Towards a Radiologist Imitation Framework for 3D CT Diagnosis with Multimodal LLMs

Kaidi Zhang[1], **Zhiyuan Yan**[1], **Gao Cheng**[2], **Zhenyang Cai**[1†]
[1] The Chinese University of Hong Kong, Shenzhen
[2] Shandong Second Medical University
{kaidizhang, zhenyangcai}@link.cuhk.edu.cn

Abstract

Three-dimensional Computed Tomography (3D CT) is a cornerstone of precision medicine. Most AI diagnostic models analyze large numbers of CT slices uniformly, treating all slices as equally important. While this has partly accelerated radiologists' workflows, it overlooks that clinically relevant information is often sparsely distributed throughout a volume. Without targeted or weighted processing, fine-grained cues may be missed and substantial computation wasted on diagnostically uninformative slices. We propose a radiologist-simulating framework for selective and efficient 3D CT interpretation. Evaluated on a 3D CT dataset covering eight thoracic lesion types, it was compared with state-of-the-art multimodal large language models such as GPT-4o and supervised visual backbones including ViT and ResNet-50. Using accuracy, F1-score, AUC, and blind radiologist assessment, Screen-CLIP achieved an AUC of 0.87 and F1-score of 0.82, surpassing ViT-Base (AUC: 0.84). For report generation, our method outperformed M3D across all metrics, reaching a BLEU-Avg of 29.03, and achieved the highest average Doctors' Score (6.16/10) in a preliminary human evaluation.

1 Introduction

Three-dimensional computed tomography (3D CT), as the mainstream three-dimensional medical imaging modality, has become the cornerstone of modern precision medicine through stereoscopic quantification of anatomical structures (Wang et al., 2025; Dayarathna et al., 2024; Chen et al., 2024a; Ouyang et al., 2022). Compared with two-dimensional imaging, 3D CT precisely delineates organ morphology (Raffy et al., 2023), spatial lesion–tissue relationships, and vascular anatomical variations (Wang et al., 2024c; Huynh et al., 2025), playing an irreplaceable role in critical scenarios such as tumor node metastasis (TNM) classification of thoracic tumors (Leduc et al., 2025), differential diagnosis of pulmonary nodules (Niu et al., 2025), and severity assessment of pulmonary infections (Rauner et al., 2025).

However, this technological potential confronts a stark clinical paradox. Projections indicate that global cancer incidence will reach 28.4 million cases by 2040 (Schlemmer, 2023), generating an unprecedented volume of CT examinations and placing unprecedented pressure on radiology departments, yet data from the *Medscape doctor Burnout & Depression Report 2023* indicate that radiologists' interpretive capacity may be approaching a cognitive limit, with 46% of radiologists reporting burnout symptoms. Moreover, a typical chest CT contains hundreds of slices requiring manual review, and diagnostic oversight under information overload has become a core bottleneck contributing to healthcare quality disparities (Reichenpfader et al., 2024; Hering et al., 2022; Mastrodicasa et al., 2025; AlSaad et al., 2024). In this context, intelligent 3D CT interpretation is not merely a technical optimization problem, but a systematic clinical challenge concerning tiered healthcare system efficacy and patient safety.

Despite the potential of Multimodal Large Language Models (MLLMs) in two-dimensional medical imaging analysis, existing MLLMs for three-dimensional scenarios reveal three core limitations. First, full-volume sequence modeling methods like *RadFM* (Wu et al., 2025), and *M3D-LaMed* (Bai et al., 2024) utilize 3D encoders to preserve spatial integrity, but suffer from high GPU overhead and long-sequence noise, struggling to balance efficiency with fine-grained detail. Second, clustering-based dimensionality reduction, such as *Vote-MI* (Wang et al., 2024b), discards intra-cluster textures and anatomical continuity across planes, compromising spatial associations and deviating from the stereoscopic reasoning essential for clinical diagnosis. To bridge the gap between complex

†Corresponding author.

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 1056–1065

AI models and clinical accessibility, researchers have explored more streamlined diagnostic workflows. Finally, challenges persist due to the scarcity of high-quality 3D annotated datasets (Hamamci et al., 2026) and the difficulty of transferring general visual knowledge from large-scale 2D pretraining (e.g., leveraging 1.6 million of biomedical image-text pairs (Lin et al., 2023)) to 3D encoders, which often necessitates inefficient training-from-scratch cycles.

These three issues fundamentally stem from a common limitation: current research focuses on adapting models to 3D data while neglecting radiologists' cognitive diagnostic processes. This paper proposes a radiologist-imitation framework that encodes this cognitive workflow into MLLM pipelines. The framework integrates a confidence-driven dynamic slice selection mechanism that simulates doctors' gaze shifts, compressing MLLM inference to top-M critical slices and significantly reducing computational overhead. A neighborhood context window preserves 3D spatial coherence while avoiding long-sequence interference.

This study seeks improvements in efficiency, performance, and usability for 3D CT interpretation. Our framework (see Figure 1) achieves a reduction in GPU memory consumption by 9.4%. In performance, it achieves an 8.8% relative improvement in macro-average F1 over the best compared baseline. In usability, generated reports outperform *M3D* and *RadFM* across BLEU, METEOR, ROUGE-L, and BERTScore metrics. Furthermore, a preliminary blinded evaluation by three radiologists using a 1–10 scale yields a subjective usability score of 6.16, suggesting improved perceived report quality.

2 Related Work

A persistent challenge in applying multimodal large language models (MLLMs) to 3D medical image analysis is the scarcity of large-scale, high-quality volumetric datasets. Early efforts such as PMC-OA (Lin et al., 2023) and MedMD (Wu et al., 2025) collected 2D and 3D image–text pairs from medical literature and professional websites, partially easing data limitations but still facing heterogeneous quality and inconsistent modality coverage. RP3D (Wu et al., 2025) expanded this scale to 51K 3D image–text pairs, but remained limited by automatic collection noise and restricted task diversity. More recently, CT-RATE (Hamamci et al., 2026) provides 25,692 non-contrast chest CT volumes (expanded to 50,188) paired with radiology reports, substantially strengthening supervised signals for 3D medical vision–language learning, though its focus on chest CT and non-contrast protocols limits broader generalization across organs and imaging modalities.

Beyond data adaptation, several medical MLLMs extend open-source 2D MLLMs to the medical domain, including LLaVA-Med (Li et al., 2023), Med-Flamingo (Moor et al., 2023), and HuatuoGPT-Vision (Chen et al., 2024b). These models mainly target 2D medical images and related reasoning or generation tasks, offering limited direct support for volumetric inputs. RadFM (Wu et al., 2025) extends multimodal modeling to 3D images, but its evaluations are largely text-centric, leaving comprehensive 3D reasoning and segmentation underexplored. SAM-Med3D-MoE (Wang et al., 2024a) introduces a Mixture-of-Experts framework integrating task-specific experts with a foundational SAM-Med3D model, but increases training and inference complexity and depends on routing quality. Med-2E3 (Shi et al., 2025) proposes a text-guided inter-slice scoring module to mimic radiologists' slice-level attention, yet remains slice-based and dependent on textual guidance quality. MLLM-For3D (Huang et al., 2025) transfers 2D MLLM reasoning to 3D scene understanding via multi-view pseudo segmentation and back-projection with spatial constraints; however, this multi-stage pipeline may propagate pseudo-label errors and requires careful geometric alignment. LLMs have also shown potential in public health feature extraction (Zhang et al., 2026).

In contrast, our framework introduces a radiologist-imitation strategy for 3D CT interpretation with MLLMs by explicitly encoding the clinical diagnostic workflow into the inference pipeline. Rather than processing full 3D volumes or relying on proxy formulations such as sparse slice selection, the proposed coarse-to-fine design performs confidence-driven abnormality screening, followed by focused diagnostic interpretation on a small set of critical slices with local 3D context and final structured report generation.

3 Methods

3.1 Data Curation

To strictly validate the proposed framework, all experiments were conducted on the CT-RATE dataset, a publicly available resource consisting of chest

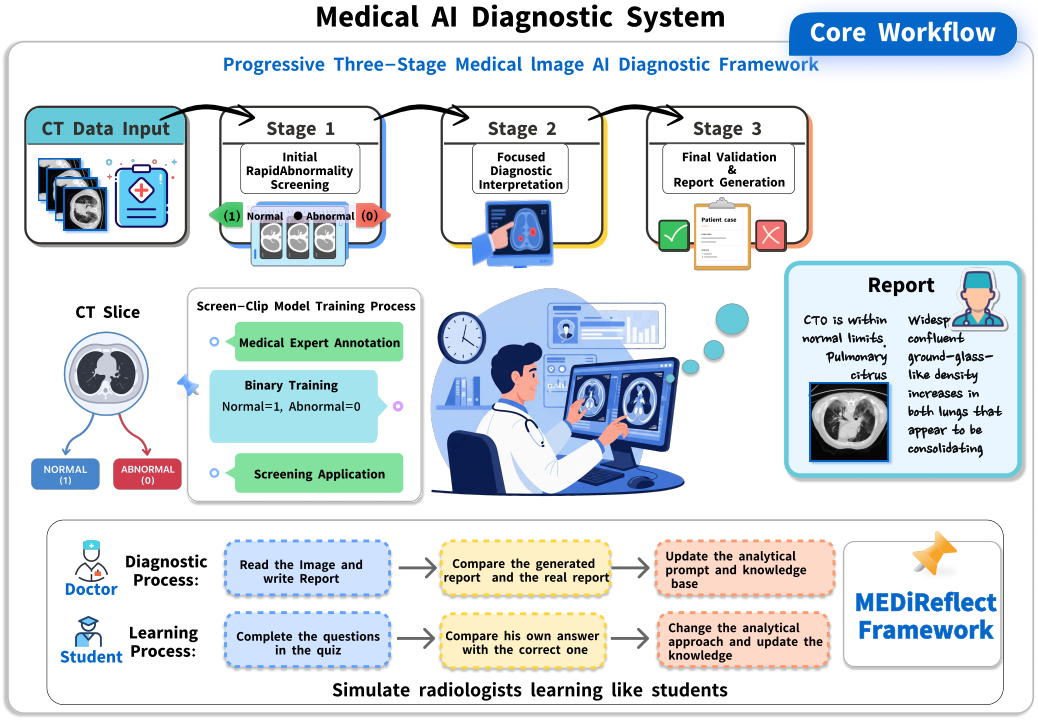

Figure 1: Overview of the proposed radiologist-imitation diagnostic framework, which comprises rapid abnormality screening, focused diagnostic interpretation, and validation through report generation, with *MediReflect* serving as a core mechanism for framework-wide self-refinement.

CT volumes paired with radiology reports, multi-abnormality labels, and associated metadata. We designed two non-overlapping evaluation splits tailored for distinct experimental purposes, with no overlap at either the patient or slice level.

To ensure the clinical reliability of the data, we invited 3 licensed doctors to annotate approximately 12,000 CT slices (from 100 patients), forming Split 1, which is primarily used for training and evaluating the CLIP-based rapid screening module, as well as for diagnostic comparison and ablation analysis. All samples were independently annotated by two doctors; cases with consistent annotations were directly retained, while discrepant cases were further reviewed and confirmed by an additional doctor. A patient-level random split was adopted, with 70% of patients used for training and the remaining 30% reserved for testing, thereby strictly preventing any data leakage between the two sets. Screening performance is evaluated on the test set using *F1-score*, *Precision*, *Recall*, and *Accuracy* (at a fixed threshold of 0.5), while diagnostic experiments rely on the official disease labels provided in the dataset.

Furthermore, Split 2 is introduced to further evaluate report generation quality using the official radiology report annotations written by professional radiologists provided by the publicly available CT-RATE dataset. Specifically, we randomly sample 1,000 CT volumes from the CT-RATE test set. A patient-level split is applied, with 90% of the patients used for updating the proposed automated *MediReflect* mechanism and the remaining 10% reserved for evaluation, ensuring no patient overlap between the two subsets. Ground-truth comparisons are conducted using the corresponding official radiology reports, and report generation quality is assessed using standard natural language generation metrics, including *BLEU*, *ROUGE-L*, *METEOR*, and *BERTScore*.

To enable mainstream MLLM with a 2D encoder to process these 3D datasets, we adopt a standardized preprocessing pipeline. Specifically, we employ percentile-based intensity normalization: voxel intensities are clipped at the 1st and 99th percentiles (P_1 and P_{99}, respectively) to suppress outliers, and then linearly rescaled to the 8-bit grayscale range $[0, 255]$:

$$x' = \frac{x - P_1}{P_{99} - P_1} \times 255\,. \tag{1}$$

followed by conversion to `np.uint8` format. In contrast, for 3D models, we input the data directly without such additional processing. Unless otherwise specified, all models share identical input

resolutions and prompting strategies to ensure fair comparison.

3.2 Overall Framework: A Radiologist-Inspired Agent Pipeline

To balance diagnostic accuracy and computational efficiency, we propose an MLLM-based framework for 3D CT diagnosis and report generation. Inspired by radiologists' clinical reading process, the framework decomposes CT interpretation into three progressive steps: Rapid Abnormality Screening, Focused Diagnostic Interpretation, and Validation and Report Generation.

In **Stage I**, Screen-CLIP rapidly scans the entire CT volume to identify potentially abnormal slices while filtering out normal ones to reduce unnecessary computation. These candidate slices are then passed to **Step II**, where the MLLM performs fine-grained analysis of pathological features. Finally, in **Step III**, the model verifies the intermediate findings and synthesizes them into a coherent, structured diagnostic report.

Step I: Initial Rapid Abnormality Screening via Screen-CLIP To reduce computational overhead, we implement **Screen-CLIP**, a high-throughput module designed to filter pathological slices from 3D volumes. We fine-tune a CLIP model (Radford et al., 2021) for binary classification (Normal: 1, Abnormal: 0) using 12,000 doctor-annotated slices. Specifically, the CLIP Vision Encoder is frozen as a fixed feature extractor, with a compact MLP classification head mounted on top to minimize trainable weights. Training was conducted on two NVIDIA A800 GPUs with a learning rate of 2×10^{-5} and a batch size of 64. During inference, the classifier has generated confidence scores, and slices predicted as "Abnormal" are prioritized for subsequent MLLM analysis.

Step II: Focused Diagnostic Interpretation via MLLM Building on Screen-CLIP, the MLLM performs localized analysis on anomalous slices and their spatial neighbors, simulating radiologists' verification of suspicious findings. For each slice, the abnormality confidence P_{abnormal} is computed by applying a softmax transformation to the CLIP logits ℓ:

$$P_{\text{abnormal}} = \frac{\exp(\ell_{\text{abnormal}})}{\exp(\ell_{\text{normal}}) + \exp(\ell_{\text{abnormal}})}. \quad (2)$$

where ℓ_{normal} and ℓ_{abnormal} represent the logits for each class.

To preserve 3D context while maintaining a manageable computational load, we adopt a sequential interpretation strategy:

1. **Target Selection**: Identify the top-M slices with the highest P_{abnormal} scores as primary targets.
2. **Context Construction**: For each target, a local contextual stack of $K = 9$ slices is formed by including its ± 4 adjacent neighbors.
3. **Sequential Inference**: The MLLM performs M separate inference steps, processing one K-slice stack at a time.

This fixed-size window approach ensures that each finding is analyzed with sufficient volumetric context while avoiding long-sequence noise and hardware memory saturation.

Step III: Final Validation and Report Generation We employ a two-phase strategy to ensure clinical accuracy and coherence. To prevent redundancy and hallucinations inherent in direct concatenation, an MLLM-based "Reflector" first audits the localized interpretations from Step II. Subsequently, the verified findings are synthesized into a standardized radiology report, structured into professional "Findings" and "Impression" sections.

3.3 MediReflect: A Self-Reflective Mechanism for MLLM Evolution

We propose ***MediReflect*** (Figure 2), a **medi**cal application–oriented **reflect**ive learning mechanism, which serves as an adaptive framework that mimics clinical reflective reasoning through two core components: multi-dimensional loss, and trajectory-based knowledge retrieval.

Component I: Multi-Dimensional Penalty Loss *MediReflect* uses a composite loss:

$$\mathcal{L}_{\text{total}} = \omega_1 \mathcal{L}_{\text{slice}} + \omega_2 \mathcal{L}_{\text{report}} + \omega_3 \mathcal{L}_{\text{disease}}, \quad (3)$$

where ω_i denotes the dynamic weight of each objective. The three components are:

1. **Slice Anomaly Loss** ($\mathcal{L}_{\text{slice}}$): Binary cross-entropy over N slices:

$$\mathcal{L}_{\text{slice}} = -\frac{1}{N} \sum_{i=1}^{N} \Big[Y^i_{\text{slice}} \log(P^i_{\text{slice}}) + (1 - Y^i_{\text{slice}}) \log(1 - P^i_{\text{slice}}) \Big]. \quad (4)$$

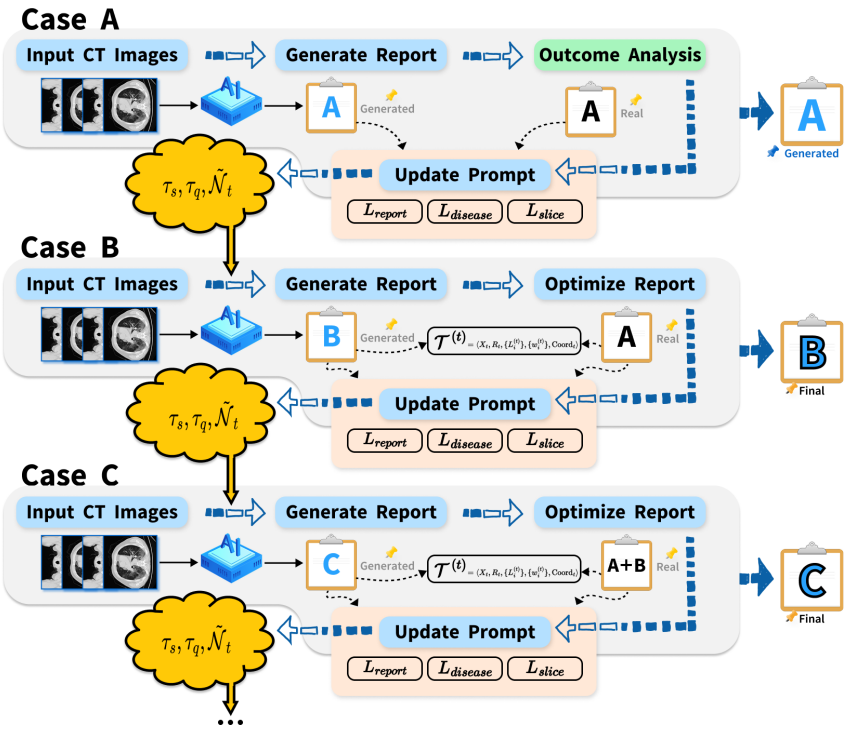

Figure 2: The MediReflect mechanism: an iterative process that alternates between reflection on visual input and retrieval of clinical knowledge to refine diagnostic reasoning.

Here, $Y^i_{\text{slice}} \in \{0, 1\}$ indicates whether the i-th slice contains a lesion, and P^i_{slice} is the predicted anomaly probability from the screening module.

2. **Disease Classification Loss** ($\mathcal{L}_{\text{disease}}$): Multi-label binary cross-entropy over M diagnostic categories:

$$\mathcal{L}_{\text{disease}} = -\frac{1}{M}\sum_{k=1}^{M}\Big[Y^k_{\text{disease}}\log(P^k_{\text{disease}}) + (1 - Y^k_{\text{disease}})\log(1 - P^k_{\text{disease}})\Big]. \tag{5}$$

Here, $Y^k_{\text{disease}} \in \{0, 1\}$ indicates the presence of the k-th disease in the ground truth, and P^k_{disease} is the predicted probability for that category.

3. **Report Quality Loss** ($\mathcal{L}_{\text{report}}$):

$$\mathcal{L}_{\text{report}} = 1 - (\lambda_{\text{sem}} \cdot \text{sim}_k + \lambda_{\text{str}} \cdot \text{comp}_s)\ . \tag{6}$$

We set $\lambda_{\text{sem}} = \lambda_{\text{str}} = 0.5$, where:

- sim_k: Jaccard similarity between clinical entities extracted by **RadGraph** (Jain et al., 2021) from the generated report ($\mathcal{E}_{\text{gen}}$) and the ground-truth report ($\mathcal{E}_{\text{gt}}$):

$$\text{sim}_k = \frac{|\mathcal{E}_{\text{gen}} \cap \mathcal{E}_{\text{gt}}|}{|\mathcal{E}_{\text{gen}} \cup \mathcal{E}_{\text{gt}}| + \epsilon},$$

where ϵ is a small constant.

- comp_s: Structural completeness ratio:

$$\text{comp}_s = \frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}\mathbb{I}(s \in R_{\text{gen}}),$$

where $\mathcal{S} = \{\text{“Findings”}, \text{“Impression”}\}$ is the set of required sections, R_{gen} is the generated report, and $\mathbb{I}(\cdot) = 1$ if section header s appears in R_{gen}, and 0 otherwise.

Component II: Trajectory Learning and Knowledge Retrieval *MediReflect* logs the evolution of each training case into a structured trajectory $T^{(t)}$, forming a searchable knowledge base $\mathcal{K}$:

$$T^{(t)} = \Big\langle X_t, R^{\text{gen}}_t, R^{\text{gt}}_t, Y^{\text{slice}}_t, Y^{\text{disease}}_t, \{L^{(t)}_i\}, \{w^{(t)}_i\}, P^t_{\text{slice}}\Big\rangle\ . \tag{7}$$

Here, X_t denotes the input CT volume; R^{gen}_t and R^{gt}_t represent the generated and ground-truth reports; $Y^{(\cdot)}_t$ are the ground-truth labels; $\{L^{(t)}_i\}$ and

$\{w_i^{(t)}\}$ store the historical loss values and dynamic weights; and P_{slice}^t is the stored anomaly probability vector generated during the training phase.

During inference, since ground-truth profiles are unavailable, we construct a **Predicted Spatial Anomaly Profile** $\hat{P} = [\hat{y}_1, \ldots, \hat{y}_N]$, where $\hat{y}_i$ is the anomaly confidence score of the i-th slice obtained from the Stage-1 Screen-CLIP. We then perform KNN retrieval to identify a set of clinically similar historical cases $\mathcal{N}_X$:

$$\mathcal{N}_X = \text{KNN}\left(\hat{P}, \mathcal{K}, k\right). \tag{8}$$

where k is the number of neighbors retrieved from knowledge base $\mathcal{K}$. These retrieved trajectories undergo an LLM-based filtration to evaluate reference-worthiness. The selected high-quality trajectories are then utilized to assist prompt tuning and optimize the current diagnostic reasoning.

4 Results

We conduct comprehensive experiments to evaluate the proposed framework across three key steps, including rapid abnormality screening, diagnostic interpretation, and report generation quality, together with a component-level ablation study to assess the contribution of each module. For the end-to-end pipeline, we first analyze the performance of the trained CLIP-based model in the rapid abnormality screening stage and examine whether it can efficiently identify suspicious slices from large 3D volumes. We further compare our framework with state-of-the-art 3D MLLMs such as M3D-LaMed-Phi-3-4B (Shi et al., 2025) and RADFM (Wu et al., 2025) to evaluate diagnostic interpretation performance and the quality of generated clinical reports. Finally, a comprehensive ablation study is conducted to quantify the contribution of each individual component to the overall performance.

Unless otherwise specified, all comparative experiments adopt a unified prompting strategy and experimental setup, with Qwen3-VL-8B-Instruct serving as the core MLLM. Detailed implementation settings and prompt templates are provided in Tables S2–S4.

4.1 Evaluation of Initial Rapid Abnormality Screening (on Step I)

Given the large volume of slices in 3D CT data and the sparsely distributed nature of clinically relevant abnormalities, an effective and efficient screening mechanism is essential to enable scalable downstream diagnostic reasoning. We evaluated the screening performance of our framework on a patient-level 7:3 split of the doctor-annotated subset to ensure no data leakage across subjects. Our **Screen-CLIP** was benchmarked against state-of-the-art MLLMs (**Qwen2.5-VL-72B** (Bai et al., 2025) and **GPT-4o** (Hurst et al., 2024)) in a zero-shot setting, and standard supervised backbones (**PMC-CLIP** (Lin et al., 2023), **ViT-Base**, and **ResNet-50**).

Method	Accuracy	Precision	Recall	F1-Score	AUC
Zero-shot MLLMs					
Qwen2.5-VL-72B	0.62	0.57	0.69	0.63	-
GPT-4o	0.68	**0.89**	0.35	0.50	-
Supervised Models					
PMC-CLIP	0.56	0.52	0.45	0.49	0.60
ViT-Base	0.77	0.77	0.71	0.74	0.84
ResNet-50	0.71	0.69	0.67	0.68	0.79
Screen-CLIP (Ours)	**0.80**	0.80	**0.83**	**0.82**	**0.87**

Table 1: Quantitative comparison of abnormality screening performance on the test set. The best results are highlighted in **bold**. For zero-shot MLLMs, AUC is not reported due to their discrete text-based outputs.

As summarized in Table 1, **Screen-CLIP** significantly outperforms all baseline models, achieving the highest **Accuracy (0.80)**, **F1-score (0.82)**, and **AUC (0.87)**. A pivotal observation is the trade-off between Precision and Recall: while **GPT-4o** yields a high Precision (0.890), its extremely low Recall (0.35) indicates an overly conservative diagnostic bias, which is clinically unacceptable for a screening task due to the heightened risk of missed diagnoses. In contrast, **Screen-CLIP** maintains a robust **Recall (0.83)** while preserving high precision, effectively minimizing false negatives. Furthermore, our model surpasses the best-performing supervised baseline (**ViT-Base**) by **3.6% in AUC** and **16.9% in Recall**. This reliable screening performance serves as a critical gatekeeper for the subsequent spatial anomaly retrieval and diagnostic interpretation stages.

4.2 Comparative Evaluation of Medical Diagnostic Capability (on Step II)

To assess whether the proposed framework achieves competitive diagnostic reasoning performance beyond initial screening, we conduct a controlled comparison against representative state-of-the-art 3D medical MLLMs.

In the M3D (Bai et al., 2024) and RadFM (Wu et al., 2025) model testing, we load the pretrained

M3D model and RadFM on an A800 GPU separately. In the image preprocessing section, the `nibabel` library is used to load NIfTI-format medical imaging files and extract the image data. Next, the image is resampled to meet the model's input requirements, and the output report is generated in the specified format. In this process, we provide the same prompt to M3D, RadFM, and ours; the details can be seen in Table S2, Table S3, and Table S4. Results are shown in Table 2. Our model demonstrates strong disease recognition capabilities, achieving a higher F1-score on 5 out of 8 diseases compared to M3D and RadFM, including significant improvements in categories like Atelectasis and Cardiomegaly.

Disease Feature	M3D citeref18	RadFM (Wu et al., 2025)	Ours
Arterial wall calcification	0.27	0.13	**0.29**
Cardiomegaly	0.09	0.09	**0.17**
Atelectasis	0.34	0.29	**0.53**
Lung nodule	0.29	**0.42**	0.40
Lung opacity	0.17	**0.72**	0.67
Pleural effusion	0.11	0.08	**0.17**
Consolidation	0.37	0.32	**0.36**
Interlobular septal thickening	0.43	**0.67**	0.40
Macro-average	0.26	0.34	**0.37**

Table 2: Comparison of **F1 Score** for Disease Feature Classification between different models. The F1-Score value of some diseases is 0 because the sample is rare or difficult to distinguish.

4.3 Quality Assessment of Medical Report Generation (on Step III)

To evaluate whether the proposed framework can translate diagnostic understanding into coherent and clinically meaningful reports, we assess its report generation quality on a shared test set. We required the MLLM to read CT images (providing 3D CT images directly to M3D and RadFM) and generate reports on the same test set. Therefore, in this experiment, we had medical personnel select several diseases that are recognizable at the current stage. The quality of the generated reports was compared against the ground truth reports using four metrics: BLEU-Avg, BERT-Score, ROUGE-L, and METEOR. As shown in Table 3, our model achieves substantial improvements over M3D and RADFM on all metrics, achieving a significant lead over all baselines. Specifically, it improves the BLEU-Avg by 19.2% and METEOR by 4.3 points compared to the strongest baseline, M3D.

In terms of the 1–10 subjective Doctors' Score, assessed by three individuals with professional medical backgrounds, our framework obtained the highest average score of 6.16, compared with 5.89 for M3D and 5.47 for RADFM, as shown in Figure 3. This result provides preliminary evidence of improved perceived report quality.

Model	BLEU Avg	ROUGE-L	METEOR	BERT Score	Doctors' Score
Ours	29.03	16.73	23.78	82.29	6.16
M3D	24.35	13.38	19.48	81.89	5.89
RADFM	11.71	9.69	13.46	79.08	5.47

Table 3: Quantitative comparison of report generation quality across different models using automatic metrics (BLEU-Avg, ROUGE-L, METEOR, BERT-Score) and human evaluation (Doctors' Score).

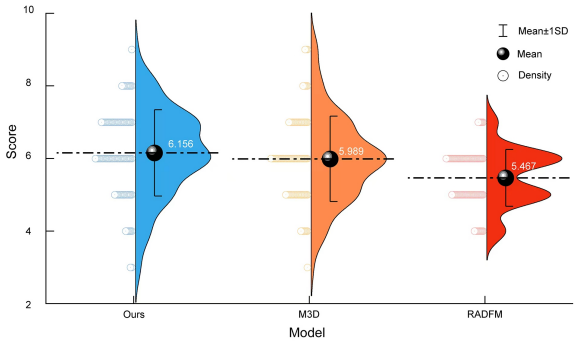

Figure 3: Distribution of doctor-assigned scores evaluating the quality and clinical applicability of model-generated medical reports. Our method achieves a higher mean score compared to other 3D medical MLLMs.

4.4 Ablation Analysis of Core Framework Components

We conduct ablation studies to evaluate the contribution of individual framework components by selectively removing Screen-CLIP, Neighbor Window, Trajectory Retrieval, and TextGrad (Reflection).

As shown in Table 4, **Screen-CLIP** is central to both efficiency and accuracy. By filtering non-informative slices, it reduces average inference latency by 61.3% (from 9.83 s to 3.80 s) and GPU memory consumption by 9.4%.

Configuration	**Latency (s)**	**GPU Memory (GB)**
Without screen-clip	9.83	19.06
With screen-clip	3.80	17.27

Table 4: Impact of screen-clip on latency and GPU memory usage.

Figure 4 further confirms its importance: remov-

ing Screen-CLIP causes a 39.7% drop in F1-score, while integrating it substantially improves throughput. Excluding the **Neighbor Window** decreases F1-score by 32.4%, indicating the importance of context for volumetric interpretation. Removing **TextGrad** reduces BLEU by 23.2%, showing its contribution to report quality. Finally, without **Trajectory Retrieval**, performance declines consistently by approximately 14%–16% across evaluation metrics.

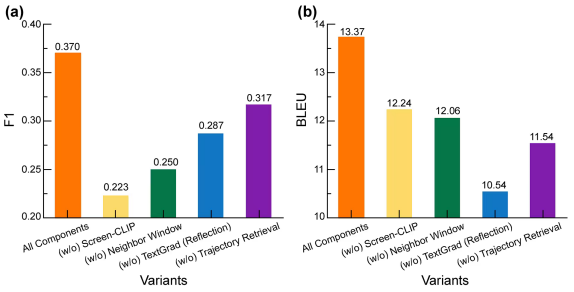

Figure 4: Component-wise ablation study. We evaluate the contribution of Screen-CLIP, Neighbor Window, Trajectory Retrieval, and TextGrad to F1 (accuracy), BLEU (quality), and latency. Removing Screen-CLIP causes the most severe degradation, with a 39.7% drop in F1.

Specifically, in the configuration without Screen-Clip, we bypassed the first-stage screening; instead, clips were input into the second-stage analysis in batches based on the maximum threshold. In the absence of Neighbor Window, rather than selecting the number of clips based on severity calculated via softmax, we solely input the clips identified as abnormal by the Screen-Clip module. For the removal of Trajectory Retrieval, the model was required to make independent judgments without referencing retrieval cases. Finally, in the experiment without TextGrad (Reflection), prompt updates were disabled.

4.5 Discussion

Although our framework achieved a solid **0.11 boost** in macro-average F1-score over M3D, the picture becomes more nuanced at the individual-disease level. This variation largely reflects the intrinsic imaging characteristics of different pathologies.

Our approach was particularly effective for volumetric lesions with clear boundaries, such as **Atelectasis (0.53)** and **Lung Nodules (0.40)**. This is consistent with our design: by mimicking how a radiologist scrolls through the Z-axis, the *MediReflect* mechanism captures spatial continuity that single-slice 2D proxies often miss. In effect, the model learned to distinguish a 3D nodule from a 2D vessel cross-section by considering the "whole picture."

However, there is a trade-off for subtle textures. We observed a slight performance drop in detecting **Consolidation** and **Interlobular Septal Thickening** compared with some baselines. These conditions are difficult because they depend on subtle textural changes that can be confused with normal pulmonary vasculature. Our *Screen-CLIP* module, while efficiently filtering healthy slices, may be somewhat too aggressive here, likely discarding slices containing faint, "threshold-level" signals before the MLLM could analyze them. This suggests that while the screening strategy works well for focal lesions, the filtering threshold may need adjustment for diffuse interstitial diseases.

Finally, like many deep learning systems, we still face the long-tail distribution problem. Although the model performed competitively on critical conditions such as **Cardiomegaly**, its generalization to extremely rare pathologies remains constrained by data scarcity. Even powerful MLLMs appear to struggle to hallucinate correct diagnoses for rare diseases without sufficient few-shot clinical examples.

In short, while our framework successfully replicates radiologist reasoning for structural abnormalities, future work should improve sensitivity to ultra-fine textural changes and robustness to class imbalance.

5 Conclusion

This study proposes a radiologist-imitation framework integrating MediReflect for closed-loop 3D CT diagnosis and reporting, with three clinical advantages. First, it improves diagnostic safety: trained on 12,000 annotated slices, Screen-CLIP achieves a recall of 0.83 and F1-score of 0.82, outperforming baselines and reducing missed diagnoses of subtle lesions. Second, it shows promising perceived report quality in a preliminary blinded human evaluation, obtaining an average Doctors' Score of 6.16 compared with 5.47–5.89 for the baselines. Third, it supports practical deployment: optimized inference reduces GPU memory cost by 9.4% per case, improving diagnostic efficiency in resource-constrained settings.

Limitations

This study has several limitations. The current experiments are mainly conducted on chest CT data, so the generalizability of the framework to other imaging modalities, anatomical regions, and multi-organ diagnostic scenarios remains to be further validated. Although Screen-CLIP improves inference efficiency, further optimization is still needed for real-time clinical deployment. In addition, the current framework does not explicitly model diagnostic uncertainty, which is important for safety-critical medical applications.

The human evaluation should be regarded as preliminary, as it involves a limited number of evaluators and does not yet include inter-rater agreement or statistical significance testing. Future work will conduct larger-scale reader studies and further improve the framework in terms of multimodal adaptation, inference efficiency, and uncertainty-aware diagnosis.

References

Rawan AlSaad, Alaa Abd-Alrazaq, Sabri Boughorbel, Arfan Ahmed, Max-Antoine Renault, Rafat Damseh, and Javaid Sheikh. 2024. Multimodal large language models in health care: applications, challenges, and future outlook. *Journal of medical Internet research*, 26:e59505.

Fan Bai, Yuxin Du, Tiejun Huang, Max Q-H Meng, and Bo Zhao. 2024. M3d: Advancing 3d medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Cheng Chen, Juzheng Miao, Dufan Wu, Aoxiao Zhong, Zhiling Yan, Sekeun Kim, Jiang Hu, Zhengliang Liu, Lichao Sun, Xiang Li, and 1 others. 2024a. Masam: Modality-agnostic sam adaptation for 3d medical image segmentation. *Medical Image Analysis*, 98:103310.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, and 1 others. 2024b. Towards injecting medical visual knowledge into multimodal llms at scale. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 7346–7370.

Sanuwani Dayarathna, Kh Tohidul Islam, Sergio Uribe, Guang Yang, Munawar Hayat, and Zhaolin Chen. 2024. Deep learning based synthesis of mri, ct and pet: Review and analysis. *Medical image analysis*, 92:103046.

Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, and 1 others. 2026. Generalist foundation models from a multimodal dataset for 3d computed tomography. *Nature Biomedical Engineering*, pages 1–19.

Alessa Hering, Lasse Hansen, Tony CW Mok, Albert CS Chung, Hanna Siebert, Stephanie Häger, Annkristin Lange, Sven Kuckertz, Stefan Heldmann, Wei Shao, and 1 others. 2022. Learn2reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning. *IEEE Transactions on Medical Imaging*, 42(3):697–712.

Jiaxin Huang, Runnan Chen, Ziwen Li, Zhengqing Gao, Xiao He, Yandong Guo, Mingming Gong, and Tongliang Liu. 2025. Mllm-for3d: Adapting multimodal large language model for 3d reasoning segmentation. *arXiv preprint arXiv:2503.18135*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

NT Huynh, E Zhang, O Francies, F Kuklis, T Allen, J Zhu, O Abeyakoon, F Lucka, M Betcke, J Jaros, and 1 others. 2025. A fast all-optical 3d photoacoustic scanner for clinical vascular imaging. *Nature Biomedical Engineering*, 9(5):638–655.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, and 1 others. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*.

Charles Leduc, Frank Detterbeck, James Huang, Anja C Roden, and Moishe Liberman. 2025. Managing thymic malignancies in 2025: Migrating from masaoka to tnm. *Journal of Thoracic Oncology*, 20(12):1760–1762.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564.

Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 525–536. Springer.

Domenico Mastrodicasa, Marly Van Assen, Merel Huisman, Tim Leiner, Eric E Williamson, Edward D Nicol, Bradley D Allen, Luca Saba, Rozemarijn Vliegenthart, and Kate Hanneman. 2025. Use of ai in cardiac ct and mri: a scientific statement from the escr, eusomii, nasci, scct, scmr, siim, and rsna. *Radiology*, 314(1):e240516.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. 2023. Med-flamingo: a multimodal medical few-shot learner. In *Machine learning for health (ML4H)*, pages 353–367. PMLR.

Chuang Niu, Qing Lyu, Christopher D Carothers, Parisa Kaviani, Josh Tan, Pingkun Yan, Mannudeep K Kalra, Christopher T Whitlow, and Ge Wang. 2025. Medical multimodal multitask foundation model for lung cancer screening. *Nature Communications*, 16(1):1523.

Cheng Ouyang, Chen Chen, Surui Li, Zeju Li, Chen Qin, Wenjia Bai, and Daniel Rueckert. 2022. Causality-inspired single-source domain generalization for medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(4):1095–1106.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Philippe Raffy, Jean-François Pambrun, Ashish Kumar, David Dubois, Jay Waldron Patti, Robyn Alexandra Cairns, and Ryan Young. 2023. Deep learning body region classification of mri and ct examinations. *Journal of Digital Imaging*, 36(4):1291–1301.

Gat Rauner, Piyush B Gupta, and Charlotte Kuperwasser. 2025. From 2d to 3d and beyond: the evolution and impact of in vitro tumor models in cancer research. *Nature Methods*, 22(9):1776–1787.

Daniel Reichenpfader, Henning Müller, and Kerstin Denecke. 2024. A scoping review of large language model based approaches for information extraction from radiology reports. *npj Digital Medicine*, 7(1):222.

Heinz-Peter Schlemmer. 2023. The cancer epidemic : Global significance of cancer and the situation in oncological imaging. *Radiologie (Heidelb.)*, 63(1):49–56.

Yiming Shi, Xun Zhu, Kaiwen Wang, Ying Hu, Chenyi Guo, Miao Li, and Ji Wu. 2025. Med-2e3: A 2d-enhanced 3d medical multimodal large language model. In *2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2754–2759. IEEE.

Guoan Wang, Jin Ye, Junlong Cheng, Tianbin Li, Zhaolin Chen, Jianfei Cai, Junjun He, and Bohan Zhuang. 2024a. Sam-med3d-moe: Towards a non-forgetting segment anything model via mixture of experts for 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 552–561. Springer.

Jinzhuo Wang, Kai Wang, Yunfang Yu, Yuxing Lu, Wenchao Xiao, Zhuo Sun, Fei Liu, Zixing Zou, Yuanxu Gao, Lei Yang, and 1 others. 2025. Self-improving generative foundation model for synthetic medical image generation and clinical applications. *Nature Medicine*, 31(2):609–617.

Yuli Wang, Jian Peng, Yuwei Dai, Craig Jones, Haris Sair, Jinglai Shen, Nicolas Loizou, Jing Wu, Wen-Chi Hsu, Maliha Imami, and 1 others. 2024b. Enhancing vision-language models for medical imaging: bridging the 3d gap with innovative slice selection. *Advances in Neural Information Processing Systems*, 37:99947–99964.

Zhaoqing Wang, Tianqing Wan, Sijie Ma, and Yang Chai. 2024c. Multidimensional vision sensors for information processing. *Nature Nanotechnology*, 19(7):919–930.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. 2025. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Nature Communications*, 16(1):7866.

Kaidi Zhang, Zimo Zhao, Yan Hu, and Thuy Le. 2026. Ai-driven feature selection using only survey variable descriptions: Large language models identify adolescent vaping predictors. *medRxiv*, pages 2026–03.