly thick-walled cysts violating BHD criteria), the model revises its differential diagnosis and shifts its verification targets in the subsequent iteration. This process is formalized as posterior update $\mathbf{p}^{(t)} = \mathcal{M}(\mathcal{H}, o_{\text{2D}}^{(t)})$, with confidence $\mathcal{C} = \max_{y \in \mathcal{Y}} \mathbf{p}^{(t)}(y)$. The loop terminates when $\mathcal{C} \geq \tau$; cases not reaching τ after T rounds are reported as uncertain. This closed-loop self-verification mitigates prediction hallucinations by preventing premature commitment to erroneous reasoning paths. Moreover, it provides a traceable evidence chain for the final diagnostic conclusion.

3 Experiments and Results

Data Curation and Implementation Details. CT scans with confirmed DCLDs diagnoses were collected from two centers following histopathological or guideline-based criteria, comprising **217 cases** (BHD 86, LAM 57, PLCH 23, LIP 51; median slice thickness 1.25 mm) with pixel-level cyst annotations. Center A (137 cases) serves as the internal cohort and Center B (80 cases) as the external test set. All supervised methods are trained on Center A using patient-level 5-fold cross-validation. Our cyst segmentation model, nnU-Net v2 [25] (3d_fullres, ResEnc L), trained on Center A with 5-fold cross-validation, achieves **Dice=0.90** on Center B. Default settings: M=4, T=2, τ=0.9. Three

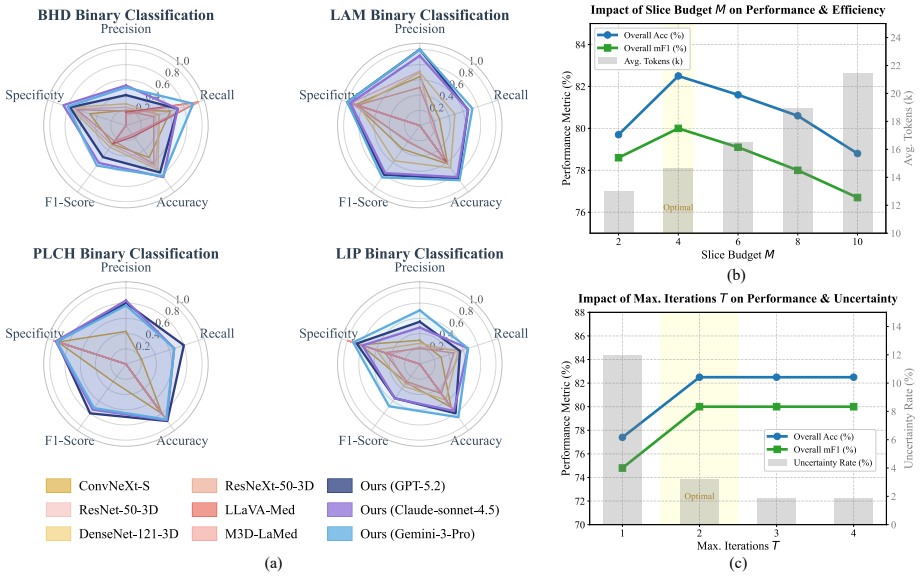

Fig. 3. (a) Per-subtype binary classification on the external test set. (b) Impact of slice budget M on performance and efficiency. (c) Impact of max iterations T on performance and uncertainty rate.

MLLM backbones are evaluated: Claude-sonnet-4.5, GPT-5.2, and Gemini-3-Pro (temperature = 0.1). GSP employs Marching Cubes [27] with mesh decimation and 1024×1024 off-screen PyVista [32] rendering.

Quantitative and Qualitative Results. Table 1 summarizes the results. Our framework with Gemini-3-Pro achieves the best accuracy on both cohorts (**87.6%**/**73.8%**), surpassing the strongest supervised baseline (ConvNeXt-S) by +**11.2**/+**14.6** points. Supervised methods suffer severe cross-center degradation, whereas our framework maintains stable performance, demonstrating generalization akin to physician-level reasoning. Notably, ConvNeXt-S trained on 3D renderings outperforms all standard 3D volumetric models, confirming the inherent discriminative power of the GSP representation. Medical MLLMs fail at differentiation (mF1<18%), as they lack pretraining on pulmonary cysts and cannot recognize them. Fig. 3(a) shows per-subtype binary results; our framework achieves the largest radar area across all four subtypes. Fig. 4 presents a qualitative comparison: given clinical metadata suggestive of BHD, standard Gemini-3-Pro produces hallucinated diagnoses biased by clinical priors. Our framework overcomes this through self-verification by retrieving axial slices and identifying morphological evidence that contradicts BHD and confirms LAM with high confidence.

Parameter Study. For the slice budget M (Fig. 3(b)), M=2 provides insufficient morphological evidence while $M \geq 8$ inflates context without commensurate gains and introduces noise that triggers hallucinations; M=4 achieves the opti-

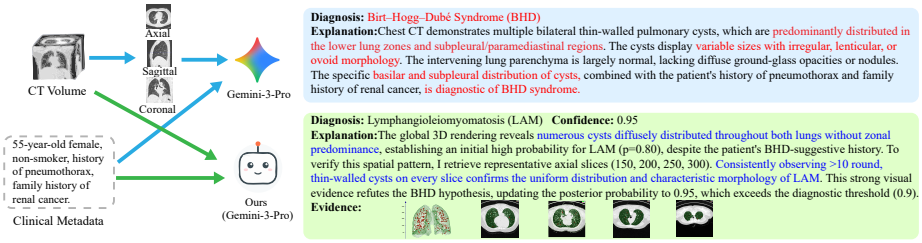

Fig. 4. Qualitative comparison between Gemini-3-Pro and our framework. Errors and hallucinations are highlighted in red; evidence-grounded diagnoses are in blue.

Table 2. Ablation study on key components.

Configuration	GSP	Retrieval	Iter.	Clin.	Accuracy	Precision	Recall	mF1
Full framework	✓	Hyp.-driven	✓	✓	**82.5**	**81.8**	**79.0**	**80.0**
w/o GSP	✗	Uniform	✓	✓	49.8	54.8	46.9	49.1
w/o MSVR	✓	✗	✗	✓	71.9	73.0	69.9	70.8
w/o Iteration (T=1)	✓	Hyp.-driven	✗	✓	77.4	77.0	73.4	74.7
w/o Clinical	✓	Hyp.-driven	✓	✗	74.2	73.2	68.6	70.0

Abbreviations: Hyp.-driven (Hypothesis-Driven), Uniform (Uniformly sampling M slices from each of 3 views), Iter. (Iterative Reasoning Loop), Clin. (Clinical Metadata), mF1 (Macro F1-Score). All metrics are computed over both centers combined and reported in %.

mal balance between performance and efficiency. For max iterations T (Fig. 3(c)), increasing from 1 to 2 yields +5.1% accuracy and +5.3% mF1 with uncertainty dropping sharply from 12.0% to 3.2%. Beyond T=2, performance plateaus; the remaining uncertain cases involve patients with extremely sparse cysts. The system appropriately refuses to commit a diagnosis rather than forcing incorrect predictions, demonstrating reliable self-awareness.

Ablation Study. Table 2 validates each component. GSP is the most critical module: its removal causes accuracy to drop from **82.5%** to 49.8%, as the model loses both global spatial understanding and the slice-index navigation map required for hypothesis-driven retrieval, degenerating to uninformative uniform sampling that fails to capture subtype-specific distribution patterns. Removing MSVR reveals that global perception alone is insufficient, as fine-grained 2D morphological verification is necessary to distinguish subtle differential features invisible in the 3D rendering. The iterative self-verification loop further enhances performance by enabling the model to revisit and correct initial hypotheses when subsequent evidence contradicts them, effectively preventing reasoning hallucinations. Clinical metadata provides complementary discriminative value.

4 Conclusion

We present an interpretable MLLM-driven framework for DCLDs diagnosis that addresses the context-length bottleneck via Global Spatial Perception and Mul-

tistep Self-Verification Reasoning, emulating radiologists' hypothetico-deductive workflow. On 217 multi-center cases, the framework achieves **82.5%** overall accuracy, significantly outperforming supervised and MLLM baselines, with ablations confirming the criticality of GSP (Δ_{Acc}=**−32.7%**) and MSVR (Δ_{Acc}=**−10.6%**). We also contribute the first pixel-level annotated multi-center DCLDs dataset and a dedicated cyst segmentation model (Dice=**0.90**). Current limitations include: GSP employs only coronal-view rendering, which inevitably introduces occlusion of overlapping structures; and the model's slice index selection relies entirely on prompt guidance and its inherent visual reasoning capability without task-specific fine-tuning. Future work will explore multi-view rendering and reinforcement learning-based retrieval optimization to further enhance performance.

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. Gupta, N., Vassallo, R., Wikenheiser-Brokamp, K.A., McCormack, F.X.: Diffuse cystic lung disease. Part I. American Journal of Respiratory and Critical Care Medicine **191**(12), 1354–1366 (2015)
2. Gupta, N., Vassallo, R., Wikenheiser-Brokamp, K.A., McCormack, F.X.: Diffuse cystic lung disease. Part II. American Journal of Respiratory and Critical Care Medicine **192**(1), 17–29 (2015)
3. Raoof, S., Bondalapati, P., Vydyula, R., *et al.*: Cystic Lung Diseases: Algorithmic Approach. Chest **150**(4), 945–965 (2016)
4. Toro, J.R., Pautler, S.E., Stewart, L., *et al.*: Lung cysts, spontaneous pneumothorax, and genetic associations in 89 families with Birt-Hogg-Dubé syndrome. American Journal of Respiratory and Critical Care Medicine **175**(10), 1044–1053 (2007)
5. Gupta, N., Meraj, R., Tanase, D., James, L.E., McCormack, F.X.: Accuracy of chest high-resolution computed tomography in diagnosing diffuse cystic lung diseases. European Respiratory Journal **46**(4), 1196–1199 (2015)
6. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021)
7. Li, C., Wong, C., Zhang, S., *et al.*: LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)
8. Bai, F., Du, Y., Huang, T., Meng, M.Q.-H., Zhao, B.: M3D: Advancing 3D Medical Image Analysis with Multi-Modal Large Language Models. arXiv preprint arXiv:2404.00578 (2024)
9. Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Towards Generalist Foundation Model for Radiology by Leveraging Web-scale 2D&3D Medical Data. arXiv preprint arXiv:2308.02463 (2023)
10. Yang, L., Xu, S., Sellergren, A., *et al.*: Advancing Multimodal Medical Capabilities of Gemini. arXiv preprint arXiv:2405.03162 (2024)

11. Saab, K., Tu, T., Weng, W.H., *et al.*: Capabilities of Gemini Models in Medicine. arXiv preprint arXiv:2404.18416 (2024)
12. Moor, M., Banerjee, O., Abad, Z.S.H., *et al.*: Foundation models for generalist medical artificial intelligence. Nature **616**(7956), 259–265 (2023)
13. Chen, Z., Varma, M., Delbrouck, J.B., *et al.*: CheXagent: Towards a Foundation Model for Chest X-Ray Interpretation. arXiv preprint arXiv:2401.12208 (2024)
14. Fallahpour, A., Ma, J., Munim, A., Lyu, H., Wang, B.: MedRAX: Medical Reasoning Agent for Chest X-ray. In: International Conference on Machine Learning (ICML). PMLR (2025)
15. Mao, Y., Xu, W., Qin, Y., Gao, Y.: CT-Agent: A Multimodal-LLM Agent for 3D CT Radiology Question Answering. arXiv preprint arXiv:2505.16229 (2025)
16. Fu, Y., Panda, R., Niu, X., Yue, X., Hajishirzi, H., Kim, Y., Peng, H.: Data Engineering for Scaling Language Models to 128K Context. In: International Conference on Machine Learning (ICML). PMLR (2024)
17. Pal, A., Umapathi, L.K., Sankarasubbu, M.: Med-HALT: Medical Domain Hallucination Test for Large Language Models. In: Conference on Computational Natural Language Learning (CoNLL) (2023)
18. Ji, Z., Yu, T., Xu, Y., *et al.*: Towards Mitigating Hallucination in Large Language Models via Self-Reflection. In: Findings of EMNLP (2023)
19. Hayden, N., Gilbert, S., Poisson, L.M., Griffith, B., Klochko, C.: Performance of GPT-4 with Vision on Text- and Image-based ACR Diagnostic Radiology In-Training Examination Questions. Radiology **312**(3), e240153 (2024)
20. Wei, J., Wang, X., Schuurmans, D., *et al.*: Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In: Advances in Neural Information Processing Systems (NeurIPS) (2022)
21. Liu, J., Wang, Y., Du, J., Zhou, J.T., Liu, Z.: MedCoT: Medical Chain of Thought via Hierarchical Expert. In: Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 17371–17389 (2024)
22. Weng, Y., Zhu, M., Xia, F., *et al.*: Large Language Models are Better Reasoners with Self-Verification. In: Findings of EMNLP (2023)
23. Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., Yao, S.: Reflexion: Language Agents with Verbal Reinforcement Learning. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)
24. Yao, S., Zhao, J., Yu, D., *et al.*: ReAct: Synergizing Reasoning and Acting in Language Models. In: International Conference on Learning Representations (ICLR) (2023)
25. Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K., Jaeger, P.F.: nnU-Net Revisited: A Call for Rigorous Validation in 3D Medical Image Segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer (2024)
26. Wasserthal, J., Breit, H.C., Meyer, M.T., *et al.*: TotalSegmentator: robust segmentation of 104 anatomic structures in CT images. Radiology: Artificial Intelligence **5**(5), e230024 (2023)
27. Lorensen, W.E., Cline, H.E.: Marching cubes: A high resolution 3D surface construction algorithm. ACM SIGGRAPH Computer Graphics **21**(4), 163–169 (1987)
28. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A ConvNet for the 2020s. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11966–11976 (2022)
29. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)

30. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely Connected Convolutional Networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2261–2269 (2017)
31. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated Residual Transformations for Deep Neural Networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5987–5995 (2017)
32. Sullivan, C.B., Kaszynski, A.A.: PyVista: 3D plotting and mesh analysis through a streamlined interface for the Visualization Toolkit (VTK). Journal of Open Source Software **4**(37), 1450 (2019)