# VEC-CT: VISUAL CUE-GUIDED ERROR-CORRECTIVE CLINICAL REASONING FOR LUNG CT REPORT GENERATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Multimodal large language models (MLLMs) provide strong capabilities in visual perception and reasoning, offering a promising approach to medical image understanding and diagnostic report generation. However, accurate interpretation of lung CT scans remains challenging due to subtle lesion characteristics, complex anatomical structures, and strong dependencies among clinical findings. Moreover, inaccurate visual observations in early clinical sub-tasks may propagate and accumulate through subsequent reasoning steps, ultimately reducing the accuracy and clinical consistency of the generated reports. To address these challenges, we construct the LungCT-CoT dataset, which decomposes lung CT report generation into six clinically ordered and interconnected sub-tasks to reflect the clinical diagnostic workflow. Based on this dataset, we propose VEC-CT, a visual cue-guided error-corrective clinical reasoning framework. The model first performs structured reasoning over the predefined clinical sub-tasks and generates an initial diagnostic report. The Error Tracer then identifies the earliest erroneous sub-task in the reasoning process, after which the Visual Cue Extractor extracts task-specific visual cues relevant to the identified sub-task from the CT image. Guided by these visual cues, the model revisits the erroneous sub-task and corrects the subsequent reasoning process to refine the final diagnostic report. Extensive experiments show that VEC-CT consistently outperforms representative methods on both conventional text generation metrics and clinically oriented semantic evaluation metrics.

## 1 INTRODUCTION

Large language models (LLMs) have demonstrated strong capabilities in language understanding and generation (Brown et al., 2020), while multimodal large language models (MLLMs) extend these capabilities to multimodal understanding and reasoning over visual and textual inputs (Bai et al., 2025; Wang et al., 2025). These advances have facilitated the development of medical MLLMs, particularly for diagnostic report generation. Early medical MLLMs primarily adapted general-purpose MLLMs to the medical domain through continued pretraining and medical instruction fine-tuning (Moor et al., 2023; Li et al., 2023a). Despite their promising performance, most approaches follow an end-to-end paradigm with supervision primarily focused on model outputs. As illustrated in Figure 1 (a), the MLLM directly generates final reports from medical images without explicit intermediate reasoning. In clinical practice, however, image interpretation is inherently structured, typically involving lesion localization, imaging appearance characterization, assessment of suspicious signs, and diagnostic judgment (MacMahon et al., 2017).

Chain-of-Thought (CoT) reasoning provides a natural paradigm for decomposing such complex decisions into a sequence of intermediate clinical sub-tasks (Wei et al., 2022; Liu et al., 2024b). However, the intermediate reasoning steps in CoT may still be unreliable. In medical imaging, intermediate clinical decisions rely heavily on fine-grained observations of lesion regions, while existing medical MLLMs may still exhibit limitations in visual perception (Sepehri et al., 2025). Moreover, inaccurate visual observations may introduce errors that can propagate across dependent clinical sub-tasks, as shown in Figure 1 (b), and ultimately affect the generated diagnostic report (Sepehri et al., 2025; Zhang et al., 2024a). This is due to the lack of correction mechanisms in existing CoT-based approaches. Generic reasoning refinement may fail to address such errors (Li et al., 2026), and retraining or fine-tuning the visual encoder is computationally expensive (Li et al., 2023b). These

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

limitations motivate a more targeted strategy that corrects erroneous reasoning by introducing visual cues relevant to specific clinical tasks rather than globally adapting the pretrained visual encoder.

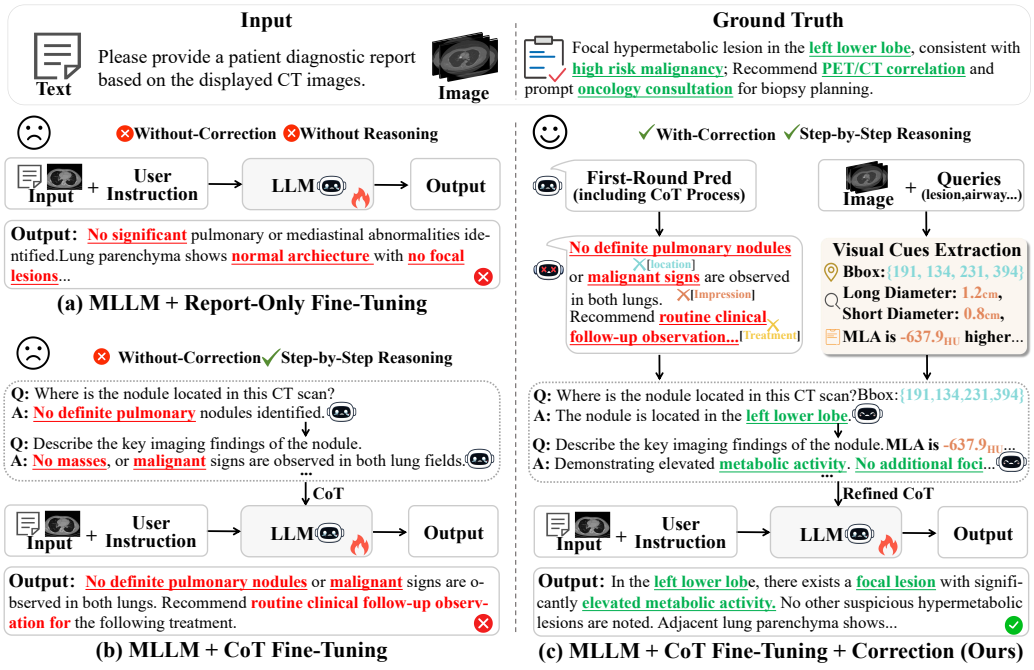



Figure 1: Comparison of VEC-CT with existing methods. (a) The end-to-end paradigm lacks explicit intermediate reasoning. (b) Conventional CoT suffers from error propagation across tasks. (c) VEC-CT traces errors to their earliest source and provides targeted visual cues for correction.

To this end, we construct **LungCT-CoT**, a structured clinical reasoning dataset for lung CT report generation that decomposes the diagnostic workflow into a sequence of predefined clinical sub-tasks with explicit dependencies among successive clinical decisions. Building on LungCT-CoT, we propose **VEC-CT**, a visual cue-guided error-corrective clinical reasoning framework that introduces targeted corrective supervision to mitigate error propagation in CoT. As illustrated in Figure 1 (c), VEC-CT follows a three-stage *reason–verify–correct* training paradigm. First, the **Reasoning Model** sequentially performs the clinically ordered sub-tasks and generates an initial diagnostic report. The **Error Tracer** then identifies the earliest erroneous clinical sub-task. Accordingly, the **Visual Cue Extractor** extracts task-specific visual cues relevant to the identified sub-task for subsequent correction. Starting from the identified sub-task, the Reasoning Model revises the erroneous prediction and regenerates the subsequent clinical decisions based on the corrected reasoning context, without restarting the entire reasoning trajectory. This process reduces the influence of early errors on subsequent clinical decisions while preserving the sequential structure of clinical reasoning. Importantly, the Error Tracer and Visual Cue Extractor are used solely to provide corrective supervision during training. At inference time, the optimized Reasoning Model independently follows the learned clinical reasoning process to generate the final diagnostic report.

Our main contributions are summarized as follows. First, we introduce the LungCT-CoT dataset to organize the clinical diagnostic workflow into a sequence of interdependent sub-tasks and provide task-level supervision beyond direct image-to-report mapping. Second, we propose VEC-CT, which traces the earliest erroneous sub-task and introduces task-specific visual cues to guide targeted correction, thereby mitigating downstream error propagation. Third, extensive experiments demonstrate that VEC-CT achieves consistent improvements over representative existing methods on both conventional generation metrics and clinically oriented semantic evaluation metrics.

## 2 Related Work

**Multimodal Large Language Models in Medicine.** MLLMs are capable of processing diverse multimodal information, laying the foundation for vision–language understanding and reasoning (Radford et al., 2021; Alayrac et al., 2022; Li et al., 2023b). Early studies employed continued

pretraining on multimodal data to enhance cross-modal representation learning (Moor et al., 2023). Subsequent studies adopted fine-tuning (FT) on medical image–text data to specialize general-purpose MLLMs for the medical domain (Li et al., 2023a; Xu et al., 2026). Despite these advances, report generation is still commonly formulated as an end-to-end mapping from medical images to final reports. The **LungCT-CoT** dataset instead organizes the diagnostic workflow into a sequence of clinically ordered sub-tasks, making the intermediate clinical reasoning process explicit.

**Chain-of-Thought Reasoning in MLLMs.** CoT prompting enables models to perform step-by-step reasoning (Wei et al., 2022; Zhang et al., 2024b). To improve the reliability of such reasoning, subsequent studies introduced expert verification and consistency-based strategies, using external feedback or agreement across multiple reasoning paths to verify intermediate predictions (Liu et al., 2024b; Sun et al., 2025). Self-refinement methods further allow MLLMs to revise their predictions, but often lack mechanisms to locate the source of errors (Li et al., 2026). In contrast, **VEC-CT** traces the earliest error in the reasoning trajectory and performs targeted refinement from the identified error point, allowing subsequent clinical decisions to be made based on the corrected context.

**Visual Enhancement in MLLMs.** Lesion-related visual information is essential for accurate clinical reasoning in medical visual question answering (VQA) tasks. Accordingly, considerable efforts have been made to enhance visual understanding in medical MLLMs. Some methods use images with annotated regions of interest (ROIs) as inputs to guide MLLMs toward clinically relevant areas (Chen et al., 2025). Others leverage large-scale medical multimodal data to improve general visual representations by jointly optimizing visual encoders and the projection layer (Zhou et al., 2026). These strategies overlook that relevant visual information varies across medical VQA tasks. **VEC-CT** extracts task-specific visual cues without requiring annotated ROIs or modifying the pre-trained visual encoder.

# 3 METHOD

To mitigate the propagation of errors arising from unreliable intermediate clinical reasoning, we propose **VEC-CT**, a visual cue-guided error-corrective clinical reasoning framework. As illustrated in Figure 2, VEC-CT comprises three main components: a **Reasoning Model**, an **Error Tracer**, and a **Visual Cue Extractor**, which jointly implement a *reason–verify–correct* training procedure. During *reason*, the Reasoning Model sequentially performs predefined clinical sub-tasks and generates diagnostic reports. Then, in the *verification* stage, the Error Tracer compares the reasoning trajectory against task-level reference annotations and identifies the earliest erroneous clinical sub-task. Based on the identified error, the Visual Cue Extractor combines CT image information with the corresponding DICOM metadata to derive task-specific, physically meaningful visual cues. Finally, during *correct*, the same Reasoning Model performs cue-guided targeted correction from the identified erroneous sub-task, followed by regeneration of the downstream clinical decisions, ultimately refining the initial diagnostic report.

## 3.1 PROBLEM FORMULATION

Under the current dataset setting, each sample is represented by a 2D lung CT slice $\boldsymbol{I}$ and its corresponding diagnostic report $R$, directly supporting image-to-report training but offering limited supervision for intermediate clinical reasoning. In clinical practice, pulmonary lesion assessment follows a structured process in which lesion location and imaging characteristics inform diagnostic judgment and malignancy risk assessment, which further guide clinical management (MacMahon et al., 2017). Motivated by this clinical workflow, we construct **LungCT-CoT** around six clinical tasks. **Lesion Location** identifies the anatomical location of the target lesion, while **Lesion Impression** describes its imaging characteristics. **Cancer Judgment** determines the case-level diagnostic status, and **Risk Assessment** estimates the malignancy risk of the target lesion. **Targeted Therapy** generates the corresponding management recommendation based on the preceding findings and diagnostic judgments, whereas **Diagnostic Finding** integrates the preceding clinical information into the final report.

Each task in LungCT-CoT is instantiated as a medical VQA task. We use *Clinical CoT* to denote this explicit sequence of clinically ordered VQA decisions rather than an unconstrained free-form rationale. Let $\mathcal{Q} = \{Q_1, \ldots, Q_K\}$, where $K = 6$, denote the ordered task sequence, and $\mathcal{Y} = \{Y_1^{\text{gold}}, \ldots, Y_K^{\text{gold}}\}$ the corresponding task-level reference answers. Together, the $K$ tasks form

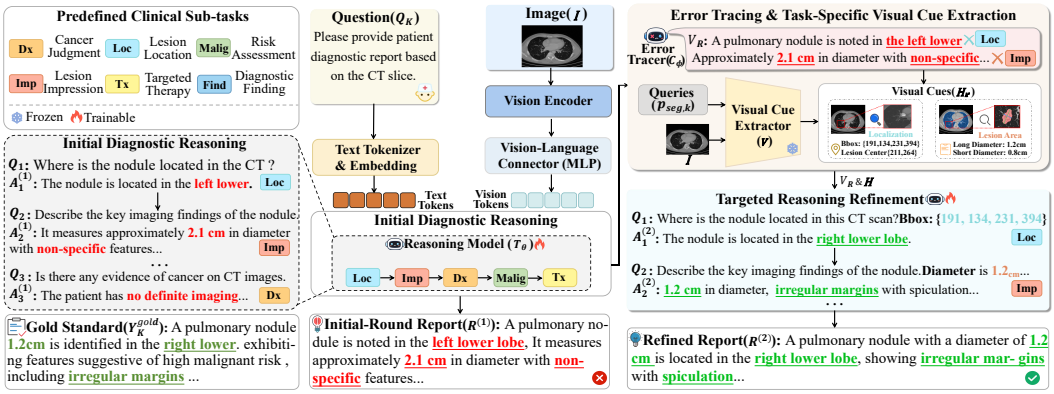



Figure 2: Overview of the VEC-CT training framework. The Reasoning Model first performs structured Clinical CoT reasoning to generate an initial diagnostic trajectory and report. The training-only Error Tracer then identifies the earliest erroneous clinical sub-task. According to the detected error, the Visual Cue Extractor provides task-specific visual cues to guide corrective reasoning from the identified sub-task onward. The Reasoning Model subsequently resumes reasoning from the corrected sub-task to regenerate the downstream clinical decisions and final report.

an ordered reasoning trajectory whose final output is the diagnostic report $R$. Since the final task corresponds to diagnostic report generation, $Y_K^{\text{gold}} = R$.

### 3.2 Initial Diagnostic Reasoning

We instantiate the Reasoning Model $\mathcal{T}_\theta$, parameterized by $\theta$, using Qwen3-VL-8B-Instruct (Bai et al., 2025). The parameters $\theta$ constitute the only trainable parameters in VEC-CT, whereas the Error Tracer and the external Visual Cue Extractor introduced in subsequent sections remain frozen throughout training. Given a CT slice $\boldsymbol{I}$, the vision encoder first converts image patches into visual tokens, which are then projected into the representation space of the language model through the vision–language connector. The current clinical question $Q_k$ and the preceding clinical responses are encoded as textual tokens. The resulting visual and textual representations are jointly processed by the Reasoning Model to generate the response for the current clinical task.

During initial reasoning, the Reasoning Model sequentially performs all predefined clinical tasks as $A_k^{(1)} = \mathcal{T}_\theta\left(\boldsymbol{I}, Q_k, \mathcal{A}_{<k}^{(1)}\right), k = 1, \ldots, K$, where $A_k^{(1)}$ denotes the initial-round prediction for task $k$, and $\mathcal{A}_{<k}^{(1)} = \{A_1^{(1)}, \ldots, A_{k-1}^{(1)}\}$ denotes the preceding initial-round predictions. Unlike teacher forcing, which conditions each task on the reference responses of preceding tasks, VEC-CT conditions each subsequent task on its own preceding predictions. This formulation preserves the sequential dependency among clinical decisions, allowing reasoning errors and their downstream effects to emerge explicitly during training.

The initial reasoning trajectory is supervised using task-level reference responses with $\mathcal{L}_{\text{reason}} = \sum_{k=1}^{K} \ell_{\text{NLL}}(Y_k^{\text{gold}}; \boldsymbol{I}, Q_k, \mathcal{A}_{<k}^{(1)})$, where $\ell_{\text{NLL}}(Y; \cdot)$ denotes the token-level autoregressive negative log-likelihood of target response $Y$ conditioned on the context following the semicolon. Each prediction directly conditions the subsequent clinical tasks, allowing errors at earlier reasoning steps to propagate to later predictions.

### 3.3 Error Tracing and Task-Specific Visual Cue Extraction

#### 3.3.1 Error Tracer

To identify errors in the initial clinical reasoning trajectory, we introduce an **Error Tracer** $\mathcal{C}_\phi$, instantiated with a pretrained Qwen3-VL-8B-Instruct model whose parameters $\phi$ remain frozen throughout training. Since clinical responses are generated in free-form natural language, exact string matching is insufficient to reliably assess their correctness. We therefore formulate error trac-

ing as a semantic consistency assessment between the generated predictions and their corresponding reference annotations. Specifically, the Error Tracer adopts a two-stage procedure: it first evaluates the clinical consistency of the generated report at the report level and, when an inconsistency is detected, further examines the intermediate predictions to localize the erroneous clinical task.

At the report level, the Error Tracer takes three inputs: the predicted diagnostic report $R^{(1)}$, its reference $Y_K^{\text{gold}}$, and a report-level verification prompt $P_{\text{report}}$. The prompt instructs the Error Tracer to focus on clinically meaningful discrepancies in imaging findings, diagnostic interpretation, and management recommendations, while ignoring non-critical linguistic differences. The report-level verification is formulated as $v_R = \mathcal{C}_\phi(R^{(1)}, Y_K^{\text{gold}}, P_{\text{report}})$, where $v_R \in \{\text{correct}, \text{incorrect}\}$. This decision serves as a gate for subsequent task-level error tracing. If $v_R = \text{incorrect}$, the Error Tracer further performs task-level verification following the predefined clinical task order. For each upstream clinical sub-task $k = 1, \ldots, K-1$, it takes the initial prediction $A_k^{(1)}$, the corresponding reference annotation $Y_k^{\text{gold}}$, and the task-specific verification prompt $P_{\text{sub},k}$ as input, yielding $v_k = \mathcal{C}_\phi(A_k^{(1)}, Y_k^{\text{gold}}, P_{\text{sub},k})$, where $v_k \in \{\text{correct}, \text{incorrect}\}$. The earliest erroneous task is then identified as $k^* = \min\{k \in \{1, \ldots, K-1\} \mid v_k = \text{incorrect}\}$. When $v_R = \text{correct}$, task-level error tracing is not activated and $k^* = \varnothing$. The corresponding report-level and task-level verification prompts are illustrated in Figure 3.

Through this two-stage verification procedure, the Error Tracer determines whether correction is required and, when necessary, localizes the earliest erroneous clinical task. Importantly, the reference annotations are used exclusively by the Error Tracer as verification targets and are never provided to the Reasoning Model as input, thereby preventing reference information leakage.

**Report-Level Verification Prompt**

You are a conservative thoracic-radiology report consistency judge. Compare the predicted CT report with the reference report. Detect clinically meaningful contradictions, not wording differences. **Check lesion presence, patient side/lobe, meaningful size category, decisive morphology, cancer/risk direction, and management tier.** A report is incorrect if a difference changes clinical characterization, risk assessment, or management. Ignore harmless extra detail, length, and sentence order. Return exactly the JSON object and nothing else: {{"report_correct": true}}
[PREDICTED REPORT]
[REFERENCE REPORT]

**Task-Level Verification Prompt**

Rules:
**Lesion Location:** compare the explicit Location answers: lesion presence, patient side, lobe/region.
**Lesion Impression:** compare lesion type, size category, and major morphology.
**Cancer Judgment:** compare Yes, No, or Indeterminate.
**Risk Assessment:** compare A-E risk category.
**Targeted Therapy:** compare management tier and follow-up interval. Return exactly the JSON object and nothing else: {{"wrong_tasks":["Location","Impression"]}}
[PREDICTED COT]
[REFERENCE COT]

Figure 3: Prompt design for the two-stage Error Tracer. The report-level prompt first guides the Error Tracer to evaluate the clinical consistency of the generated report, while the task-level prompt further guides it to identify the earliest erroneous clinical sub-task for targeted correction.

### 3.3.2 Error-Localized Visual Cue Extraction

Identifying the erroneous clinical task alone is insufficient for reliable correction, as the Reasoning Model may still lack task-relevant visual information needed to revise the error. We therefore associate each clinical task $k$ with a predefined set of segmentation text queries, $\mathcal{P}_{\text{seg},k} = \{P_{\text{seg},k,1}, \ldots, P_{\text{seg},k,J_k}\}$, where $J_k$ denotes the number of queries associated with task $k$. Unlike the verification prompts used by the Error Tracer, which contain complete instructions for assessing clinical consistency, the segmentation queries are short semantic phrases referring to anatomical or lesion-related regions, such as `lung lesion`, `pulmonary vessels`, and `bronchi`. These queries are processed by a frozen visual model $\mathcal{V}$ with text-guided semantic segmentation capability (Carion et al., 2026). Once the earliest erroneous task $k^*$ is identified, the task-specific query sets for $k^*$ and the subsequent clinical tasks are used to generate task-relevant masks as $\boldsymbol{M}_{k,j} = \mathcal{V}(\boldsymbol{I}, P_{\text{seg},k,j}), j = 1, \ldots, J_k,\ k = k^*, \ldots, K$. Each mask $\boldsymbol{M}_{k,j}$ localizes the region associated with the corresponding text query without requiring additional point or bounding-box annotations. Representative queries are summarized in Table 1.

The resulting masks are further combined with the original CT image and its DICOM metadata $D$ to derive clinically interpretable visual cues, including anatomical location, lesion geometry, spatial relationships, and attenuation characteristics. Pixel spacing is used to convert image-space measurements into physical dimensions, while the DICOM rescale parameters are used to recover

Hounsfield unit (HU) values. The task-specific visual cue extraction process is formulated as $\mathcal{H}_k = \mathcal{G}_k\left(\boldsymbol{I}, \{\boldsymbol{M}_{k,j}\}_{j=1}^{J_k}, D\right),\ k = k^*, \ldots, K$, where $\mathcal{G}_k$ denotes the cue extraction function associated with task $k$, and $\mathcal{H}_k$ represents the resulting visual cue used for corrective reasoning.

Table 1: Representative text queries used for error-localized visual cue extraction. The complete query library is provided in Appendix C.

| Segmentation Region | Example Text Queries |
|---|---|
| **Lung Field** | `bilateral lungs; left and right lung fields...` |
| **Lesion** | `lung lesion; pulmonary nodule...` |

**(a) Pulmonary and Lesion Regions**

| Segmentation Region | Example Text Queries |
|---|---|
| **Pulmonary Vessels** | `pulmonary vessels near the lung lesion...` |
| **Airways** | `bronchial structures near the lung lesion...` |

**(b) Perilesional Vessel and Airway Regions**

### 3.4 Targeted Reasoning Refinement

Once $k^*$ is identified, VEC-CT resumes reasoning from the erroneous task while retaining the initial predictions for all preceding tasks. The corrective reasoning process is formulated as $A_k^{(2)} = \mathcal{T}_\theta\left(\boldsymbol{I}, Q_k, \mathcal{A}_{<k^*}^{(1)}, \mathcal{A}_{k^*:k-1}^{(2)}, \mathcal{H}_k\right)$, for $k = k^*, \ldots, K$. Here, $\mathcal{A}_{<k^*}^{(1)}$ denotes the initial predictions for tasks before the earliest erroneous task, which are retained during correction, while $\mathcal{A}_{k^*:k-1}^{(2)} = \{A_{k^*}^{(2)}, \ldots, A_{k-1}^{(2)}\}$ contains the predictions generated during the corrective reasoning stage before task $k$. For $k = k^*$, $\mathcal{A}_{k^*:k-1}^{(2)} = \varnothing$. Each task from $k^*$ to $K$ receives its corresponding visual cue $\mathcal{H}_k$. Here, $A_k^{(2)}$ denotes the corrected prediction for task $k$.

For a sample with $k^* \neq \varnothing$, the corrective reasoning objective supervises the erroneous task and all downstream tasks as $\mathcal{L}_{\text{correct}} = \ell_{\text{NLL}}\left(Y_{k^*}^{\text{gold}}; \boldsymbol{I}, Q_{k^*}, \mathcal{A}_{<k^*}^{(1)}, \mathcal{H}_{k^*}\right) + \sum_{k=k^*+1}^{K} \ell_{\text{NLL}}\left(Y_k^{\text{gold}}; \boldsymbol{I}, Q_k, \mathcal{A}_{<k^*}^{(1)}, \mathcal{A}_{k^*:k-1}^{(2)}, \mathcal{H}_k\right)$. The overall training objective combines supervision from the initial reasoning trajectory and the targeted corrective trajectory as $\mathcal{L} = \mathcal{L}_{\text{reason}} + \lambda \mathbb{I}[k^* \neq \varnothing]\mathcal{L}_{\text{correct}}$, where $\lambda$ controls the contribution of corrective supervision and $\mathbb{I}[\cdot]$ is an indicator function. Samples whose initial reports are verified as correct contribute only to $\mathcal{L}_{\text{reason}}$.

The same Reasoning Model parameters $\theta$ are shared across the initial and corrective reasoning stages. Task-specific visual cues are used only as corrective supervision during training, helping the model revise erroneous clinical sub-tasks. At inference, only the optimized Reasoning Model is retained, while the Error Tracer and Visual Cue Extractor are not used. The model independently performs the predefined clinical tasks to generate the final diagnostic report.

## 4 Experiments

### 4.1 Dataset and Implementation

We train and evaluate VEC-CT on LungCT-CoT, comprising two public lung CT datasets (Armato et al., 2011; Li et al., 2020) and one private dataset. Selected 2D CT slices are paired with task-specific VQA pairs, yielding over 100,000 pairs. The dataset is constructed through generation and verification, with original clinical annotations and manual review used to ensure data quality. We further adopt patient-level splitting to prevent data leakage. The Reasoning Model is trained with LoRA while keeping its pretrained visual encoder frozen, using four NVIDIA V100 GPUs. Dataset construction and training details are provided in Appendices A and B, respectively.

### 4.2 Main Results

To comprehensively evaluate VEC-CT, we compare it with three categories of representative baselines: **(1) General-purpose MLLMs**, **(2) Medical Specialist MLLMs**, and **(3) Radiology-oriented MLLMs**. The generated diagnostic reports are evaluated from both lexical and semantic perspectives. BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) are used to mea-

sure lexical overlap with the reference reports, while METEOR (Banerjee & Lavie, 2005) and BERTScore (Zhang et al., 2020) provide complementary evaluation of semantic consistency.

Table 2: Comparison of VEC-CT with representative methods on the diagnostic report generation task. The best results are highlighted in bold.

| Method | BLEU-2 | BLEU-3 | BLEU-4 | ROUGE-L | METEOR | BERT-F1 |
|---|---|---|---|---|---|---|
| *General-purpose MLLMs* | | | | | | |
| Qwen3-VL-8B (2025) | 0.13±0.09 | 0.07±0.06 | 0.05±0.06 | 0.21±0.09 | 0.25±0.10 | 0.69±0.05 |
| InternVL3.5-8B (2025) | 0.07±0.04 | 0.03±0.03 | 0.02±0.02 | 0.14±0.04 | 0.23±0.06 | 0.64±0.03 |
| LLaVA-v1.6-7B (2024a) | 0.05±0.02 | 0.03±0.02 | 0.01±0.01 | 0.11±0.04 | 0.21±0.06 | 0.63±0.03 |
| Qwen3.5-9B (2026) | 0.06±0.04 | 0.02±0.02 | 0.01±0.01 | 0.15±0.05 | 0.16±0.06 | 0.61±0.02 |
| *Medical Specialist MLLMs* | | | | | | |
| Lingshu-7B (2026) | 0.11±0.06 | 0.06±0.04 | 0.03±0.03 | 0.17±0.05 | 0.23±0.08 | 0.68±0.04 |
| MedGemma-1.5-4B-IT (2026) | 0.06±0.05 | 0.03±0.04 | 0.02±0.03 | 0.12±0.06 | 0.14±0.06 | 0.64±0.06 |
| Qwen3.5-2B-MedVL (2026) | 0.09±0.06 | 0.04±0.04 | 0.02±0.02 | 0.16±0.07 | 0.17±0.06 | 0.67±0.05 |
| Hulu-Med-4B (2025) | 0.12±0.05 | 0.07±0.04 | 0.03±0.03 | 0.18±0.04 | 0.27±0.07 | 0.68±0.03 |
| LLaVAMed1.5-7B (2023a) | 0.06±0.04 | 0.03±0.03 | 0.02±0.02 | 0.14±0.05 | 0.11±0.04 | 0.60±0.04 |
| MedVLThinker-7B (2026) | 0.06±0.04 | 0.02±0.02 | 0.01±0.01 | 0.13±0.07 | 0.15±0.06 | 0.63±0.07 |
| *Radiology-oriented MLLMs* | | | | | | |
| OmniCT (2026) | 0.06±0.03 | 0.03±0.02 | 0.02±0.02 | 0.11±0.04 | 0.20±0.05 | 0.62±0.04 |
| BiMediX2 (2025) | 0.10±0.05 | 0.05±0.04 | 0.03±0.03 | 0.17±0.05 | 0.19±0.05 | 0.64±0.03 |
| RaDialog (2026) | 0.09±0.05 | 0.04±0.03 | 0.02±0.02 | 0.13±0.04 | 0.19±0.05 | 0.66±0.03 |
| *Ours* | | | | | | |
| **VEC-CT** | **0.39**±0.25 | **0.33**±0.25 | **0.28**±0.24 | **0.47**±0.23 | **0.47**±0.24 | **0.81**±0.09 |

Table 3: Performance on the Risk Assessment (*Malignancy*) and Cancer Judgment (*Cancer*) sub-tasks of Clinical CoT.

| Method | Malignancy Acc. | Cancer Acc. |
|---|---|---|
| *General-purpose MLLMs* | | |
| Qwen3-VL-8B (2025) | 0.14 | 0.26 |
| InternVL3.5-8B (2025) | 0.27 | 0.35 |
| LLaVA-v1.6-7B (2024a) | 0.19 | 0.31 |
| Qwen3.5-9B (2026) | 0.10 | 0.39 |
| *Medical Specialist MLLMs* | | |
| Lingshu-7B (2026) | 0.25 | 0.36 |
| MedGemma-1.5-4B-IT (2026) | 0.15 | 0.39 |
| Qwen3.5-2B-MedVL (2026) | 0.14 | 0.26 |
| Hulu-Med-4B (2025) | 0.16 | 0.25 |
| LLaVAMed1.5-7B (2023a) | 0.14 | 0.40 |
| MedVLThinker-7B (2026) | 0.25 | 0.25 |
| *Radiology-oriented MLLMs* | | |
| OmniCT (2026) | 0.14 | 0.35 |
| BiMediX2 (2025) | 0.25 | 0.41 |
| RaDialog (2026) | 0.33 | 0.33 |
| **VEC-CT (Ours)** | **0.62** | **0.59** |

As shown in Table 2, VEC-CT achieves consistent improvements across both lexical and semantic evaluation metrics. The gains in BLEU and ROUGE reflect closer conformity to the reference reporting style, as further illustrated by the report examples in Figure 4 (a). Meanwhile, higher METEOR and BERTScore scores indicate stronger semantic consistency with the reference findings, suggesting that the improvement extends beyond surface-level wording to clinically relevant report content. The qualitative results further reveal representative failure modes in the compared models, including missing subtle lesions and overestimating suspicious morphological findings. The attention maps in Figure 4 (b) show that several comparison models focus on broader pulmonary regions rather than the target lesion, corresponding to these erroneous findings. This suggests that VEC-CT also achieves more lesion-focused visual grounding and clinically relevant reasoning. Overall, VEC-CT produces reports that are more consistent with the reference in both expression and clinical content.

Beyond report generation, we further evaluate the Cancer Judgment and Risk Assessment tasks in Table 3. Across the baselines, the Risk Assessment task generally remains more challenging, reflecting the additional difficulty of distinguishing fine-grained malignancy levels from imaging findings. VEC-CT improves both tasks, with a particularly clear gain on Risk Assessment, supporting the

benefit of structured intermediate clinical reasoning for diagnostic judgment beyond final report generation.

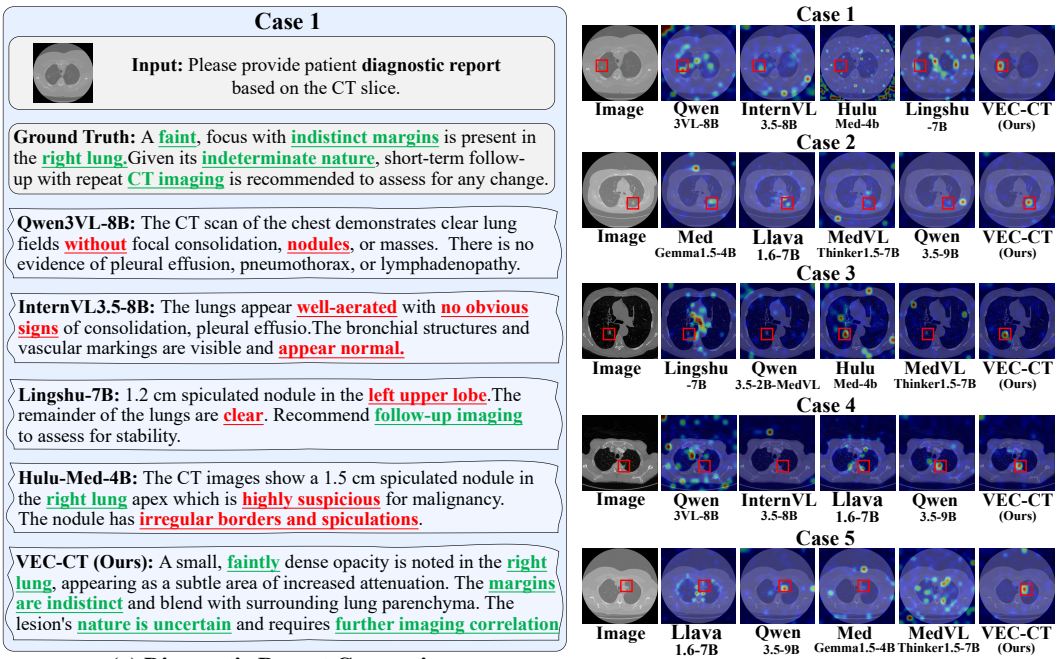



Figure 4: Qualitative comparison of VEC-CT and representative MLLMs. Case 1 corresponds to the same clinical case in (a) and (b), while Cases 2–5 provide additional attention-map examples.

### 4.3 Ablation Study

To assess the contribution of each component in VEC-CT, we conduct progressive ablation studies on fine-tuning (FT), Clinical CoT, and error correction. As shown in Table 4, VEC-CT (w/o FT) performs the worst, while report-only fine-tuning in VEC-CT (w/o CoT) yields only limited improvements. These results indicate that relying solely on the model's pretrained capabilities is insufficient for complex diagnostic tasks. Although FT provides some improvement, supervision only on the final output, without explicitly modeling the intermediate reasoning process, leads to limited gains. Introducing Clinical CoT in VEC-CT (w/o Correction) further improves performance, demonstrating the benefit of explicit supervision for intermediate clinical reasoning. With error correction, the full VEC-CT (Ours) achieves the best overall performance, showing that structured supervision can be further strengthened by targeted correction of erroneous reasoning steps.

Table 4: Ablation study on different components of VEC-CT for diagnostic report generation. The best results are highlighted in bold.

| Method | BLEU-2 | BLEU-3 | BLEU-4 | ROUGE-L | METEOR | BERT-F1 |
|---|---|---|---|---|---|---|
| VEC-CT (w/o FT) | 0.13±0.09 | 0.07±0.06 | 0.05±0.06 | 0.20±0.08 | 0.24±0.09 | 0.69±0.04 |
| VEC-CT (w/o CoT) | 0.20±0.21 | 0.16±0.21 | 0.12±0.21 | 0.31±0.19 | 0.34±0.22 | 0.71±0.08 |
| VEC-CT (w/o Correction) | 0.33±0.20 | 0.26±0.20 | 0.20±0.19 | 0.40±0.19 | 0.42±0.20 | 0.78±0.08 |
| VEC-CT (Pre-Cue) | 0.33±0.27 | 0.27±0.26 | 0.23±0.25 | 0.42±0.23 | 0.41±0.26 | 0.78±0.09 |
| **VEC-CT (Ours)** | **0.39**±0.25 | **0.33**±0.25 | **0.28**±0.24 | **0.47**±0.23 | **0.47**±0.24 | **0.81**±0.09 |

We further investigate the timing of visual cue provision. Table 4 shows that providing visual cues throughout a single-pass reasoning process does not consistently lead to substantial improvements. VEC-CT (Pre-Cue) remains close to VEC-CT (w/o Correction) on both the overall metrics and the Lesion Location, Lesion Impression, and Targeted Therapy sub-tasks in Figure 5 (b)–(d), whereas

VEC-CT (Ours) achieves more consistent improvements. This comparison suggests that the effectiveness of visual cues depends not only on their availability, but also on their timing, with cue provision after error localization fostering internalized error-aware reasoning.

Table 5: Ablation study of VEC-CT on the Risk Assessment (*Malignancy*) and Cancer Judgment (*Cancer*) tasks. The best results are highlighted in bold.

| Method | Malignancy Acc. | Cancer Acc. |
|---|---|---|
| VEC-CT (w/o FT) | 0.14 | 0.26 |
| VEC-CT (w/o CoT) | 0.35 | 0.27 |
| VEC-CT (w/o Correction) | 0.37 | 0.43 |
| VEC-CT (Pre-Cue) | 0.41 | 0.47 |
| **VEC-CT (Ours)** | **0.62** | **0.59** |

The textual attention patterns in Figure 5 (a) provide further insight into this performance difference. Both VEC-CT (w/o Correction) and VEC-CT (Pre-Cue) exhibit stronger attention toward the beginning and end of the reasoning sequence, while showing limited focus on fine-grained intermediate clinical attributes, particularly those in the Lesion Impression task. This suggests that visual cues alone may not sufficiently highlight key intermediate clinical attributes.

In contrast, VEC-CT (Ours) exhibits clearer task-dependent attention across the clinical sub-tasks, with greater emphasis on fine-grained intermediate clinical information relevant to the current task. This pattern indicates that the Reasoning Model has learned to selectively attend to task-relevant clinical information during inference, even without visual cues or explicit correction.

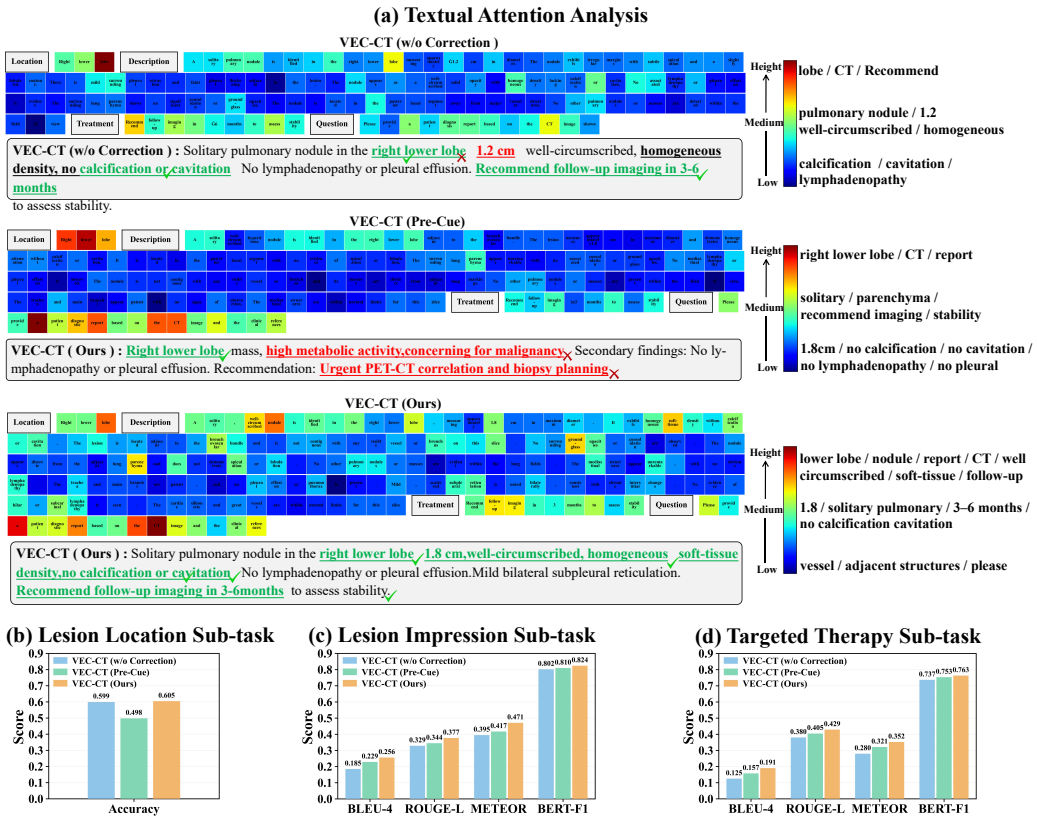



Figure 5: Analysis of VEC-CT variants. (a) Textual attention comparison among VEC-CT (w/o Correction), VEC-CT (Pre-Cue), and VEC-CT (Ours). (b)–(d) Comparative results for the Lesion Location, Lesion Impression, and Targeted Therapy sub-tasks, respectively.

## 5 Conclusion

Lung CT report generation with MLLMs remains challenged by unreliable visual perception and the propagation of errors across dependent clinical reasoning steps. To address these issues, we introduce LungCT-CoT for structured clinical supervision and propose VEC-CT, which combines task-specific visual cues with targeted error correction to improve intermediate clinical decisions. Experiments on multiple lung CT datasets demonstrate consistent improvements in diagnostic report generation and clinical decision tasks, supporting the effectiveness of structured reasoning and targeted correction.

## AI USE STATEMENT

Generative AI tools were used to assist with language editing, LaTeX formatting, and software code development. They were not used to generate experimental data or fabricate experimental results. All AI-assisted text and code were reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

## ETHICS STATEMENT

This work complies with the ethical requirements of ICLR. The study does not involve any human or animal experiments. All datasets, including LungCT-CoT, were obtained and used in strict accordance with relevant regulations and data-use requirements, with careful protection of data privacy. We ensure that the research process is free from bias and discrimination, does not include personally identifiable information, and does not involve experiments that pose safety risks. We are committed to maintaining transparency and integrity throughout the research process.

## REPRODUCIBILITY STATEMENT

To support reproducibility, we provide detailed descriptions of the LungCT-CoT data sources, construction and verification procedures, task composition, and patient-level splits in Appendix A. Appendix B provides model configurations, training settings, visual cue extraction procedures, and inference details. Anonymized source code, configuration files, and evaluation scripts are provided as supplementary material. Public datasets can be obtained from their original sources, while the private clinical dataset is subject to the corresponding data-use and privacy restrictions.

## REFERENCES


Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/960a172bc7fbf0177ccccbb411a7d800-Abstract-Conference.html.

Samuel G. Armato, Geoffrey McLennan, Luc Bidaut, Michael F. McNitt-Gray, Charles R. Meyer, Anthony P. Reeves, Binsheng Zhao, Denise R. Aberle, Claudia I. Henschke, Eric A. Hoffman, et al. The lung image database consortium (LIDC) and image database resource initiative (IDRI): A completed reference database of lung nodules on CT scans. *Medical Physics*, 38(2): 915–931, 2011. doi: 10.1118/1.3528204. URL https://pubmed.ncbi.nlm.nih.gov/21452728/.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72. Association for Computational Linguistics, 2005. URL https://aclanthology.org/W05-0909/.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. SAM 3: Segment anything with concepts. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=r35clVtGzw.

Xupeng Chen, Zhixin Lai, Kangrui Ruan, Shichu Chen, Jiaxiang Liu, and Zuozhu Liu. R-LLaVA: Improving Med-VQA understanding through visual region of interest. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–10, 2025. doi: 10.1109/IJCNN64981.2025.11227685.

Xiaoke Huang, Juncheng Wu, Hui Liu, Xianfeng Tang, and Yuyin Zhou. Medvlthinker: Simple baselines for multimodal medical reasoning. In *Proceedings of the Fifth Machine Learning for Health Symposium*, volume 297 of *Proceedings of Machine Learning Research*, pp. 384–398. PMLR, 2026.

Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, et al. Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*, 2025. URL https://arxiv.org/abs/2510.08668.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems*, volume 36, pp. 28541–28564, 2023a. doi: 10.52202/075280-1240. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/5abcdf8ecdcacba028c6662789194572-Abstract-Datasets_and_Benchmarks.html.

Jinsong Li, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Jiaqi Wang, and Dahua Lin. Visual self-refine: A pixel-guided paradigm for accurate chart parsing. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 85499–85522, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/89b0e466b46292ce0bfe53618aadd3de-Paper-Conference.pdf.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023b. URL https://proceedings.mlr.press/v202/li23q.html.

Peihua Li, Shijie Wang, Tao Li, Jian Lu, Yuyang HuangFu, and Dazhe Wang. A large-scale CT and PET/CT dataset for lung cancer diagnosis (Lung-PET-CT-Dx), 2020. URL https://www.cancerimagingarchive.net/collection/lung-pet-ct-dx/.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81. Association for Computational Linguistics, 2004. URL https://aclanthology.org/W04-1013/.

Tianwei Lin, Zhongwei Qiu, Wenqiao Zhang, Jiang Liu, Yihan Xie, Mingjian Gao, Zhenxuan Fan, Zhaocheng Li, Sijing Li, Zhongle Xie, Peng Lu, Yueting Zhuang, Ling Zhang, Beng Chin Ooi, and Yingda Xia. Omnict: Towards a unified slice-volume lvlm for comprehensive ct analysis. In *International Conference on Learning Representations*, 2026.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024a.

Jiaxiang Liu, Yuan Wang, Jiawei Du, Joey Tianyi Zhou, and Zuozhu Liu. MedCoT: Medical chain of thought via hierarchical expert. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17371–17389. Association for Computational Linguistics, 2024b. doi: 10.18653/v1/2024.emnlp-main.962. URL https://aclanthology.org/2024.emnlp-main.962/.

Heber MacMahon, David P. Naidich, Jin Mo Goo, Kyung Soo Lee, Ann N. C. Leung, John R. Mayo, Atul C. Mehta, Yoshiharu Ohno, Charles A. Powell, Mathias Prokop, Geoffrey D. Rubin, Cornelia M. Schaefer-Prokop, William D. Travis, Paul E. Van Schil, and Alexander A. Bankier. Guidelines for management of incidental pulmonary nodules detected on ct images: From the fleischner society 2017. *Radiology*, 284(1):228–243, 2017. doi: 10.1148/radiol.2017161659.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pp. 353–367. PMLR, 2023. URL https://proceedings.mlr.press/v225/moor23a.html.

Sahal Shaji Mullappilly, Mohammed Irfan Kurpath, Sara Pieri, Saeed Yahya Alseiari, Shanavas Cholakkal, Khaled M Aldahmani, Fahad Shahbaz Khan, Rao Muhammad Anwer, Salman Khan, Timothy Baldwin, and Hisham Cholakkal. Bimedix2: Bio-medical expert lmm for diverse medical modalities. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 14051–14071. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.756.

OpenMed. Qwen3.5-2B-MedVL model card. Hugging Face model card, 2026. URL https://huggingface.co/OpenMed/Qwen3.5-2B-MedVL. Accessed 2026-08-28.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318. Association for Computational Linguistics, 2002. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Chantal Pellegrini, Ege Özsoy, Benjamin Busam, Benedikt Wiestler, Nassir Navab, and Matthias Keicher. Radialog: Large vision-language models for x-ray reporting and dialog-driven assistance. In *Proceedings of the 8th International Conference on Medical Imaging with Deep Learning*, volume 301, pp. 1294–1312. PMLR, 2026.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Andrew Sellergren, Chufan Gao, Fereshteh Mahvar, Timo Kohlberger, Fayaz Jamil, Madeleine Traverse, Alberto Tono, Bashir Sadjad, Lin Yang, Charles Lau, et al. Medgemma 1.5 technical report. *arXiv preprint arXiv:2604.05081*, 2026. URL https://arxiv.org/abs/2604.05081.

Mohammad Shahab Sepehri, Zalan Fabian, Maryam Soltanolkotabi, and Mahdi Soltanolkotabi. MediConfusion: Can you trust your AI radiologist? probing the reliability of multimodal medical foundation models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=H9UnNgdq0g.

Haoran Sun, Yankai Jiang, Wenjie Lou, Yujie Zhang, Wenjie Li, Lilong Wang, Mianxin Liu, Lei Liu, and Xiaosong Wang. Chiron-o1: Igniting multimodal large language models towards generalizable medical reasoning via mentor-intern collaborative search. In *Advances in Neural Information Processing Systems*, volume 38, pp. 115328–115365, 2025.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. URL https://arxiv.org/abs/2508.18265.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, Yu Sun, Junao Shen, Chaojun Wang, Can Hu, Siwei Pan, Jie Tan, Tingyang Xu, Deli Zhao, Hao Zhang, and Yu Rong. Lingshu: Generalist foundation model for unified multimodal medical understanding and reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP, 2026. doi: 10.1109/TPAMI.2026.3730310.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. How language model hallucinations can snowball. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 59670–59684. PMLR, 2024a. URL https://proceedings.mlr.press/v235/zhang24ay.html.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*, 2024b. URL https://openreview.net/forum?id=y1pPWFVfuR.

Feizhong Zhou, Xingyue Liu, Zhiying Yang, Zhipeng Li, and Hanguang Xiao. Enhancing medical mllms with dual vision encoders and moe-based modality projector. *Knowledge-Based Systems*, 336:115275, 2026. doi: 10.1016/j.knosys.2026.115275.

To provide a quick overview of the supplementary material, we briefly summarize the content of each appendix section below for easier reference.

**LungCT-CoT Dataset Details** describes the data sources and the complete construction and verification process of LungCT-CoT, together with the resulting task-level data statistics (Appendix A).

**Implementation Details** presents the model configurations and the complete training and inference procedure of VEC-CT (Appendix B).

**Prompt Templates** presents the prompts used for Clinical CoT reasoning and error tracing (Appendix C).

**Additional Experimental Results and Analysis** extends the main experiments with complete quantitative results, qualitative comparisons, and failure-case analysis (Appendix D).

**Limitations and Additional Discussion** summarizes the current limitations of VEC-CT and discusses potential directions for future work (Appendix E).

# A LungCT-CoT Dataset Details

## A.1 Dataset Sources

LungCT-CoT is constructed from two publicly available datasets and one private clinical dataset. The details of these datasets are described below.

LIDC-IDRI (Armato et al., 2011) is a public chest CT dataset designed for pulmonary nodule detection and diagnosis, comprising 1,018 cases. Each case includes a chest CT scan and its corresponding XML annotation file. The dataset was annotated by four experienced thoracic radiologists using a two-stage process: initially, each radiologist independently identified and marked pulmonary lesions. Subsequently, they re-evaluated the CT scans and provided a final judgment based on the anonymized annotations of their peers. Lesions were categorized into three types: nodules with diameters of 3 mm or greater, nodules with diameters smaller than 3 mm, and non-nodules with diameters of 3 mm or greater. These comprehensive nodule-level annotations provide a foundation for developing clinical tasks such as lesion localization, morphological characterization, and malignancy risk assessment.

Lung-PET-CT-Dx (Li et al., 2020) is a large-scale CT and PET/CT dataset for lung cancer diagnosis, containing 436 imaging studies and 251,135 images from 355 subjects. The dataset comprises patients with suspected lung cancer who underwent standard clinical lung biopsies and PET/CT scans, with cases classified according to histopathological diagnoses. Unlike LIDC-IDRI, which focuses primarily on pulmonary nodules, Lung-PET-CT-Dx provides bounding-box annotations for tumor locations alongside corresponding pathological diagnosis information. Tumor location annotations were reviewed by five thoracic radiologists experienced in lung cancer diagnosis. One radiologist performed the initial annotation, while the others verified the results. This information provides clearer supervision for lesion localization and diagnosis, facilitating the development of clinical tasks such as lesion localization and diagnostic reporting.

In addition, we incorporated a private pulmonary CT dataset derived from clinical practice to further enrich the diversity of imaging appearances and clinical scenarios within LungCT-CoT. To protect patient privacy, we do not publicly disclose patient-level information or further details regarding this private dataset.

## A.2 Data Construction

LungCT-CoT contains six clinically ordered tasks, as defined in Section 3.1. This section details the construction and verification process used to obtain task-specific VQA pairs.

As shown in Figure 6 (a), we employ Qwen3-VL-8B-Instruct to automatically construct initial task-specific visual question answering (VQA) pairs. For CT data from diverse sources, we first perform standardized image preprocessing and select representative CT images from each case as visual inputs. We then design task-specific prompts for each clinical task, in which specific clinical expert roles and task requirements are defined to guide the model in analyzing the image according to

the corresponding task. For instance, in the *Lesion Location* task, the model assumes the role of a tumor imaging specialist and focuses on lesion localization. Similar roles and task requirements are defined for the remaining tasks.

For each CT sample, the image and corresponding task prompt are jointly fed into the model to generate a clinically relevant question and its corresponding answer. By adopting a unified structured output format across tasks, each CT image is transformed into multiple task-specific VQA pairs corresponding to different clinical objectives. This MLLM-based automated construction reduces the labor required for large-scale medical VQA data construction and enables efficient expansion of diverse clinical questions through standardized task-specific prompts.

However, the VQA pairs generated at this stage serve merely as initial data. Their clinical reliability is not fully guaranteed. On one hand, variations across data sources may introduce source-specific biases into the generated results. On the other hand, MLLMs themselves are susceptible to hallucinations, potentially leading to issues such as incorrect lesion localization, inaccurate descriptions of imaging features, or diagnoses lacking sufficient visual support. Therefore, rather than using the automatically generated QA pairs directly as the final LungCT-CoT dataset, we subject these initial samples to further verification and refinement during a subsequent validation phase.

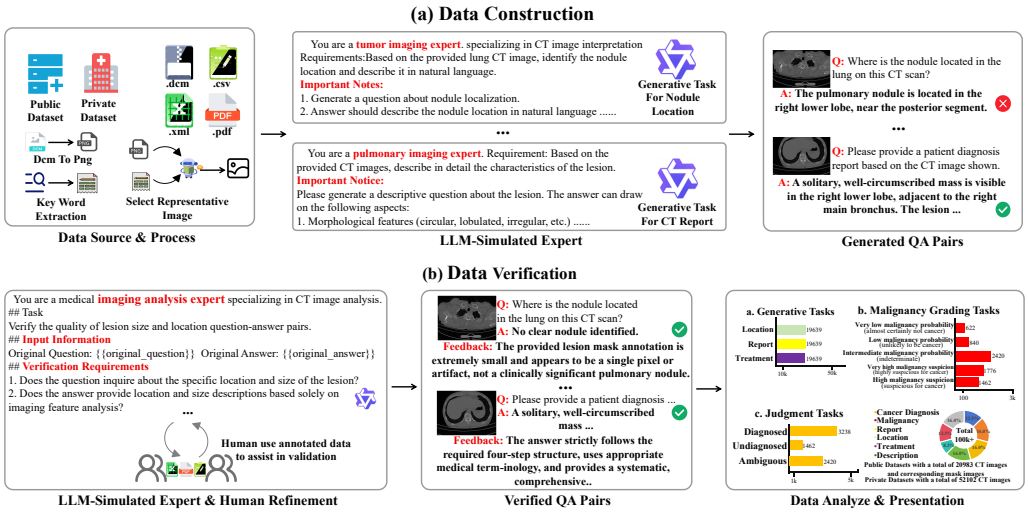



Figure 6: Overview of the LungCT-CoT data construction and verification pipeline. (a) Task-specific prompts guide the MLLM to generate initial VQA pairs from multi-source lung CT images. (b) Source annotations and verification prompts are used to verify and refine the generated pairs, followed by human review.

### A.3 Data Verification

We designed a data verification process to inspect and correct the initial QA pairs generated in the previous stage, thereby enhancing the data quality of LungCT-CoT.

As shown in Figure 6 (b), we utilize annotations from various sources as the basis for verification and design specific verification prompts tailored to each task. For each initial QA pair, the model receives the original question, the original answer, and the relevant annotation information. It is then instantiated as a medical imaging verification expert to assess the consistency between the question and the answer, as well as the clinical consistency between the answer and existing annotations. The verification process focuses on task-relevant details such as lesion location, size, morphological characteristics, and diagnostic information.

When a generated QA pair aligns with the corresponding annotations, we retain the original sample. When significant inconsistencies are detected, the model corrects the problematic answer based on the annotation information rather than regenerating the entire data sample. In this way, the verification stage specifically addresses erroneous samples arising from the automated construction process, correcting inconsistencies with existing medical annotations while preserving valid QA pairs.

Given that automated verification itself may still yield errors, we further conduct manual inspection and refinement of the verified data, focusing on samples that are difficult for the model to assess reliably or that contain significant ambiguities. Ultimately, the QA pairs that have undergone both MLLM-based verification and human refinement are retained for the construction of LungCT-CoT.

### A.4 Dataset Analysis

Following data construction and verification, we obtained a total of over 100,000 clinical QA pairs. LungCT-CoT encompasses six categories of clinical tasks: Lesion Location, Lesion Impression, Cancer Judgment, Risk Assessment, Targeted Therapy, and Diagnostic Finding. Based on their output formats, these tasks can be further classified into four generative tasks and two categorical tasks. To prevent patient-level data leakage, all samples from the same patient are assigned to the same split, resulting in 21,043 training samples, 927 validation samples, and 751 test samples.

Among these, Lesion Location, Lesion Impression, Targeted Therapy, and Diagnostic Finding are generative tasks, with each task comprising 19,639 QA pairs for lesion localization, imaging feature description, clinical treatment recommendations, and final diagnostic report generation, respectively. Cancer Judgment determines the cancer diagnosis status of each case and consists of three categories: *Yes*, *No*, and *Indeterminate*, with 3,238, 1,462, and 2,420 samples, respectively. The Risk Assessment task estimates the malignancy risk of lesions across five levels from low to high, with 622, 840, 2,420, 1,776, and 1,462 samples, respectively.

## B Implementation Details

### B.1 Reasoning Model Configuration

We employ Qwen3-VL-8B-Instruct as the backbone of the Reasoning Model and perform parameter-efficient fine-tuning using Low-Rank Adaptation (LoRA). During training, the pretrained visual encoder is kept frozen, while only the LoRA parameters on the language model side are optimized. Specifically, LoRA is applied to the `q_proj`, `k_proj`, `v_proj`, and `o_proj` modules in the attention layers, as well as the `gate_proj`, `up_proj`, and `down_proj` modules in the feed-forward layers. Both the first-round Clinical Chain-of-Thought (CoT) reasoning and the second-round refined reasoning are performed by the same Reasoning Model with shared parameters, allowing the model to jointly learn structured clinical reasoning and error-corrective reasoning.

### B.2 Error Tracer Configuration

The Error Tracer is built upon a frozen Qwen3-VL-8B-Instruct model and is used exclusively for verification and error tracing during training, without participating in parameter optimization. During training, the Error Tracer reviews the Reasoning Model's first-round predictions and uses the corresponding gold-standard annotations as internal references to identify erroneous clinical subtasks and locate the earliest clinical task at which the initial reasoning deviates. The Error Tracer only provides error localization and does not directly reveal the correct answers to the Reasoning Model. The subsequent refined reasoning is therefore still performed by the Reasoning Model based on the CT image and the corresponding visual cues.

### B.3 Visual Cue Extractor Configuration

We employ a locally deployed SAM3 model as the Visual Cue Extractor and keep it frozen throughout visual cue extraction. All visual cues are extracted offline and stored before VEC-CT training. The precomputed results are directly loaded during training, and the Visual Cue Extractor is therefore not jointly optimized with the Reasoning Model. The segmentation confidence threshold is set to 0.4. The minimum and maximum area fractions for local candidate regions are set to 0.0001 and 0.12, respectively, while the minimum area fraction for the global region is set to 0.05. The crop padding around the lesion is set to 30 pixels, and the search radius for auxiliary regions near the lesion is set to 24 pixels. For samples with paired DICOM information, the width of the perilesional region is set to 5.0 mm.

For CT samples with valid paired DICOM information, we use `PixelSpacing` to convert lesion measurements from pixel space into physical units, including the long diameter, short diameter, equivalent diameter, lesion area, and distance to the lung boundary. In addition, the original pixel values are converted into Hounsfield units (HU) using `RescaleSlope` and `RescaleIntercept`, enabling quantitative characterization of the lesion and its surrounding lung tissue. For samples without valid paired DICOM information, geometric measurements, including lesion area, diameters, and spatial location, are retained in pixel space without conversion to physical units.

### B.4 Training Hyperparameters

VEC-CT is trained for one epoch on four NVIDIA V100 GPUs using distributed training. The batch size per device and gradient accumulation steps are both set to 1. We use the AdamW optimizer with a LoRA learning rate of $5 \times 10^{-6}$, a warmup ratio of 0.03, and a maximum gradient norm of 1.0. The LoRA rank $r$ is set to 16, with $\alpha = 32$ and a dropout rate of 0.05. The random seed is fixed to 42. Training is performed using FP16 precision with non-reentrant gradient checkpointing to reduce GPU memory consumption.

### B.5 Inference Settings

During inference, we use only the trained **Reasoning Model**, without involving the Error Tracer, Visual Cue Extractor, or visual cues used during training. For each test sample, the Reasoning Model takes the original CT image as visual input and performs Clinical CoT reasoning following the predefined sequence of clinical tasks. Predictions from preceding tasks are provided as contextual information for subsequent relevant tasks, progressively leading to the final diagnostic report. No gold-standard annotations are provided to the model during inference. They are retained only for performance evaluation after prediction.

For the specific inference configuration, we load the trained LoRA checkpoint and set the model to evaluation mode. All generative tasks use deterministic decoding (`do_sample=False`) with KV caching enabled. No parameter updates are performed throughout the inference process.

### B.6 Training and Inference Algorithms

**Algorithm 1** Training Procedure of VEC-CT

**Require:** Training set $\mathcal{S}_{\text{train}}$; initialized Reasoning Model $\mathcal{T}_\theta$; frozen Error Tracer $\mathcal{C}_\phi$; frozen Visual Cue Extractor $\mathcal{V}$; clinical task sequence $\mathcal{Q} = \{Q_1, \ldots, Q_K\}$; each training sample contains a CT image $\boldsymbol{I}$, DICOM metadata $D$, and gold-standard responses $\{Y_k^{\text{gold}}\}_{k=1}^K$

**for** each training sample $(\boldsymbol{I}, D, \{Y_k^{\text{gold}}\}_{k=1}^K) \in \mathcal{S}_{\text{train}}$ **do**
  Initialize first-round reasoning context $\mathcal{A}_{<1}^{(1)} \leftarrow \varnothing$
  **for** $k = 1$ to $K$ **do**
    $A_k^{(1)} \leftarrow \mathcal{T}_\theta(\boldsymbol{I}, Q_k, \mathcal{A}_{<k}^{(1)})$
  **end for**
  $R^{(1)} \leftarrow A_K^{(1)}$
  Compute first-round reasoning loss $\mathcal{L}_{\text{reason}}$ between $\{A_k^{(1)}\}_{k=1}^K$ and $\{Y_k^{\text{gold}}\}_{k=1}^K$
  $v_R \leftarrow \mathcal{C}_\phi(R^{(1)}, Y_K^{\text{gold}}, P_{\text{report}})$
  **if** $v_R = \text{incorrect}$ **then**
    Identify the earliest erroneous task
    $k^* \leftarrow \min\left\{k \mid \mathcal{C}_\phi(A_k^{(1)}, Y_k^{\text{gold}}, P_{\text{sub},k}) = \text{incorrect}\right\}$
    **for** $k = k^*$ to $K$ **do**
      Activate the task-specific prompt set $\mathcal{P}_{\text{seg},k} = \{P_{\text{seg},k,1}, \ldots, P_{\text{seg},k,J_k}\}$
      **for** $j = 1$ to $J_k$ **do**
        $\boldsymbol{M}_{k,j} \leftarrow \mathcal{V}(\boldsymbol{I}, P_{\text{seg},k,j})$
      **end for**
      Extract task-specific visual cues $\mathcal{G}_k\left(\boldsymbol{I}, \{\boldsymbol{M}_{k,j}\}_{j=1}^{J_k}, D\right)$
      **if** $k = k^*$ **then**
        $A_k^{(2)} \leftarrow \mathcal{T}_\theta\left(\boldsymbol{I}, Q_k, \mathcal{A}_{<k^*}^{(1)}, \mathcal{H}_k\right)$
      **else**
        $A_k^{(2)} \leftarrow \mathcal{T}_\theta\left(\boldsymbol{I}, Q_k, \mathcal{A}_{<k^*}^{(1)}, \mathcal{A}_{k^*:k-1}^{(2)}, \mathcal{H}_k\right)$
      **end if**
    **end for**
    Compute correction loss $\mathcal{L}_{\text{correct}}$ between $\{A_k^{(2)}\}_{k=k^*}^K$ and $\{Y_k^{\text{gold}}\}_{k=k^*}^K$
  **else**
    $\mathcal{L}_{\text{correct}} \leftarrow 0$
  **end if**
  $\mathcal{L} \leftarrow \mathcal{L}_{\text{reason}} + \lambda \mathcal{L}_{\text{correct}}$
  Update Reasoning Model parameters $\theta$ using $\mathcal{L}$
**end for**
**return** Trained Reasoning Model $\mathcal{T}_\theta$

## C Prompt Templates

### C.1 Prompts for Initial Diagnostic Reasoning

During initial reasoning, the Reasoning Model sequentially performs the six predefined clinical sub-tasks in LungCT-CoT. Each sub-task is associated with a fixed task-specific prompt specifying its clinical objective and output requirements. The complete prompt templates are shown in Figure 7.

**Location Prompt (Initial Diagnostic Reasoning)**

You are a senior thoracic radiologist specialized in lung CT interpretation.
Follow a step-by-step clinical workflow. At the current step, answer only the requested subtask.
The [Previous Answers] section contains earlier step answers. Re-check the CT image and do not invent unsupported findings.
Return only the requested answer, without markdown headings, bullet points, or hidden reasoning.
Question:Where is the nodule located in the lung based on the CT scan?
Output:Return only the anatomical location.
Rule:Use patient anatomical left/right, not screen position.
Answer:

(a) Lesion Location

**Description Prompt (Initial Diagnostic Reasoning)**

You are a senior thoracic radiologist specialized in lung CT interpretation.
Follow a step-by-step clinical workflow. At the current step, answer only the requested subtask.
The [Previous Answers] section contains earlier step answers. Re-check the CT image and do not invent unsupported findings.
Return only the requested answer, without markdown headings, bullet points, or hidden reasoning.
[Previous Answers] :{predicted_location}
Question:Please describe the characteristics of the lesion.
Output:Return a concise morphology description including size, shape, margin, density, and surrounding changes when visible.
Rule:
Do not report pixel values, pixel areas, or convert pixels to mm/cm.
If physical scale is unavailable, describe size qualitatively from the CT image.
Answer:

(b) Lesion Impression

**Cancer Diagnosis Prompt (Initial Diagnostic Reasoning)**

You are a senior thoracic radiologist specialized in lung CT interpretation.
Follow a step-by-step clinical workflow. At the current step, answer only the requested subtask.
The [Previous Answers] section contains earlier step answers. Re-check the CT image and do not invent unsupported findings.
Return only the requested answer, without markdown headings, bullet points, or hidden reasoning.
[Previous Answers] :{predicted_description}
Question:Is there any evidence of cancer based on CT scan images of the lungs?
Output:Return exactly one of: Yes, No, Indeterminate.
Answer:

(c) Cancer Judgment

**Malignancy Prompt (Initial Diagnostic Reasoning)**

You are a senior thoracic radiologist specialized in lung CT interpretation.
Follow a step-by-step clinical workflow. At the current step, answer only the requested subtask.
The [Previous Answers] section contains earlier step answers. Re-check the CT image and do not invent unsupported findings.
Return only the requested answer, without markdown headings, bullet points, or hidden reasoning.
[Previous Answers] :{predicted_description}{predicted_cancer_diagnosis}
Question:
Based on the lung CT imaging, assess malignancy.
Select A. Highly Unlikely B. Moderately Unlikely
C. Indeterminate D. Moderately Suspicious
E. Highly Suspicious.
Output:Return exactly one letter: A, B, C, D, or E.
Answer:

(d) Risk Assessment

**Treatment Prompt (Initial Diagnostic Reasoning)**

You are a senior thoracic radiologist specialized in lung CT interpretation.
Follow a step-by-step clinical workflow. At the current step, answer only the requested subtask.
The [Previous Answers] section contains earlier step answers. Re-check the CT image and do not invent unsupported findings.
Return only the requested answer, without markdown headings, bullet points, or hidden reasoning.
[Previous Answers] :{predicted_description} {diagnosis_risk_summary}
Question:What is the recommended follow-up treatment plan for the nodule detected in CT?
Output:Return a concise management recommendation appropriate to the imaging risk level.
Answer:

(e) Targeted Therapy

**Report Prompt (Initial Diagnostic Reasoning)**

You are a senior thoracic radiologist specialized in lung CT interpretation.
Follow a step-by-step clinical workflow. At the current step, answer only the requested subtask.
The [Previous Answers] section contains earlier step answers. Re-check the CT image and do not invent unsupported findings.
Return only the requested answer, without markdown headings, bullet points, or hidden reasoning.
[Previous Answers] :{predicted_description}{diagnosis_risk_summary} {predicted_treatment}
Question:Please provide a patient diagnostic report based on the CT image and the clinical references.
Output:Return one concise natural-language report combining findings, impression,and recommendation, without headings or lists.
Answer:

(f) Diagnostic Finding

Figure 7: Task-specific prompt templates used during initial reasoning: (a) Lesion Location, (b) Lesion Impression, (c) Cancer Judgment, (d) Risk Assessment, (e) Targeted Therapy, and (f) Diagnostic Finding.

### C.2 Complete Text Queries for Visual Cue Extraction

The Visual Cue Extractor uses a predefined library of text queries to guide SAM3 toward anatomical and lesion-related regions in lung CT images. The query library is organized into four categories: **Global**, **Local**, **Vessel**, and **Airway**. Global queries provide lung-level anatomical context, while local queries focus on pulmonary lesions and their imaging appearances. Vessel and airway queries are additionally used to identify candidate anatomical structures surrounding the lesion. These queries are fixed textual inputs and are not optimized during training. The complete query library is provided in Table 6.

Table 6: Complete SAM3 text query library used for visual cue extraction.

| Segmentation Region | Complete Text Queries |
|---|---|
| **Lung Field** | `left and right lung fields; bilateral lungs; entire lungs; lung parenchyma; all suspicious pulmonary abnormalities; pulmonary nodules consolidation ground-glass opacity and other lung lesions; lungs excluding chest wall mediastinum heart and extrapulmonary tissue` |
| **Lesion** | `lung lesion; pulmonary nodule; lung nodule; ground-glass pulmonary opacity; pulmonary consolidation; solid component of a lung lesion; pulmonary lesion with spiculation or lobulation` |
| **Pulmonary Vessels** | `pulmonary blood vessels; lung blood vessels; pulmonary vessels near the lung lesion; pulmonary vasculature; lung vasculature; branching pulmonary vessels; pulmonary arterial and venous branches; intrapulmonary vessels; vessels within the lung fields; blood vessel structures in the lungs` |
| **Airways** | `bronchi; airway; bronchial structures near the lung lesion; bronchial tree; intrapulmonary bronchi; airway branches in the lungs; bronchovascular structures; segmental bronchi` |

The extracted task-specific visual cues are further illustrated in Figure 8. These cues summarize lesion-related information derived from the segmented regions and provide targeted visual guidance for corrective reasoning.

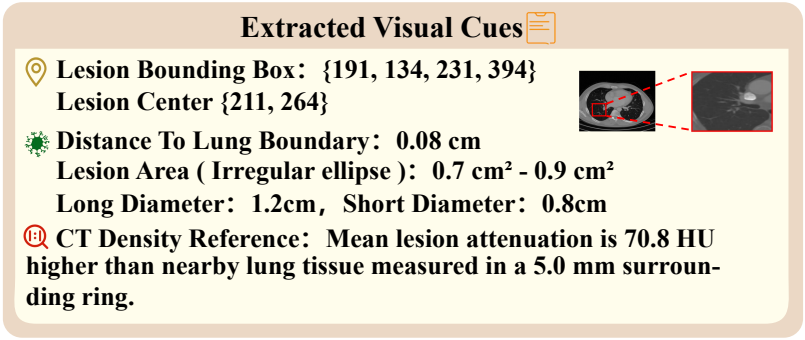



Figure 8: Representative visual cues extracted from task-specific anatomical and lesion-related regions for corrective reasoning.

### C.3 Prompts for Targeted Reasoning Refinement

During targeted reasoning refinement, the Reasoning Model resumes the clinical reasoning process from the erroneous sub-task identified by the Error Tracer. Predictions preceding the identified error are retained as context, while the current and subsequent tasks are reconsidered with additional visual information extracted for correction. Compared with the initial-round prompts, the refinement prompts explicitly instruct the model to re-examine the CT image, treat the provided visual cues as auxiliary evidence, and avoid mechanically inheriting potentially unreliable previous predictions. The task-specific refinement templates are shown in Figure 9.

**Location Prompt (Targeted Reasoning Refinement)**

You are a senior thoracic radiologist specialized in lung CT interpretation.
Follow a step-by-step clinical workflow. At the current step, answer only the requested subtask.
The [Previous Answers] section contains earlier step answers. Re-check the CT image and do not invent unsupported findings.
Return only the requested answer, without markdown headings, bullet points, or hidden reasoning.
You are revising one unreliable step in a lung-nodule reasoning chain.
Assess the displayed CT image(s) independently.
External lesion evidence is only an auxiliary reference for localization or quantitative context; it may be incomplete or imperfect and must not replace image-based clinical judgment.

[Reference Evidence]

Question:Where is the nodule located in the lung based on the CT scan?
Output:Return only the anatomical location.
Rule:
Use patient anatomical left/right, not screen position.
Use the CT image as the primary basis for judgment.
The reference evidence is auxiliary; use it only when consistent with the visible CT appearance.
Do not copy it mechanically or infer unsupported clinical conclusions.
Answer:

(a) Lesion Location

**Description Prompt (Targeted Reasoning Refinement)**

You are a senior thoracic radiologist specialized in lung CT interpretation.
Follow a step-by-step clinical workflow. At the current step, answer only the requested subtask.
The [Previous Answers] section contains earlier step answers. Re-check the CT image and do not invent unsupported findings.
Return only the requested answer, without markdown headings, bullet points, or hidden reasoning.
You are revising one unreliable step in a lung-nodule reasoning chain.
Assess the displayed CT image(s) independently.
External lesion evidence is only an auxiliary reference for localization or quantitative context; it may be incomplete or imperfect and must not replace image-based clinical judgment.
[Previous Answers] :{current_location}

[Reference Evidence]

Question:Please describe the characteristics of the lesion.
Output:Return a concise morphology description including size, shape, margin, density, and surrounding changes when visible.
Rule:
Do not report pixel values, pixel areas, or convert pixels to mm/cm.
If physical scale is unavailable, describe size qualitatively from the CT image.
Use the CT image as the primary basis for judgment.
The reference evidence is auxiliary; use it only when consistent with the visible CT appearance.
Do not copy it mechanically or infer unsupported clinical conclusions.
The supplied DICOM measurements provide quantitative context only.
They do not by themselves establish morphology, malignancy, or treatment.
Answer:

(b) Lesion Impression

**CancerDiagnosis Promp (Targeted Reasoning Refinement)**

You are a senior thoracic radiologist specialized in lung CT interpretation.
Follow a step-by-step clinical workflow. At the current step, answer only the requested subtask.
You are revising an unreliable Cancer Diagnosis decision.
Re-assess the CT image and lesion region independently.
Use the preceding lesion description as clinical context, but do not simply inherit its diagnostic conclusion.
[Previous Answers]: {current_description}

[Reference Evidence]

Question:Is there any evidence of cancer based on CT scan images of the lungs?
Output:Return exactly one of : Yes / No / Indeterminate
Rule:
Base the decision primarily on the visible CT findings.
Use the lesion crop only as additional visual support.
Do not infer a definite cancer diagnosis when the imaging evidence is insufficient or ambiguous.
Do not introduce findings that are not supported by the CT image.
Answer:

(c) Cancer Judgment

**Malignancy Prompt (Targeted Reasoning Refinement)**

You are a senior thoracic radiologist specialized in lung CT interpretation.
Follow a step-by-step clinical workflow. At the current step, answer only the requested subtask.
You are revising an unreliable malignancy assessment.
Re-assess the CT image and lesion region independently.
Use the preceding lesion description and cancer diagnosis as clinical context, while making the malignancy judgment from the visible imaging findings.
[Previous Answers]: {current_description}{current_cancer_diagnosis}

[Reference Evidence]

Question:Based on the lung CT imaging, assess malignancy.
Options: A. Highly Unlikely B. Moderately Unlikely
C. Indeterminate D. Moderately Suspicious E. Highly Suspicious
Output:Return exactly one letter: A, B, C, D, or E.
Rule:
Base the malignancy assessment primarily on the visible CT findings.
Consider lesion morphology and other suspicious imaging characteristics.
Use the lesion crop as additional visual support.
Do not mechanically infer malignancy from the preceding CancerDiagnosis.
Do not introduce unsupported clinical findings.
Answer:

(d) Risk Assessment

**Treatment Prompt (Targeted Reasoning Refinement)**

You are a senior thoracic radiologist specialized in lung CT interpretation.
Follow a step-by-step clinical workflow. At the current step, answer only the requested subtask.
The [Previous Answers] section contains earlier step answers. Re-check the CT image and do not invent unsupported findings.
Return only the requested answer, without markdown headings, bullet points, or hidden reasoning.
You are revising one unreliable step in a lung-nodule reasoning chain.
Assess the displayed CT image(s) independently.
External lesion evidence is only an auxiliary reference for localization or quantitative context; it may be incomplete or imperfect and must not replace image-based clinical judgment.
[Previous Answers]: {current_description}{current_diagnosis_risk_summary}

[Reference Evidence]

Question:What is the recommended follow-up treatment plan for the nodule detected in CT?
Output:Return a concise management recommendation appropriate to the imaging risk level.
Rule:Use the CT image as the primary basis for judgment.
The reference evidence is auxiliary; use it only when consistent with the visible CT appearance.
Do not copy it mechanically or infer unsupported clinical conclusions.
The supplied DICOM measurements provide quantitative context only.
They do not by themselves establish morphology, malignancy, or treatment.
Answer:

(e) Targeted Therapy

**Report Prompt (Targeted Reasoning Refinement)**

You are a senior thoracic radiologist specialized in lung CT interpretation.
Follow a step-by-step clinical workflow. At the current step, answer only the requested subtask.
The [Previous Answers] section contains earlier step answers. Re-check the CT image and do not invent unsupported findings.
Return only the requested answer, without markdown headings, bullet points, or hidden reasoning.
You are revising one unreliable step in a lung-nodule reasoning chain.
Assess the displayed CT image(s) independently.
External lesion evidence is only an auxiliary reference for localization or quantitative context; it may be incomplete or imperfect and must not replace image-based clinical judgment.

[Previous Answers]: {current_description}{current_diagnosis_risk_summary}
{current_treatment}

[Reference Evidence]

Question:Please provide a patient diagnostic report based on the CT image and the clinical references.
Output:Return one concise natural-language report combining findings, impression, and recommendation, without headings or lists.
Rule:
Use the CT image as the primary basis for judgment.
The reference evidence is auxiliary; use it only when consistent with the visible CT appearance.
Do not copy it mechanically or infer unsupported clinical conclusions.
The supplied DICOM measurements provide quantitative context only.
They do not by themselves establish morphology, malignancy, or treatment.
Answer:

(f) Diagnostic Finding

Figure 9: Task-specific prompt templates used during targeted reasoning refinement: (a) Lesion Location, (b) Lesion Impression, (c) Cancer Judgment, (d) Risk Assessment, (e) Targeted Therapy, and (f) Diagnostic Finding.

# D Additional Experimental Results and Analysis

## D.1 Evaluation Protocol

To ensure consistent and fair comparisons across different models, all baselines are evaluated on the same test set under a unified task setting. We directly use the publicly released pretrained or instruction-tuned weights of each baseline without additional training or fine-tuning on LungCT-CoT. Since different models employ distinct architectures and visual processing pipelines, we retain their native image preprocessing and visual input interfaces while keeping the test samples, task definitions, and output spaces consistent.

Specifically, we evaluate all baselines on three tasks: Diagnostic Finding, Cancer Judgment, and Risk Assessment. For each test sample, a baseline model receives only the original CT image and the prompt corresponding to the current task. The three tasks are evaluated independently, without using predictions from other tasks as additional context. To reduce evaluation bias caused by unconstrained output formats, we impose consistent task-specific output constraints. For Cancer Judgment, the output space is restricted to *Yes*, *No*, and *Indeterminate*, while Risk Assessment requires a single label from *A* to *E*. For the Diagnostic Finding task, models are instructed to output only the final diagnostic report without additional reasoning traces, JSON structures, or other auxiliary content. Classification tasks use deterministic decoding with a maximum generation length of 16 tokens, while the maximum generation length for Diagnostic Finding is set to 512 tokens.

## D.2 Evaluation Metrics

We evaluate model performance from three complementary perspectives: lexical overlap, semantic consistency, and clinical prediction accuracy. For the Diagnostic Finding task, we employ BLEU, ROUGE, METEOR, and BERTScore. BLEU measures n-gram precision between the generated and reference reports, and we report BLEU-1 through BLEU-4. ROUGE evaluates lexical and sequential overlap from a recall-oriented perspective, for which ROUGE-1, ROUGE-2, and ROUGE-L are reported. METEOR further evaluates the overall correspondence between generated and reference texts based on word-level matching. BERTScore complements these lexical metrics by measuring semantic similarity between generated and reference reports using contextualized language representations.

For the clinical prediction tasks, we use accuracy as the evaluation metric. Cancer Judgment evaluates predictions over three diagnostic categories: *Yes*, *No*, and *Indeterminate*. The Risk Assessment task evaluates predictions over five malignancy-risk categories from *A* to *E*. For all report-generation and clinical prediction metrics, higher values indicate better performance. All metrics are computed by comparing model predictions with the corresponding gold-standard annotations on the same test set.

## D.3 Complete Experimental Results

To provide a more comprehensive evaluation, we additionally report the full set of BLEU and ROUGE metrics that are omitted from the main paper. These supplementary results enable a more complete assessment of report generation performance across different evaluation dimensions.

For traditional text generation metrics, including BLEU-1–4 and ROUGE-1/2/L, VEC-CT consistently outperforms the compared methods, demonstrating higher textual consistency between the generated and reference reports. For semantic evaluation metrics, including METEOR and BERTScore, VEC-CT also achieves the best performance, indicating that its advantages extend beyond textual matching to better preserve the semantic consistency of diagnostic reports. Overall, the complete experimental results further support the main findings reported in the main paper.

Table 7: Complete comparison of VEC-CT with three categories of baseline models on diagnostic report generation. We report BLEU-1–4, ROUGE-1, ROUGE-2, ROUGE-L, METEOR, and BERT-F1. The best results are highlighted in bold.

| Method | BLEU | | | | ROUGE | | | Semantic | |
|---|---|---|---|---|---|---|---|---|---|
| | B-1 | B-2 | B-3 | B-4 | R-1 | R-2 | R-L | MET | BERT-F1 |
| *General-purpose MLLMs* | | | | | | | | | |
| Qwen3-VL-8B (2025) | 0.25±0.12 | 0.13±0.09 | 0.07±0.07 | 0.05±0.06 | 0.29±0.11 | 0.09±0.08 | 0.21±0.09 | 0.25±0.10 | 0.69±0.05 |
| InternVL3.5-8B (2025) | 0.16±0.05 | 0.07±0.04 | 0.03±0.03 | 0.02±0.02 | 0.21±0.06 | 0.05±0.03 | 0.14±0.04 | 0.23±0.06 | 0.64±0.03 |
| LLaVA-v1.6-7B (2024a) | 0.11±0.04 | 0.05±0.02 | 0.03±0.02 | 0.01±0.01 | 0.17±0.06 | 0.04±0.02 | 0.11±0.04 | 0.21±0.06 | 0.63±0.03 |
| Qwen3.5-9B (2026) | 0.19±0.06 | 0.06±0.04 | 0.02±0.02 | 0.01±0.01 | 0.22±0.07 | 0.04±0.03 | 0.15±0.05 | 0.16±0.06 | 0.61±0.02 |
| *Medical Specialist MLLMs* | | | | | | | | | |
| Lingshu-7B (2026) | 0.23±0.07 | 0.11±0.06 | 0.06±0.04 | 0.03±0.03 | 0.26±0.08 | 0.07±0.05 | 0.17±0.05 | 0.23±0.08 | 0.68±0.04 |
| MedGemma-1.5-4B-IT (2026) | 0.12±0.09 | 0.06±0.05 | 0.03±0.03 | 0.02±0.02 | 0.13±0.09 | 0.03±0.03 | 0.12±0.06 | 0.14±0.06 | 0.64±0.06 |
| Qwen3.5-2B-MedVL (2026) | 0.21±0.11 | 0.09±0.06 | 0.04±0.04 | 0.02±0.02 | 0.22±0.09 | 0.04±0.04 | 0.16±0.07 | 0.17±0.06 | 0.67±0.05 |
| Hulu-Med-4B (2025) | 0.25±0.07 | 0.12±0.05 | 0.07±0.04 | 0.03±0.03 | 0.28±0.06 | 0.09±0.04 | 0.18±0.04 | 0.27±0.07 | 0.68±0.03 |
| LLaVAMed1.5-7B (2023a) | 0.15±0.06 | 0.06±0.04 | 0.03±0.03 | 0.02±0.02 | 0.18±0.07 | 0.04±0.03 | 0.14±0.05 | 0.11±0.04 | 0.60±0.04 |
| MedVLThinker-7B (2026) | 0.18±0.11 | 0.06±0.04 | 0.02±0.02 | 0.01±0.01 | 0.18±0.09 | 0.03±0.03 | 0.13±0.07 | 0.15±0.06 | 0.63±0.07 |
| *Radiology-oriented MLLMs* | | | | | | | | | |
| OmniCT (2026) | 0.13±0.06 | 0.06±0.03 | 0.03±0.02 | 0.02±0.02 | 0.18±0.07 | 0.05±0.03 | 0.11±0.04 | 0.20±0.05 | 0.62±0.04 |
| BiMediX2 (2025) | 0.22±0.07 | 0.10±0.05 | 0.05±0.04 | 0.03±0.03 | 0.24±0.07 | 0.06±0.04 | 0.17±0.05 | 0.19±0.05 | 0.64±0.03 |
| RaDialog (2026) | 0.23±0.07 | 0.09±0.05 | 0.04±0.03 | 0.02±0.02 | 0.21±0.06 | 0.04±0.03 | 0.13±0.04 | 0.19±0.05 | 0.66±0.03 |
| *Ours* | | | | | | | | | |
| **VEC-CT** | **0.49**±0.25 | **0.39**±0.25 | **0.33**±0.25 | **0.28**±0.25 | **0.54**±0.22 | **0.35**±0.25 | **0.47**±0.23 | **0.47**±0.24 | **0.81**±0.09 |

### D.4 Comprehensive Qualitative Analysis

To provide a more complete view of model behavior under the same input, we present the full diagnostic reports generated by VEC-CT and all comparison models on the same test sample, together with their corresponding attention maps. We further present the outputs and attention maps of different VEC-CT ablation variants on the same sample, allowing differences in both report generation and visual responses to be examined.

#### D.4.1 Comparison with Baseline Models

Figure 10 presents the complete diagnostic reports generated by VEC-CT and all comparison models on the same test sample. Although different models are able to identify some pulmonary imaging findings, their generated reports exhibit varying degrees of inconsistency with the ground truth. Some models incorrectly identify the lesion location, while others generate clinical descriptions that are inconsistent with the observed imaging findings. In comparison, the report generated by VEC-CT shows greater consistency with the ground truth, preserving the major lesion-related findings while reducing erroneous descriptions observed in the comparison models.

We further visualize the attention maps of the models on the same test sample. As shown in Figure 11, the comparison models exhibit noticeably different visual responses to the same CT image. Several models show dispersed attention or strong responses outside the target lesion region. In comparison, VEC-CT shows a stronger response around the target lesion indicated by the red bounding box, with better spatial correspondence to the target region.

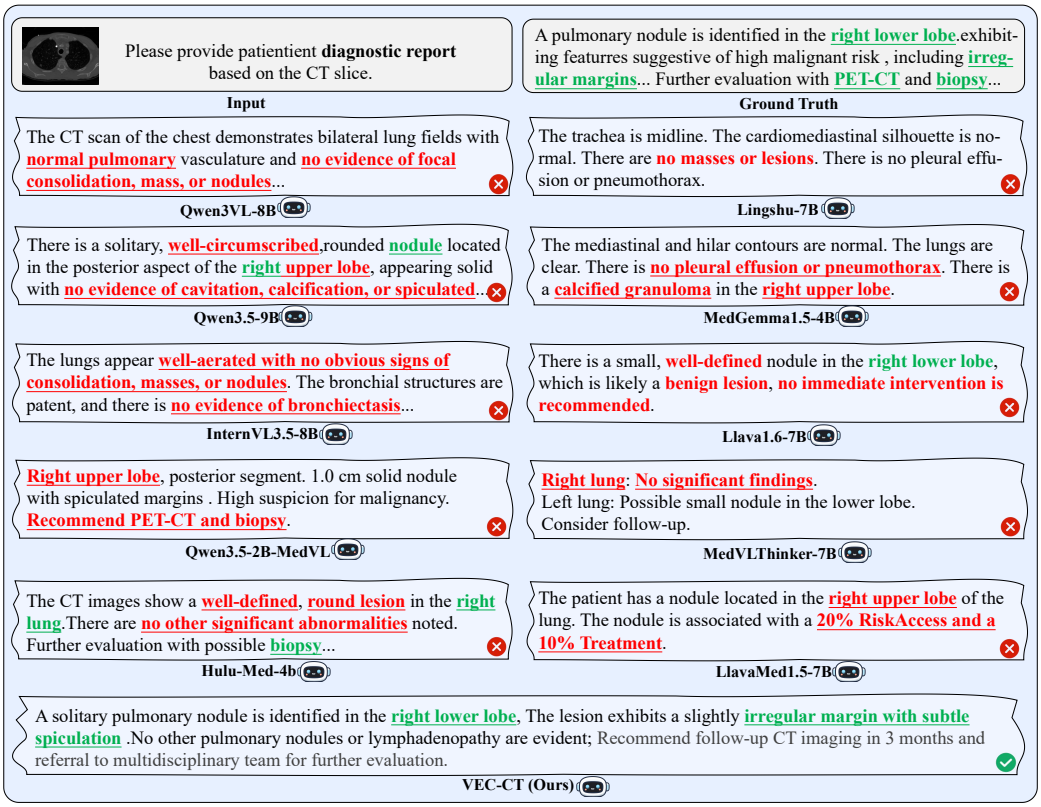



Figure 10: Diagnostic reports generated by VEC-CT and all comparison models on the same test sample. Green highlights indicate findings consistent with the ground truth, while red highlights indicate erroneous or inconsistent clinical findings.

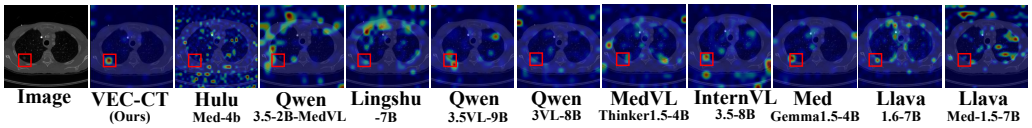



Figure 11: Attention map comparison between VEC-CT and comparison models. The red boxes indicate the target lesion region.

### D.4.2 Ablation Variant Analysis

Figure 12 further presents the complete diagnostic reports generated by different VEC-CT ablation variants on the same sample. Without task-specific fine-tuning, the generated report shows substantial differences from the ground truth. After introducing Clinical CoT, more findings in the generated report are consistent with the ground truth, while some erroneous descriptions are reduced. With the further incorporation of the correction mechanism and visual cues, additional inconsistencies are removed. Among these variants, the complete VEC-CT produces a report with greater consistency with the ground truth and preserves more lesion-related information.

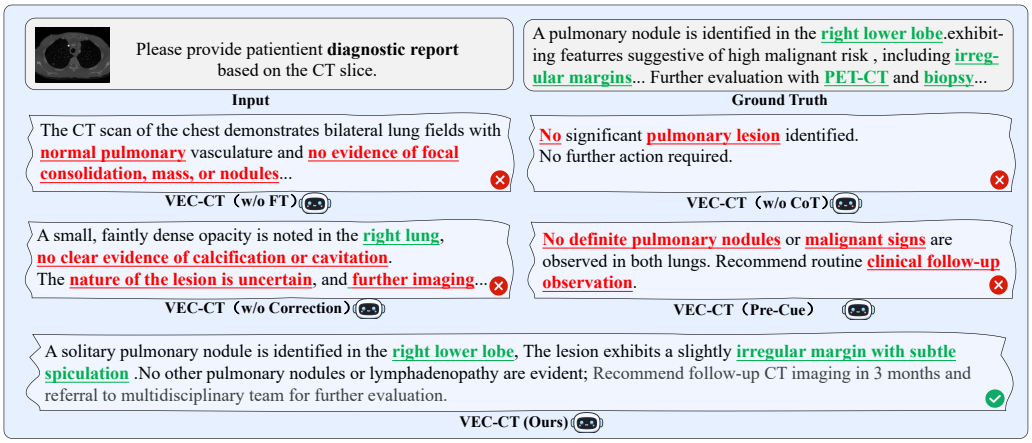



Figure 12: Diagnostic reports generated by different VEC-CT ablation variants on the same test sample. Green highlights indicate findings consistent with the ground truth, while red highlights indicate erroneous or inconsistent clinical findings.

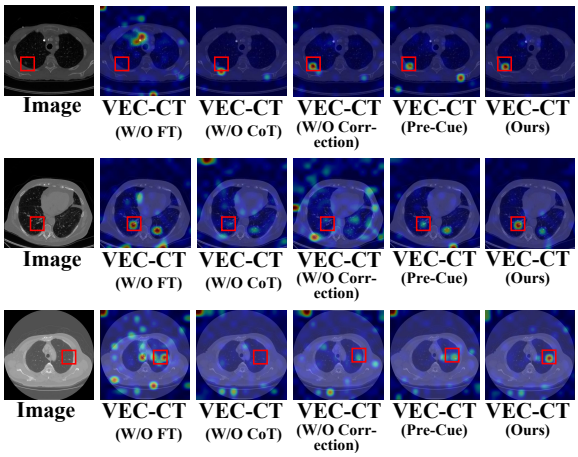



Figure 13: Attention maps of different VEC-CT ablation variants. The red boxes indicate the target lesion.

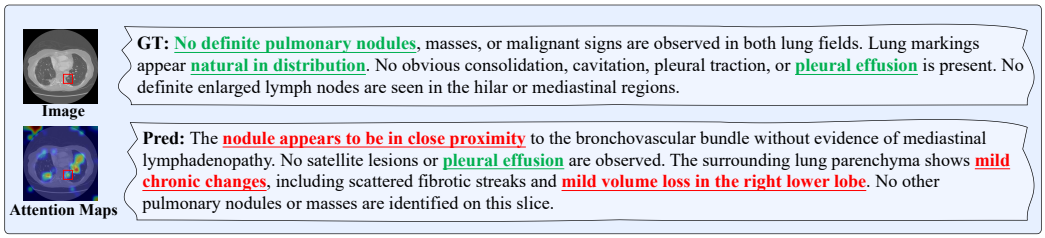
Figure 13 further presents the attention maps of different VEC-CT ablation variants. The variants exhibit noticeable differences in their visual responses around the target lesion. As different components are incorporated, the visual responses become less dispersed and show greater correspondence with the target region. The complete VEC-CT exhibits a clearer and more consistent lesion-centered response across the three cases. This consistent pattern suggests that the benefit of the complete framework is not restricted to a single sample. Together with the corresponding report outputs, these visualizations suggest that differences in visual responses are accompanied by changes in the generated diagnostic content, providing complementary visual evidence for the ablation results.

It should be noted that attention maps primarily reflect the image regions to which a model responds during reasoning and do not directly indicate that all lesion characteristics have been correctly understood. A model may attend to the correct lesion region while still producing inaccurate descriptions or clinical judgments. Therefore, the attention maps are used here as complementary visual cues to illustrate differences in visual responses across comparison models and VEC-CT variants, rather than as direct evidence of diagnostic correctness.

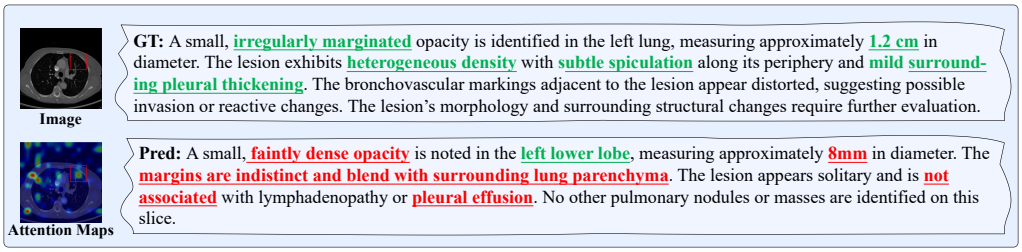
### D.5 Failure Case Analysis

Despite the overall improvements achieved by VEC-CT, we still observe four representative failure patterns, as illustrated in Figure 14.

The first failure case corresponds to inaccurate lesion localization and recognition. Although the ground truth indicates that no definite pulmonary nodule is present, the model still produces a strong response over a local region and incorrectly interprets it as a pulmonary nodule. This suggests that VEC-CT may still generate false-positive lesion predictions in negative cases or in regions with complex anatomical structures.



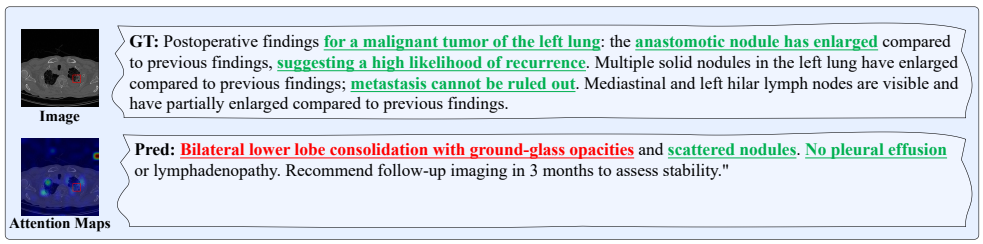

**(a) Lesion Localization Failure**



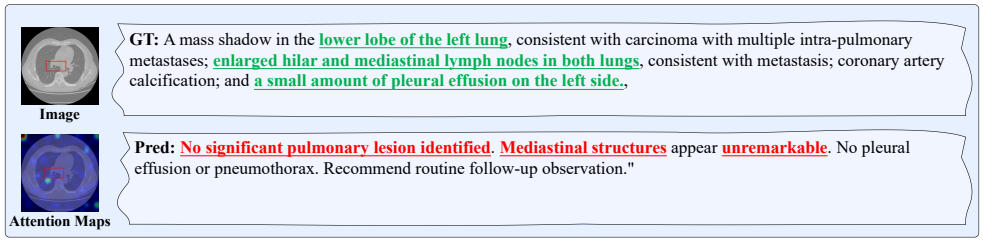

**(b) Fine-Grained Characterization Failure**

Figure 14: Representative failure cases of VEC-CT. (a) Lesion localization failure. (b) Fine-grained characterization failure. Green text indicates findings consistent with the ground truth, while red text highlights incorrect or inconsistent predictions.

The second failure case shows that correct lesion localization does not necessarily guarantee accurate lesion characterization. In this example, the model successfully attends to the target lesion region and identifies its approximate location, but the generated description remains inconsistent with the ground truth in fine-grained lesion characteristics. This indicates that the model may still have difficulty extracting and interpreting subtle visual details even when the lesion region is correctly localized.




**(c) Longitudinal Context Limitation**




**(d) Disease Extent Underestimation**

Figure 15: Additional failure cases of VEC-CT. (c) Longitudinal context limitation. (d) Disease extent underestimation. Green text indicates findings consistent with the ground truth, while red text highlights incorrect or inconsistent predictions.

The third case reflects a limitation of the current single-image setting. The reference report relies on longitudinal information, including interval enlargement of the anastomotic lesion, pulmonary nodules, and lymph nodes. Such temporal changes cannot be reliably inferred from the current CT

image alone without access to prior examinations. This failure therefore reflects missing longitudinal context rather than solely inaccurate visual perception.

The fourth case shows substantial underestimation of disease extent. The reference describes a large left lower lobe mass with multiple intrapulmonary metastases, hilar and mediastinal lymphadenopathy, and pleural effusion, whereas the model produces an almost negative report. This suggests that VEC-CT may still fail to integrate extensive abnormalities distributed across multiple anatomical regions into a complete assessment of disease burden.

Overall, these cases reveal limitations in lesion recognition, fine-grained characterization, longitudinal reasoning, and assessment of extensive multi-region disease. They further suggest that accurate spatial attention alone is not sufficient for reliable clinical reasoning.

## E Limitations and Additional Discussion

**Limitations.** Despite the promising performance of VEC-CT in report generation, several limitations remain. First, our current framework primarily relies on 2D CT images, which limits its ability to capture cross-slice continuity and three-dimensional spatial structures of lesions. Second, the current error detection and visual cue mechanisms are mainly employed during training, while dynamic error tracing and iterative refinement have not yet been incorporated into the test stage. Finally, the current training data remain limited in terms of diversity across medical centers, imaging devices, and disease types, and the generalization capability of the model in more complex clinical scenarios requires further validation.

**Future Challenges.** Future work will explore 3D volumetric CT and multi-slice inputs to enhance the model's perception of lesion spatial structures. We also aim to extend error tracing to the test stage, enabling the model to identify errors arising during clinical reasoning and progressively refine its predictions and final diagnostic reports through multiple rounds of iterative correction. Furthermore, we plan to incorporate data from more medical centers, imaging devices, and disease types to improve the generalization and robustness of VEC-CT across diverse clinical data distributions.