SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

May 2026, Vol. 69, Iss. 5, 150107:1–150107:20
https://doi.org/10.1007/s11432-025-4818-7

Special Topic: Large Multimodal Models

CT-Agent: a multimodal-LLM agent for 3D CT radiology question answering

Yuren MAO[1,2], Wenyi XU[1], Yuyang QIN[1] & Yunjun GAO[1,2*]

[1]*School of Software Technology, Zhejiang University, Hangzhou 310007, China*
[2]*Zhejiang Key Laboratory of Big Data Intelligent Computing, Hangzhou 310007, China*

Received 18 May 2025/Revised 18 September 2025/Accepted 3 November 2025/Published online 23 April 2026

Abstract Computed tomography (CT) scans can produce 3D volumetric medical data, which is viewed as hundreds of cross-sectional images (slices) and provides detailed anatomical information for diagnosis. Creating CT radiology reports is time-consuming and error-prone for radiologists. A visual question answering (VQA) system is needed to answer radiologists' anatomical questions about CT scans and to automatically generate radiology reports. However, existing VQA systems cannot adequately handle the CT radiology question answering (CTQA) task due to anatomic complexity, which makes CT images difficult to understand, and spatial relationships across hundreds of slices, which are difficult to capture. To address these challenges, this study proposes CT-Agent, a multimodal agentic framework for CTQA. CT-Agent uses anatomically independent tools to break down anatomic complexity and captures across-slice spatial relationships via global-local token compression. Experimental results on the CT-RATE and RadGenome-Chest CT datasets verify its superior performance.

Keywords LLM agent, CT radiology question answering, visual question answering, token compression, LoRA fine-tuning

Citation Mao Y R, Xu W Y, Qin Y Y, et al. CT-Agent: a multimodal-LLM agent for 3D CT radiology question answering. Sci China Inf Sci, 2026, 69(5): 150107, https://doi.org/10.1007/s11432-025-4818-7

1 Introduction

Computed tomography (CT), a three-dimensional (3D) X-ray-based medical imaging technique, is widely used in diagnosing various diseases, such as tumors, fractures, and lung diseases [1, 2]. CT provides volumetric data, offering a more complete view of a lesion's 3D structure and its spatial relationship with surrounding tissues than two-dimensional (2D) medical images. CT volumes are typically reformatted into hundreds of cross-sectional slices for volumetric data processing [3–5]. It complicates the analysis and documentation of radiology reports.

In writing radiology reports, radiologists have to examine hundreds of slices individually to identify potential abnormalities. It is time-consuming, labor-intensive, and error-prone. Radiologists increasingly need intelligent tools that provide auxiliary information to accelerate report writing and improve quality. A visual question answering (VQA) system capable of answering radiologists' anatomical questions on CT scans and automatically generating radiology reports is desired.

However, existing VQA methods [6,7] cannot meet the needs of CT volume question answering. Although recent VQA methods have significantly advanced, they mainly focus on 2D images or videos. Developing VQA systems for 3D CT volumetric data remains challenging. In CT radiology question answering (CTQA), it is crucial to accurately model anatomical structures and capture spatial relationships across slices, which presents two main challenges. First, anatomical structures are highly complex. Boundaries between organs are often unclear, and their shapes vary widely. This prevents CTQA models from understanding semantics and localizing abnormalities. Second, the spatial relationship across slices is difficult to capture. Many lesions span multiple regions and slices. Existing methods [8] that rely on single images or sparse sampling struggle to maintain a consistent global representation. In addition, CT scans contain hundreds of slices. Encoding them produces an overwhelming number of visual tokens, exceeding the capacity of current multimodal models.

To address the above challenges, we propose CT-Agent, an agentic CTQA framework based on multimodal large language models (MLLMs). CT-Agent is anatomy-aware and token-efficient. It addresses anatomical complexity using an organ-specific sub-model ensemble strategy, employing a low-rank adaptation (LoRA) plugin for each

* Corresponding author (email: gaoyj@zju.edu.cn)

anatomical region to obtain fine-grained, localized clinical clues. Furthermore, CT-Agent captures spatial relationships and reduces the number of input tokens by adopting a global-local dual-path token compression method, which reduces token length by approximately 75% while preserving semantic integrity. Using these methods, we designed the CT-Agent action space. Driven by the planning ability of LLMs, CT-Agent dynamically identifies tasks, dispatches reasoning paths, retrieves memory-based exemplars, and executes actions to generate coherent and clinically stylized responses. Experimental results on the public 3D chest CT datasets CT-RATE and RadGenome-Chest CT demonstrate that CT-Agent consistently outperforms existing methods across report generation and question answering tasks, achieving superior semantic fluency and clinical efficacy (CE).

2 Related work

Medical VQA has seen significant progress in recent years. Early methods relied on template retrieval or Transformer-based sequence generation models such as R2Gen [9] and BPI-MVQA [10], which primarily focused on 2D images such as chest X-rays. With the rise of large language models (LLMs) in the medical domain, approaches such as MAIRA [11], XrayGPT [12], and PMC-VQA [7] have introduced instruction-tuned generation to enable more flexible free-text outputs. However, these methods remain limited to 2D inputs and lack effective modeling of spatial relationships and anatomical semantics inherent in 3D volumetric data.

Recent studies on 3D CT report generation have transitioned from traditional 2D modeling paradigms toward volumetric representations that offer stronger spatial reasoning capabilities. CT2Rep [13] introduced a hierarchical memory mechanism to enable global volumetric data encoding. Models such as 3D-CT-GPT [14] and ViT3D [15] integrate 3D vision encoders with LLMs, achieving significant progress in cross-modal semantic understanding. CT-AGRG [16] adopts an abnormality-centric generation strategy to enhance clinical specificity, while Argus [17] establishes a high-resolution 3D dataset to improve fidelity to real anatomical detail. Similarly, MS-VLM [18] and M3D [19] simulate the slice-by-slice review process used by radiologists, capturing contextual dependencies across slices. Although these methods have laid a solid foundation for 3D report generation, their regional semantic integration and task generalization remain limited.

The role of LLMs in 3D medical imaging is extended by agent-based systems. Designed to automate and enhance clinical data interaction, these agents are increasingly used for tasks such as radiology report generation and answering diagnostic queries [20]. M3Builder, MedAgent-Pro, and PathFinder exemplify this trend from different perspectives [21–23]. M3Builder streamlines the end-to-end medical machine learning workflow through coordinated multi-agent collaboration; MedAgent-Pro establishes a reasoning-centered diagnostic pipeline grounded in multimodal clinical evidence; and PathFinder emulates the decision-making process of pathologists by integrating various agents to analyze whole-slide images with interpretability. Despite these advances, existing models often treat volume data as a homogeneous entity, lacking both anatomical decomposition and efficient token handling. Our proposed CT-Agent addresses these limitations through anatomy-aware reasoning and a dual-path compression strategy, making inference more scalable and interpretable.

Moreover, advancements in medical VQA and report generation have been strongly supported by diverse, well-annotated imaging datasets. In the 2D domain, resources such as VQA-RAD [24], MedMNIST [25], and MedSeg-Bench [26] provide extensive collections of X-ray, ultrasound, and pathological images paired with structured labels or question-answer (QA) pairs, enabling foundational research in image-language understanding. In 3D analysis, datasets such as CT-RATE [27] and RadGenome-Chest CT [28] have been instrumental, providing large-scale chest CT volumes with detailed radiology reports, anatomical masks, and millions of aligned QA pairs. Additionally, datasets such as MedMNISTV2 [29] and MedShapeNet [30] offer structured 3D imaging data with clinical annotations, facilitating spatially aware modeling of volumetric medical data. CT-Agent is built on two representative 3D chest CT datasets, CT-RATE [27] and RadGenome-Chest CT [28]. These provide large-scale volumetric data with region-level annotations and structured QA supervision, enabling anatomical reasoning and task-specific training.

3 Preliminaries

This section defines the 3D CTQA task, outlines the 3D CT data preprocessing procedure, and provides background on LoRA.

3.1 Problem definition

3D CTQA aims to generate clinically meaningful answers A for a user-issued natural language query Q and a volumetric CT scan I. The objective of CTQA is to learn a multimodal mapping function:

$$f : (Q, I) \to A. \tag{1}$$

Typically, CTQA systems should support two functional modes. (i) Radiology report generation. The system should conduct a comprehensive analysis of the entire CT volume and provide a radiology report. (ii) Region-guided question answering. The system should answer users' questions about one or more specific anatomical regions on a CT scan. The former emphasizes global contextual integration and summarization, while the latter demands localized semantic understanding and targeted reasoning. CTQA output should satisfy two evaluation criteria: linguistic quality and clinical correctness. Linguistic quality, which measures the fluency, coherence, and lexical similarity of the generated text with respect to ground truth, is assessed using BLEU, ROUGE-L, and METEOR. In contrast, clinical correctness measures the accuracy and relevance of the output. It is evaluated using CE metrics that assess the presence and accuracy of medically significant entities and relations in the generated output.

3.2 3D CT preprocessing

To prepare the 3D CT volumes for multimodal learning, we preprocess the data to ensure spatial consistency, standardized resolution, and anatomical focus [13]. Each volumetric scan is first resampled to a uniform voxel spacing (typically 1.5 mm along the z-axis and 1.0 mm in the axial plane) to reduce variation across scans and ensure consistent modeling. The spatial orientation of each scan is aligned with a canonical coordinate system; irrelevant background regions are removed by cropping to the foreground. Intensity values are converted into physical Hounsfield units (HU) using the rescale parameters in the DICOM metadata, enabling a consistent representation of tissue densities [2].

Following spatial and intensity normalization, each 3D CT volume is decomposed into axial slices. To extract semantically rich visual features, each slice is encoded independently using a pretrained vision transformer backbone, such as CLIP ViT-B/16 [31]. Each slice is partitioned into non-overlapping patches and mapped into a high-dimensional token space, producing a structured representation that preserves local anatomical detail. The full volume is represented as a sequence of slice-level token matrices, which serve as input for the CT-Agent's action-tool reasoning modules.

3.3 Low-rank adaptation

To efficiently adapt LLMs to domain-specific tasks such as CT radiology reasoning, full fine-tuning is often prohibitively expensive in computation and memory. To address this, parameter-efficient fine-tuning (PEFT) techniques update only a small subset of parameters while keeping the pretrained model weights fixed. Among these techniques, LoRA [32] has emerged as a widely recognized and effective method, achieving competitive performance across various tasks while significantly decreasing resource requirements [33].

LoRA modifies the weight update process by constraining updates to a low-rank decomposition. For a pretrained weight matrix $W_0 \in \mathbb{R}^{d_1 \times d_2}$, LoRA represents the update ΔW as a low-rank product $\Delta W = BA$, where $B \in \mathbb{R}^{d_1 \times d_r}$ and $A \in \mathbb{R}^{d_r \times d_2}$, with $d_r \ll \min(d_1, d_2)$. The forward pass is adjusted as $h' = W_0 x + \Delta W x = W_0 x + B(Ax)$, where x is the input. During training, W_0 remains frozen, while only the low-rank matrices B and A are updated, significantly reducing the number of trainable parameters and overall computational cost. The trained LoRA weights can be merged with the original pretrained weights or used independently as a plug-and-play module during inference. Our framework uses LoRA to develop lightweight, region-specific anatomical reasoning adapters. Each anatomical region (e.g., the lung and heart) is equipped with a LoRA module, enabling computationally efficient, anatomically specialized fine-tuning.

4 Methodology

This section introduces the overall architecture and key components of CT-Agent, including how it plans tasks, takes actions, and utilizes memory.

4.1 The framework of CT-Agent

CT-Agent is designed for 3D chest CT question answering and report generation. Built on an LLM backbone, the system dynamically plans tasks, invokes anatomy-aware expert tools, and augments context via memory. It is capable of performing targeted clinical reasoning based on task type and of generating outputs that are semantically consistent and professionally expressed.

The system consists of three core components: a planning module, an action space, and a memory module. The overall architecture of CT-Agent can be formally expressed as

$$\text{CT-Agent} = (\mathcal{P}, \mathcal{T}, \mathcal{M}). \tag{2}$$

Here, $\mathcal{P}$ denotes the planning module responsible for identifying the task (report generation or question answering) and planning the execution path; $\mathcal{T}$ represents the action space, which includes multiple region-specific models and a few-shot selection tool; and $\mathcal{M}$ refers to the agent's memory module.

4.2 Planning module

CT-Agent's planning module serves as the system's central scheduler, managing task recognition, reasoning path planning, and tool invocation. Driven by an LLM, this module dynamically guides the downstream processing flow based on the input query and CT image. Its operational mechanism can be formalized as the following state transition function:

$$S_{t+1} = f(S_t, A_t, E_t). \tag{3}$$

Here, S_t denotes the system state at time t; A_t represents the action taken at that moment; and E_t refers to the external environment at time t, which includes the user-input question Q and corresponding CT volume I. The state transition function f governs the system's evolution by determining which tool modules to activate and whether to invoke the memory module. The action space covers task type classification, anatomical region identification, query rewriting or selection, anatomical model inference, and few-shot selection.

The planner internally generates structured prompts that encode the user intent and available tools to support reasoning and decision-making. These prompts are processed by the LLM, such as DeepSeek-V3, to determine the appropriate downstream execution strategy. Based on the parsed semantics, the planner interacts with the action space by invoking anatomy-aligned LoRA reasoning modules and retrieval components to perform the designated operations. Figure 1 shows that the planning module supports two primary task types: radiology report generation and region-guided question answering.

In report generation, the planning module routes the visual input in parallel to ten region-specific reasoning tools, each responsible for localized inference within its designated anatomical area. A predefined question from a curated query pool is selected for each region and is answered by the corresponding tool to produce a diagnostic statement. Then, these regional outputs are aggregated by the system to generate a complete radiology report. During aggregation, a memory-enhanced mechanism is employed. The planning module encodes the intermediate regional predictions into a semantic embedding and retrieves semantically similar historical exemplars from the memory module. These exemplars are incorporated into the prompt to enhance the fluency and clinical relevance of the final report.

In the region-guided question answering task, the planning module performs query parsing to identify the anatomical focus of the user's question. A question such as "Is there fluid around the heart?" is mapped to the heart region, as defined in the system's ten-region anatomical taxonomy. When the relevant region is identified, the planner invokes the query rewriting tool, which rewrites the original query into a standardized format aligned with the model's training distribution. Afterward, the planner activates only the reasoning module associated with the identified region, avoiding unnecessary computation over irrelevant areas. Then, the activated module performs targeted inference based on the user's intent and extracted visual features of that region. The final answer is generated using a standardized template based on the reasoning output, ensuring fluency and consistency.

4.3 The action space of CT-Agent

CT-Agent's action space serves as the functional repository that enables the agent to conduct anatomical reasoning and context-aware generation. It comprises three key tool categories: anatomy-aware reasoning tools, few-shot retrieval tools, and query normalization tools.

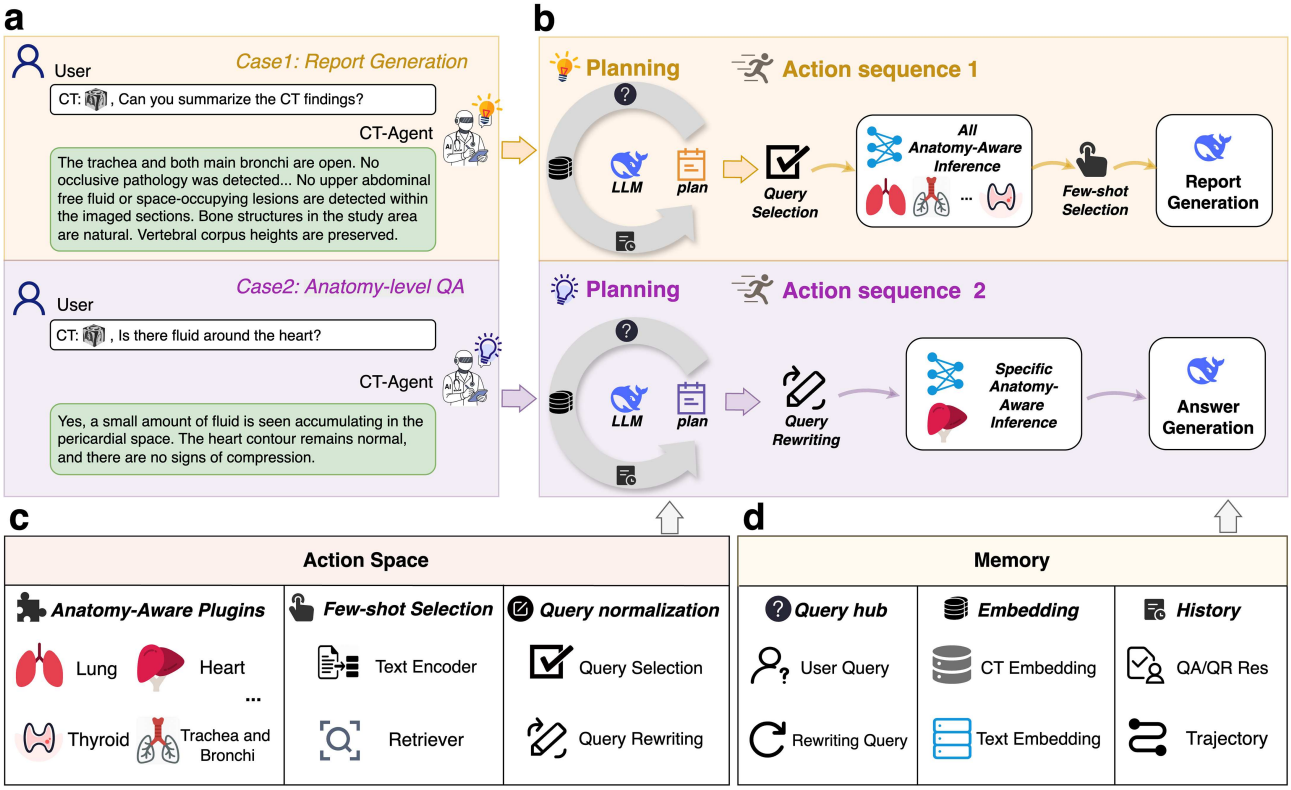


Figure 1 (Color online) The overall architecture of CT-Agent consists of three modules: the planning module, the action space, and the memory module. (a) Example user inputs for report generation and anatomy-level QA tasks; (b) the planning module is driven by an LLM and is responsible for identifying the task, parsing user inputs, locating the involved anatomical region, dispatching appropriate tool modules, and planning how to select few-shot exemplars; (c) the action space includes anatomy-aware plugins specialized for different anatomical regions, as well as few-shot selection and query normalization tools for query selection and rewriting; (d) the memory module stores historical queries, planning paths, and prior radiology reports or question-answering results.

4.3.1 *Anatomy-aware reasoning tools*

The anatomy-aware reasoning tools include LoRA-weighted plugins, each specialized for a distinct anatomical region, such as the lung, trachea and bronchi, mediastinum, heart, esophagus, pleura, bone, thyroid, breast, and abdomen. These models have a common multimodal LLM backbone and are selectively activated based on the task region. To ensure computational efficiency and anatomical precision, each model is supported by a standardized processing pipeline composed of three components: hierarchical token compression, projection into the LLM embedding space, and LoRA-based training. Each component is described in detail below. The overall architecture is illustrated in Figure 2.

Hierarchical token compression. A two-stage token compression consisting of a global token aggregation (GTA) and a local token selection (LTS) pathway was developed to alleviate the contextual burden caused by long token sequences in 3D volumes.

(1) GTA. We first extract slice-level visual tokens using a pretrained CLIP encoder to obtain a compact semantic representation of the CT volume and then apply a token-wise mixture-of-experts (MoE) to the token embeddings of each slice, followed by slice-level averaging to aggregate global features. Let the visual token embeddings extracted from the $T = 240$ axial CT slices be denoted as a sequence of matrices $\{Z_t \in \mathbb{R}^{N\times d}\}_{t=1}^{T}$, where $N = 256$ is the number of visual tokens per slice and $d = 1024$ is the hidden dimension of each token.

Each Z_t represents the visual token matrix of the t-th slice. A shared token-wise MOE is applied to each slice-level token matrix Z_t. It consists of E independent expert MLPs, and a learned gating function assigns each input token to its top-k experts, enabling conditional computation through selective expert activation. Each token vector $z_{t,i} \in \mathbb{R}^d$ (where $i = 1, \ldots, N$) is routed to the top-k experts according to a learned gate:

$$\alpha_{t,i} = \mathrm{Softmax}(W_g z_{t,i} + b_g) \in \mathbb{R}^E, \tag{4}$$

where $W_g \in \mathbb{R}^{E\times d}$ and $b_g \in \mathbb{R}^E$ are learnable gating parameters. The top-k entries of $\alpha_{t,i}$ (with $k \ll E$) identify the

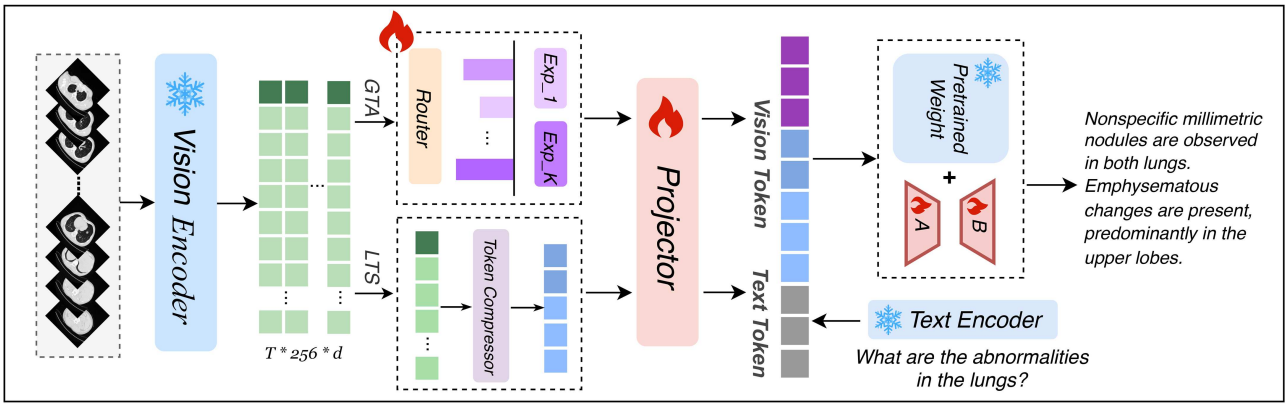


Figure 2 (Color online) The pipeline of anatomy-aware reasoning tools. Given a set of axial CT slices, a frozen vision encoder extracts slice-level visual tokens. The token representations are processed through two parallel pathways: GTA and LTS. The fused tokens are projected through a linear layer and combined with text tokens from the query, forming a multimodal input sequence. The final response is generated by a pretrained language model augmented with LoRA plugins, enabling anatomy-specific reasoning and accurate output. During training, the vision encoder and text encoder remain frozen, while the MoE module, Projector, and LoRA adapters are optimized.

selected experts. Then, the token is processed by the selected experts, and the outputs are aggregated as follows:

$$z'_{t,i} = \sum_{e \in \text{top-}k(\alpha_{t,i})} \alpha_{t,i}^{(e)} \cdot \text{Expert}_e(z_{t,i}). \tag{5}$$

After applying the MOE to all tokens in slice t, we obtain an updated matrix $Z'_t = \{z'_{t,1}, z'_{t,2}, \ldots, z'_{t,N}\} \in \mathbb{R}^{N \times d}$, which preserves the original structure of Z_t, but incorporates expert-guided semantic refinement at the token level. After processing all slices through MoE, the slice-wise mean is computed to aggregate the global representation:

$$Z_{\text{global}} = \frac{1}{T} \sum_{t=1}^{T} Z'_t \in \mathbb{R}^{N \times d}. \tag{6}$$

This produces a final global token matrix $Z_{\text{global}} \in \mathbb{R}^{256 \times 1024}$, which encapsulates the aggregated semantic content across the 3D volume.

(2) LTS. We adopt a VisionZip-inspired [34] two-stage local token compression strategy consisting of dominant token selection and contextual token merging to address the intra-slice redundancy due to patch-based tokenization.

Dominant token selection. First, a subset of highly informative visual tokens is identified based on their attention interactions with the CLS token. Let $A \in \mathbb{R}^{H \times N \times N}$ be the self-attention map from a selected transformer layer in the CLIP encoder, where H is the number of heads and N is the number of visual tokens per slice. For each token j, its attention score is computed as follows:

$$\text{score}_j = \sum_{h=1}^{H} A_{h,\text{CLS},j}, \quad j = 1, \ldots, N. \tag{7}$$

The top-k tokens with the highest attention scores are selected as dominant tokens. Denote these selected token vectors as $\{z_{\text{dom}}^{(1)}, \ldots, z_{\text{dom}}^{(K)}\} \subset \mathbb{R}^d$.

Contextual token merging. The remaining $N-K$ tokens are compressed by merging them based on key similarity to preserve long-tail or background semantics. Let the remaining tokens be split into two sets, a target set $T \in \mathbb{R}^{M \times d}$ and a merge set $M' \in \mathbb{R}^{(N-K-M) \times d}$, where d is the token embedding dimension and M is a predefined hyperparameter specifying the number of output contextual tokens. Similarity is calculated using the dot product between the keys of the merge and target tokens

$$\text{sim}(m_i, t_j) = \langle K_{m_i}, K_{t_j} \rangle. \tag{8}$$

Each merge token m_i is assigned to the most similar target token t_j via the arg max operation. Then, the assigned tokens are aggregated using average pooling

$$\text{merged}_j = \frac{1}{|\mathcal{A}_j|} \sum_{m_i \in \mathcal{A}_j} m_i, \quad \mathcal{A}_j = \left\{ m_i \mid t_j = \arg\max_{t_k \in T} \text{sim}(m_i, t_k) \right\}. \tag{9}$$

The final local token matrix $Z_{\text{local}} \in \mathbb{R}^{(K+M)\times d}$ is obtained by concatenating the selected dominant tokens and generated contextual tokens

$$Z_{\text{local}} = \left[z_{\text{dom}}^{(1)}, \ldots, z_{\text{dom}}^{(K)}, z_{\text{ctx}}^{(1)}, \ldots, z_{\text{ctx}}^{(M)}\right]. \tag{10}$$

This attention-driven compression preserves salient visual cues while significantly reducing token length, retaining only the most informative features for efficient integration with textual queries in the anatomy-aware reasoning.

Projection. To align the compressed visual representation with the LLM's input space, a linear projection is applied to the concatenated global and local tokens. The combined visual token sequence is first constructed as follows:

$$Z_{\text{vision}} = [Z_{\text{global}}; Z_{\text{local}}] \in \mathbb{R}^{L\times d}, \tag{11}$$

where $Z_{\text{global}} \in \mathbb{R}^{G\times d}$ and $Z_{\text{local}} \in \mathbb{R}^{L'\times d}$ represent the global and local token matrices, respectively; $L = G + L'$; and $d = 1024$ denotes the original CLIP token dimension. To match the hidden space of the LLM, each token embedding is projected from $d = 1024$ to $d' = 4096$ using a learnable linear transformation

$$Z_{\text{proj}} = Z_{\text{vision}} W_p + b_p, \quad W_p \in \mathbb{R}^{1024\times 4096},\ b_p \in \mathbb{R}^{4096}. \tag{12}$$

The output $Z_{\text{proj}} \in \mathbb{R}^{L\times 4096}$ serves as the visual input to the multimodal LLM, ensuring integration with text embeddings for anatomy-aware reasoning tasks.

Training and inference. CT-Agent adopts a modular training strategy to support anatomy-specific reasoning while keeping the base multimodal LLM backbone frozen. Each anatomical region r_i is associated with a LoRA plugin ϕ_{r_i}, which is injected into selected layers of the transformer to enable parameter-efficient adaptation. These plugins operate within a shared vision-language model (e.g., LLaVA-Med) and are only activated during region-relevant tasks.

Training phase. Given a CT volume that has been processed by the preprocessing pipeline and a region-level task prompt (a question or region-guided report instruction), the visual input is first transformed into a token matrix $Z_{\text{proj}} \in \mathbb{R}^{L\times d'}$ through hierarchical token compression and projection modules. The task prompt and visual tokens are then concatenated to form the input sequence for the language model:

$$X_{\text{input}} = [T_{\text{task}}, \texttt{<im\_start>}, Z_{\text{proj}}, \texttt{<im\_end>}]. \tag{13}$$

Only the LoRA adapter ϕ_{r_i} corresponding to the annotated anatomical region is activated, while the backbone parameters θ remain frozen. The model is optimized using a standard language modeling loss $\mathcal{L} = -\log P(Y|X_{\text{input}}; \theta, \phi_{r_i})$, where Y is the expected output text, such as a region-specific diagnostic statement or answer.

Inference phase. At inference time, the planner first parses the user query and identifies the target anatomical region. The corresponding LoRA plugin is dynamically loaded into the model. Then, the system encodes the preprocessed CT slides, generates Z_{proj}, and constructs the input sequence X_{input} as in training. The LLM generates the output sequence in an autoregressive manner by predicting the next token conditioned on the previous context. Given the input X_{input}, the model outputs the predicted text $\hat{Y} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_T\}$ as follows:

$$P\left(\hat{Y}|X_{\text{input}}; \theta, \phi_{r^*}\right) = \prod_{t=1}^{T} P\left(\hat{y}_t|X_{\text{input}}, \hat{y}_{<t}; \theta, \phi_{r^*}\right), \tag{14}$$

where $\hat{y}_t$ is the t-th predicted token, and ϕ_{r^*} denotes the LoRA adapter corresponding to the selected anatomical region. This modular approach enables fine-grained specialization across regions while maintaining scalability and low adaptation overhead.

4.3.2 *Few-shot selection tools*

A memory-guided exemplar retrieval mechanism is established to enhance reasoning quality in report generation. This tool retrieves semantically aligned few-shot exemplars by conditioning on the model's own anatomy-level outputs. The overall architecture is presented in Figure 3.

Encoding. Before encoding, the planner selects a predefined question q_i for each anatomical region and invokes the corresponding reasoning tool to generate a diagnostic statement s_i. The resulting s_i are concatenated into a sequence $s_{\text{query}} = [s_1; s_2; \ldots; s_{10}]$. This sequence is encoded using a frozen sentence embedding model to obtain a semantic query vector v_{query}. To support retrieval, a semantic few-shot corpus $\mathcal{C} = \{(v_k, x_k)\}_{k=1}^{K}$ is constructed, initialized from historical reports in the training set. Each x_k is decomposed into ten anatomical findings $\{s_1^{(k)}, \ldots, s_{10}^{(k)}\}$. The

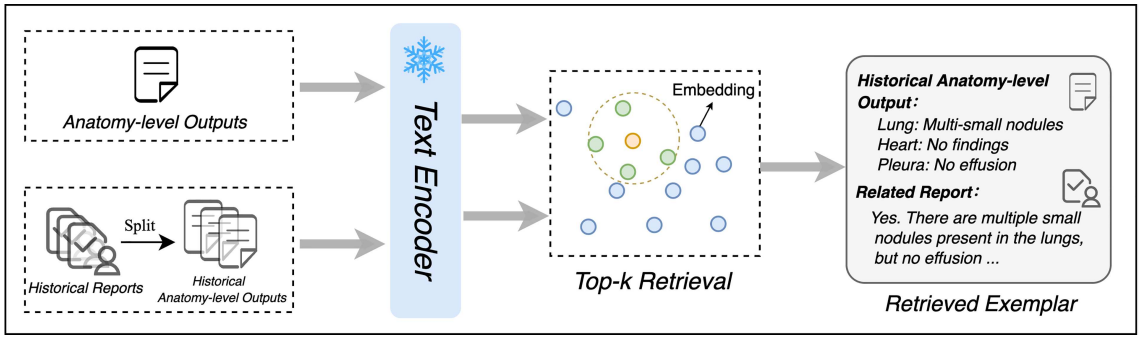


Figure 3 (Color online) Semantic retrieval pipeline for few-shot prompting. The current case's anatomy-level outputs are encoded into a semantic vector and matched against a vector index constructed from historical reports. Top-matching exemplars are retrieved and prepended to the input prompt for final report generation.

sequence $s_k = [s_1^{(k)}; \ldots; s_{10}^{(k)}]$ is then encoded to obtain the semantic key $v_k = \text{Encoder}(s_k)$. All $\{v_k\}$ vectors are stored in a vector index for efficient similarity search.

Retrieval. At inference time, cosine similarity is computed between v_{query} and all stored $\{v_k\}$:

$$\text{sim}(v_{\text{query}}, v_k) = \frac{v_{\text{query}}^\top v_k}{\|v_{\text{query}}\| \cdot \|v_k\|}. \tag{15}$$

The top-K' most similar exemplars $\mathcal{X}_{\text{shot}} = \{x_1, \ldots, x_{K'}\}$ are retrieved and prepended to the input prompt as few-shot demonstrations. This anatomy-conditioned retrieval strategy grounds CT-Agent's generation in clinically relevant cases, enhancing factual accuracy, stylistic fluency, and robustness to anatomical variation.

4.3.3 *Query normalization tools*

The query normalization tools bridge the gap between free-form user questions and the query expected by anatomy-specific reasoning modules. This component includes two distinct mechanisms: query selection and query rewriting.

Query selection. In the report generation task, CT-Agent generates a diagnostic statement for each of the ten anatomies. A predefined query is assigned to each anatomy in advance. These queries address common clinical concerns and remain fixed across all report generation cases.

Query rewriting. In the question-answering setting, users' questions vary widely in style and specificity. CT-Agent infers the user's intent and target anatomy, maps surface expressions to a canonical lexicon, and rewrites the question into one of four structured templates: presence, abnormality, location, and size. Each template specifies how many slots are required, and the system fills abnormality and/or region selectively. When the template semantics depend on a single slot, only that slot is filled; when both the lesion and its location should be specified, both slots are filled. Slot values are determined by the recognized entities and normalized through synonym consolidation (including term unification and abbreviation expansion) to align with the planner's regional taxonomy. When abnormality is missing and the template permits, the system completes it with the common abnormality associated with the given region; when region is underspecified, the planner resolves it to the most appropriate entry in the ten-region set. The rewritten canonical text preserves the original intent while aligning with the training distribution; the final normalized query is passed to the corresponding anatomy-specific reasoning tool to maintain stable prompt formats and predictable system behavior.

4.4 Memory module

Complementing the planning and action space modules, the memory module in CT-Agent serves as a persistent knowledge base to support anatomy-aware reasoning, few-shot selection, and decision traceability. It comprises three main components: the query hub, the embedding store, and the history log.

The query hub stores the user-issued queries and a predefined question pool aligned with anatomical regions. For each region, the system maintains curated canonical questions that guide the reasoning tools during report generation and anatomy-specific QA.

The embedding store manages semantic vector representations that support few-shot selection and vision-language alignment. It maintains two types of embeddings. Region-level CT embeddings, obtained by encoding axial CT slices through the vision encoder and the hierarchical token compression pipeline, followed by MoE and projection; and anatomy-level text embeddings derived from a concatenated sequence using a frozen sentence encoder.

The history log records previously generated QA responses, report outputs, and reasoning trajectories during multistep inference. This log facilitates session-level memory, supports agent interpretability, and provides context-aware grounding for future decisions.

5 Experiments

In this section, we perform comprehensive experiments to validate the effectiveness of our proposed CT-Agent.

5.1 Experimental settings

5.1.1 *Datasets*

The experiments were conducted on two public chest CT datasets. CT-RATE [27] is a large-scale collection of 50188 raw chest CT volumes (47149 for training; 3039 for testing) from 21304 patients, covering a wide range of clinical cases and imaging variations. RadGenome-Chest CT [28] extends CT-RATE by providing region-wise segmentation masks for ten anatomical structures and GPT-generated pathological labels, enabling region-aligned multimodal learning. The dataset includes 25692 non-contrast 3D chest CT scans (24128 for training; 1564 for testing) from 20000 patients, 665k (624876 for training; 40342 for testing) multi-granularity grounded reports, and 1.3 M (1343057 for training; 84625 for testing) grounded visual question-answering pairs. Based on the two datasets, a QA pair dataset was constructed for chest CT images. This dataset comprises 2033648 QA pairs (1914448 for training; 119200 for testing), including question and answer labels for individual organs in each CT image from the CT-RATE and RadGenome-Chest CT datasets. The question templates are categorized into two: presence detection and abnormality identification. Unified templates were developed for each question type, and the anatomical region names were filled using the RadGenome anatomical hierarchy. A presence detection question template may be "Is there any abnormality in the {anatomical region}?" The answer labels are derived from the GPT annotations provided by RadGenome-Chest CT and serve as ground-truth supervision during training. This structured design ensures medical semantic accuracy and large-scale generation of QA pairs to support our subsequent training tasks.

5.1.2 *Robustness considerations*

Robustness analysis leverages the heterogeneity of the public datasets used. CT-RATE aggregates chest CT scans from multiple vendors, such as Philips and Siemens, and covers diverse reconstruction kernels ranging from lung-sharp to mediastinal-smooth, a spread of in-plane matrix sizes, and varied slice thickness and counts, with natural variation in voxel spacing along all three axes. These factors show realistic differences in noise, sharpness, and texture appearance. Building on this foundation, RadGenome-Chest CT adds fine-grained anatomical supervision and large-scale region-aligned text and QA pairs. Consequently, training and evaluation span morphology from large structures (e.g., the lobes, heart, and mediastinum) to small structures (e.g., the segmental bronchi, small vessels, and pleural reflections).

Given this multidimensional heterogeneity, CT-Agent is trained and tested with routine variability in acquisition protocols and hardware, including vendor, reconstruction kernel, resolution, slice thickness, and count. Without additional assumptions, the system shows stable performance for report generation and region-level QA across typical protocol differences.

We also note the limits of external validity. All the study cases are non-contrast chest CTs; enhancement-related signs (e.g., vascular or lesion enhancement patterns) are neither learned nor evaluated. Moreover, the public datasets lack systematic labels for ultra-low-dose scans, pronounced motion or metal artifacts, extremely rare or atypical abnormalities, and cases at the far end of disease severity. Therefore, no robustness claims were made for these out-of-distribution settings. Safe clinical deployment will require additional multicenter, multi-protocol validation and physician-in-the-loop review workflows.

5.1.3 *Baselines*

To validate the effectiveness of the proposed method, it was compared to five representative baselines for 3D medical report generation. (i) CT2Rep [13] is the first approach specifically designed for 3D chest CT report generation. It employs an auto-regressive transformer with a memory-driven decoder and integrates longitudinal multimodal data through a hierarchical memory module. (ii) 3D-CT-GPT [14] builds upon vision-language integration by pairing a CT-specific vision transformer with a large language model. It uses average-pooled 3D features projected into the LLM space, supporting VQA and report generation with strong semantic alignment. (iii) MS-VLM [18] simulates radiologists' slice-by-slice workflow, combining a 2D ViT, Z-former for inter-slice modeling, and a perceiver resampler

to generate compact visual prompts for LLMs, enabling efficient volumetric reasoning without 3D redundancy. (iv) M3D [19] blends local and global 3D features via cross-modal attention in a hierarchical encoder, enhancing lesion details and diagnostic accuracy. Together, these baselines encompass global volumetric encoding, vision-language integration, slice-level processing, and multimodal large-language modeling, providing a comprehensive foundation for evaluating our proposed approach. (v) LLaVA-CT builds on the LLaVA-Med [6] by directly using CT images and medical reports provided by CT-RATE [27] for end-to-end fine-tuning. (vi) CT-Agent (Qwen) follows the same architecture and training procedure as CT-Agent, with the only change being the replacement of the language backbone by Qwen2.5-VL-7B[1)]. The training data, visual encoder, tokenizer, prompt templates, loss functions, and optimization hyperparameters are unchanged. This baseline isolates the effect of the language backbone on report generation and region-level question answering. (vii) CT-Agent (Uni) replaces region LoRA adapters with a single adapter trained on the union of all anatomical regions. At inference time, the same adapter is applied to any region. This baseline tests whether a shared adapter can match or approach the performance of region-specific adapters.

5.1.4 *Evaluation metrics*

To evaluate the effectiveness, two distinct evaluation frameworks are adopted for report generation and region-guided question answering. For the report generation task, first, standard natural language generation (NLG) metrics are employed, including BLEU-1 to BLEU-4 [35], ROUGE-L [36], and METEOR [37]. These metrics evaluate the lexical similarity and fluency of the generated reports with respect to the ground-truth reports. Second, to assess the clinical reliability of the generated content, we incorporate the CE metric proposed in CT2Rep [13], which evaluates the model's performance at the medical semantic level by comparing the consistency of key clinical abnormalities and diagnostic terms between the generated and reference reports. In this study, the metric was implemented using an improved exact match (EM) method, if an abnormal term or key medical keyword that appears in the ground-truth report is accurately mentioned in the generated report, it is considered a successful match. Abnormalities identified in the generated report are compared against those annotated in the ground-truth report, focusing on the 18 common chest CT abnormalities defined in CT2Rep. Each abnormality is treated as an independent category, and micro-averaged precision, recall, and F1 score are computed based on per-abnormality agreement with the ground-truth findings. In the region-guided question answering task, the CE metric is adopted based on the improved EM method to evaluate the medical accuracy of generated answers across 10 anatomical regions, including the lungs, heart, trachea and bronchi, mediastinum, pleura, bones, thyroid, breasts, esophagus, and abdomen.

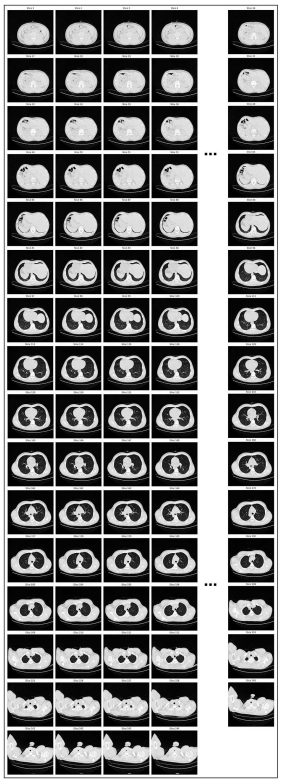
5.1.5 *Implementation details*

3D CT data preprocess. Each 3D CT volume is preprocessed using the MONAI framework[2)] by aligning spatial orientation, cropping the anatomical foreground, and resampling to a fixed voxel spacing of $(1.5, 1.0, 1.0)$ mm. Voxel intensities are converted to Hounsfield Units using DICOM metadata. Each volume is decomposed into 240 axial slices, which are resized and encoded independently using a frozen CLIP ViT-B/16 model. Each slice yields 256 patch tokens of 1024 dimensions, forming a tensor of shape $240 \times 256 \times 1024$ as visual input for the CT-Agent reasoning module.

Tools in action space. Anatomy-aware reasoning tools are built on LLaVA-Med-v1.5[3)], inheriting the core parameters of its language architecture. Each LoRA plugin is injected into all attention and linear layers of the transformer backbone, with rank $r = 16$, scaling factor $\alpha = 16$, and dropout rate 0.05. These adapters are implemented and managed using the PEFT library. In token compression, each CT slice is compressed into 64 visual tokens, including 54 dominant tokens and 10 contextual tokens. During training, the CLIP visual is frozen. The MoE modules, projection layers, and LoRA parameters are optimized using AdamW ($\text{lr} = 2 \times 10^{-4}$, batch size 8) with a 2-epoch schedule and 500-step warm-up. Training is conducted on $8 \times$ A40 GPUs using FSDP with FP16 and gradient accumulation. For the few-shot selection tool, OpenAI's text-embedding-3-small[4)] is used to retrieve the top-3 semantically similar cases from a corpus of 20000 pre-embedded clinical examples.

CT-Agent planning. DeepSeek-V3 is used as the backbone of the CT-Agent planning module. Task classification, anatomy identification, and other planning operations are implemented via structured prompts. All prompt templates are provided in Appendix B.

1) https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct.
2) https://github.com/Project-MONAI/MONAI.
3) https://huggingface.co/microsoft/llava-med-v1.5-mistral-7b.
4) https://platform.openai.com/docs/models/text-embedding-3-small.



Ground Truth	LLAVA-CT (Baseline)
Trachea and both main bronchi are open. No occlusive pathology was detected in the trachea and both main bronchi. There are several millimetric nonspecific nodules in both lungs. There are sometimes linear atelectasis in both lungs. No mass or infiltrative lesion was detected in both lungs. Mediastinal structures cannot be evaluated optimally because contrast material is not given. As far as can be observed: Heart contour and size are normal. There is minimal pericardial effusion. No pleural effusion was detected. The widths of the mediastinal main vascular structures are normal. There are atheromatous plaques in the aorta and coronary arteries. There are no pathologically enlarged lymph nodes in the mediastinum and hilar regions. No pathological increase in wall thickness was detected in the esophagus within the sections. No upper abdominal free fluid-collection or pathologically enlarged lymph nodes were observed in the sections. No fractures or lytic-destructive lesions were detected in the bone structures within the sections.	Trachea and both main bronchi are open. No occlusive pathology was detected in the trachea and lumen of both mediastinum. The heart size has increased (cardiomegaly). Mediastinal vascular structures could not be evaluated optimally due to the lack of contrast, and they have an appearance compatible with hepatosteatosis. Atherosclerotic wall calcifications were observed on the thoracic-abdominal sections in millimeter sizes that can be distinguished by this examination. Calibration of other abdominal organs is normal as far as can be seen; Heart contour and size are natural. Pericardial thickening-effusion was not detected. Thoracic esophagus calibration was normal and no significant tumoral wall thickening was detected. In the mediobasal segment of the lower lobe of the right lung, there is focal consolidation accompanied by ground glass density increases around it. Findings consistent with atypical pneumonia suggest Covid infection. Upper abdominal organs included in the sections are normal. No space-occupying lesion was detected in the liver that entered the cross-sectional area. Bilateral adrenal glands were normal and no space-occupying lesion was detected. Mild degenerative changes are present in bone structures. Hypertrophic osteophytic taperings showing convergence tendencies secondary to previous TB fractures are evident on bilaterals. It is recommended to evaluate metastic sutures together with 6 mm diameter nonspecific nodules at leveling materials secondary to previous examinations?). Clinical evaluation for hypodense appearances cannot be excluded because of multiple myeloma) within the image will explain this later if necessary. Irregular interlobular septums are stable. Lymphedema was observed caudotly at the anterior-posterior corners of the left breast.
Ground Truth	**CT-Agent (Ours)**
Trachea and both main bronchi are open. No occlusive pathology was detected in the trachea and both main bronchi. There are several millimetric nonspecific nodules in both lungs. There are sometimes linear atelectasis in both lungs. No mass or infiltrative lesion was detected in both lungs. Mediastinal structures cannot be evaluated optimally because contrast material is not given. As far as can be observed: Heart contour and size are normal. There is minimal pericardial effusion. No pleural effusion was detected. The widths of the mediastinal main vascular structures are normal. There are atheromatous plaques in the aorta and coronary arteries. There are no pathologically enlarged lymph nodes in the mediastinum and hilar regions. No pathological increase in wall thickness was detected in the esophagus within the sections. No upper abdominal free fluid-collection or pathologically enlarged lymph nodes were observed in the sections. No fractures or lytic-destructive lesions were detected in the bone structures within the sections.	The trachea and both main bronchi are open. No occlusive pathology was detected. The thyroid gland is enlarged and demonstrates a heterogeneous appearance. Nonspecific millimetric nodules are observed in both lungs. Emphysematous changes are present, predominantly in the upper lobes. No mass lesion with distinguishable borders or active infiltrative process is detected. The heart contour and size are normal. A pericardial effusion is present, measuring wider than expected. Mediastinal main vascular structures are normal in caliber. No pathologically enlarged lymph nodes are detected. Pleural effusion or thickening was not observed in either hemithorax. The thoracic esophagus demonstrates normal calibration without significant wall thickening. No upper abdominal free fluid or space-occupying lesions are detected within the imaged sections. Bone structures in the study area are natural. Vertebral corpus heights are preserved.

Figure 4 (Color online) Sentence-level comparison among the CT-Agent generated report, baseline generated report, and reference report (ground_truth). Green highlights indicate consistent findings between the two reports. Blue highlights represent mismatches or deviations. When applied to the ground truth, they mark crucial details that the model failed to include. When appearing in the generated report, they highlight content that does not exist in the reference report, which may indicate redundancy or hallucination. Red highlights indicate statements in the generated report that are factually incorrect or contradictory to the ground truth. Yellow background indicates internally contradictory content within the same generated report.

5.2 Overall performance

Report generation quality. The experimental results of report generation quality are detailed in Table 1. Our method achieves state-of-the-art performance on CE metrics and all NLG metrics, except BLEU-4. While BLEU-4 emphasizes exact 4-gram overlap, our agent favors clinically faithful, longer-form phrasing and valid paraphrases encouraged by anatomy-aware planning and prediction-guided exemplar retrieval. This decreases 4-gram matches against a single reference but improves CE metrics, which better reflect clinical utility. Figure 4 presents cases where semantic equivalence is preserved despite surface variation. Compared to the strongest baseline, our approach yields improvements of 0.06 in BLEU-1, 0.029 in BLEU-2, 0.004 in BLEU-3, 0.022 in ROUGE-L, and 0.004 in METEOR, indicating enhanced fluency, coherence, and semantic alignment. Regarding clinical consistency, our method demonstrates substantial gains in CE metrics, with absolute improvements of 0.016, 0.148, and 0.159 in precision, recall, and F1 score, respectively. These results indicate the model's superior ability to accurately capture critical medical findings. Figure 4 provides a comparison example between our generated report, the baseline, and the ground-truth. The performance gains can be attributed to our anatomy-aware architecture, which integrates region-specific LoRA sub-models, hierarchical token compression to preserve semantic integrity, intelligent task routing for focused reasoning, and prediction-guided exemplar retrieval to enhance contextual relevance. Moreover, our additional baselines show that replacing the language backbone with Qwen2.5-VL-7B leads to weaker performance on NLG and CE metrics. This gap is linked to the fact that the final CT-Agent backbone has been further adapted on medical corpora, providing domain priors that Qwen2.5-VL-7B lacks under identical vision and training settings. A single unified LoRA adapter across all regions remains competitive for general fluency but lags on clinically grounded metrics because it cannot specialize to organ-level patterns and attributes. These observations emphasize the importance of a clinically tuned language backbone and anatomy-aware parameterization.

Region-guided question answering quality. To further evaluate our system's QA capability, its performance is measured across anatomical regions for two tasks: presence detection and abnormality identification. Table 2 shows the P, R, and F1 scores for each anatomical region and task type. Analysis of these metrics reveals that the

Table 1 Performance comparison of our method with baseline models on radiology report generation. We report standard NLG metrics: BLEU-1 to BLUE-4 (BL-1 to BL-4), ROUGE-L (RL), and METEOR (M) and CE metrics: precision (P), recall (R), and F1 score (F1) on the generated reports. Bold values indicate the best performance for that metric.

Method	BL-1	BL-2	BL-3	BL-4	RL	M	P	R	F1
3D-CT-GPT	–	–	–	0.133	0.145	0.140	–	–	–
CT2Rep	0.442	0.344	0.279	0.235	0.401	0.309	0.355	0.132	0.175
MS-VLM	–	–	–	0.232	0.438	0.396	0.222	0.329	0.261
M3D	0.435	0.345	0.286	0.245	0.400	0.326	0.407	0.009	0.148
LLaVA-CT	0.369	0.309	0.275	**0.252**	0.468	0.421	0.323	0.182	0.221
CT-Agent (Qwen)	0.403	0.252	0.168	0.132	0.353	0.327	0.308	0.218	0.233
CT-Agent (Uni)	0.496	0.338	0.278	0.225	0.461	0.389	0.418	0.336	0.351
CT-Agent	**0.502**	**0.374**	**0.290**	0.231	**0.490**	**0.425**	**0.423**	**0.477**	**0.420**

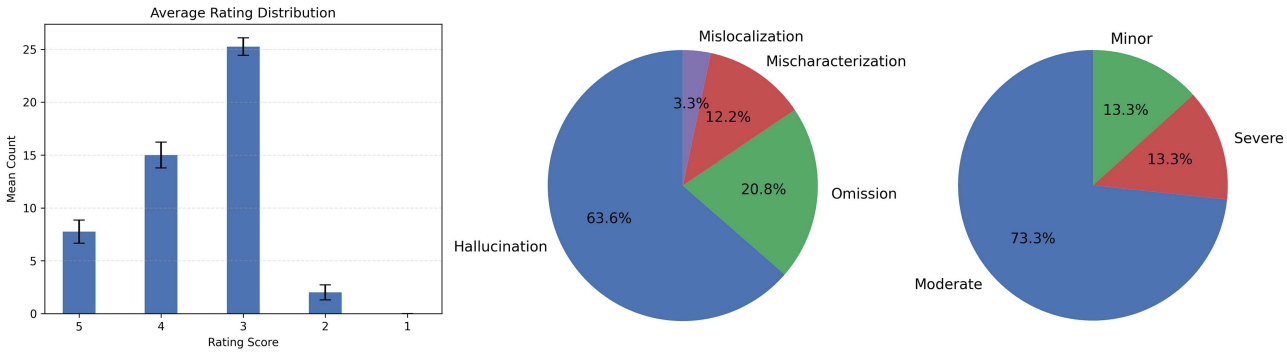
Table 2 Performance comparison between CT-Agent (denoted as OM) and the end-to-end baseline LLaVA-CT (denoted as BL) on visual question-answering tasks across 10 anatomical regions.

Anatomical part	Presence						Abnormality					
	P		R		F1		P		R		F1	
	BL	OM	BL	OM	BL	OM	BL	OM	BL	OM	BL	OM
Lung	0.973	0.962	0.904	0.941	0.937	0.952	0.934	0.917	0.362	0.410	0.521	0.566
Trachea & bronchi	0.188	0.145	0.043	0.349	0.070	0.205	0.055	0.887	0.011	0.713	0.018	0.790
Mediastinum	0.720	0.931	0.663	0.418	0.690	0.577	0.633	0.622	0.446	0.969	0.523	0.758
Heart	0.848	0.868	0.857	0.894	0.853	0.880	0.827	0.842	0.733	0.728	0.777	0.781
Esophagus	0.279	0.219	0.241	0.247	0.259	0.232	0.276	0.217	0.165	0.245	0.207	0.230
Pleura	0.676	0.713	0.232	0.246	0.345	0.366	0.598	0.623	0.182	0.165	0.279	0.260
Bone	0.721	0.763	0.271	0.723	0.393	0.743	0.589	0.559	0.150	0.285	0.239	0.377
Thyroid	0.981	0.962	0.830	0.934	0.899	0.948	0.953	0.906	0.377	0.302	0.540	0.453
Breast	0.964	0.960	0.823	0.812	0.888	0.880	0.905	0.934	0.325	0.487	0.478	0.640
Abdomen	0.933	0.812	0.396	0.584	0.556	0.679	0.882	0.716	0.214	0.341	0.345	0.462
Overall	0.728	0.734	0.526	0.615	0.589	0.646	0.665	0.722	0.297	0.465	0.393	0.532

proposed CT-Agent consistently achieves superior performance compared to the end-to-end baseline, LLaVA-CT, across both evaluated tasks, with the greatest gains observed in abnormality identification. The performance of CT-Agent varies with anatomical complexity. It demonstrates pronounced advantages in recall for high-complexity regions such as the mediastinum and trachea and bronchi, better capturing sparse abnormal signals while maintaining robust performance in standardized regions including the lung and heart, reducing missed diagnoses. The improvement is more substantial for the abnormality task, where CT-Agent overcomes LLaVA-CT's tendency for failure in complex abnormal patterns. These results indicate the effectiveness of our method in accurately identifying anatomical structures and detecting abnormalities, improving its clinical utility.

Clinical reliability analysis of CT-Agent-generated reports. To assess the clinical reliability of CT-Agent, a structured evaluation was conducted with four board-certified thoracic radiologists. Each radiologist independently reviewed the same set of 50 CT-Agent-generated chest CT reports, yielding 200 quality assessments. Using a five-point rating scale, the radiologists scored the overall quality and clinical usefulness of each report. As shown in Figure 5(a), the majority of ratings were 3 or above, indicating that CT-Agent's outputs are generally acceptable as auxiliary clinical references, with a substantial portion approaching the readability and completeness of human-written radiology reports. The scoring was performed through a dedicated web interface (Appendix C, Figure C1).

Furthermore, beyond global quality scoring, the radiologists performed fine-grained sentence-level annotations to characterize specific error patterns. Figure 5(b) shows that each identified error was assigned to one of four categories (hallucination, omission, mislocalization, or mischaracterization) and further labeled with a clinical severity level (minor, moderate, or severe) based on its potential diagnostic impact (Figure 5(c)). Most errors were classified as minor or moderate, with only a small fraction severe enough to cause misdiagnosis or missed critical abnormalities. This distribution indicates that, under appropriate human oversight, CT-Agent-generated reports demonstrate a generally acceptable degree of clinical safety. Moreover, a region-specific analysis reveals that CT-Agent is prone to hallucinations and omissions in descriptions of the thyroid and lungs, manifesting as the introduction of findings unsupported by imaging evidence or the failure to report abnormalities that should have been mentioned. These



(a) Distribution of the mean rating scores averaged across four radiologists

(b) Proportion of error types identified in CT-Agent-generated CT reports

(c) Distribution of clinical severity levels among the identified errors

Figure 5 (Color online) Clinical reliability evaluation of CT-Agent. (a) The five-level rating scale reflects radiologists' overall assessment of report quality. 1, very poor and unusable reports; 2, poor outputs that provide only minimal reference value; 3, reports that are acceptable as auxiliary clinical references; 4, good-quality reports approaching certain human-written ones; and 5, outputs that closely match the readability and completeness of expert-written radiology reports. (b) Distribution of four error categories identified in CT-Agent-generated reports, hallucination denotes descriptions unsupported by imaging evidence, omission indicates missing clinically relevant findings, mislocalization refers to assigning a correct finding to an incorrect anatomical region, and mischaracterization corresponds to an inaccurate description of an existing abnormality. (c) Clinical severity levels of annotated errors, minor errors have negligible clinical impact; moderate errors may influence diagnostic decisions, but typically do not cause major harm; and severe errors are those that could lead to misdiagnosis or missed critical abnormalities.

observations underscore that although CT-Agent performs reasonably well overall, improving the model's ability to accurately recognize and describe abnormalities in critical anatomical regions such as the thyroid and lungs is critical for future development.

5.3 Ablation study

5.3.1 *Effectiveness of CT-Agent planning*

To evaluate the effect of anatomical planning, CT-Agent is compared with a standard end-to-end baseline LLaVA-CT that lacks region-specific routing. As shown in Tables 1 and 2, CT-Agent outperforms the baseline by 0.199 CE-F1 in report generation and achieves a 0.057 gain in average F1 for presence and 0.139 gain for abnormality QA.

Furthermore, error analysis revealed that the general model often suffers from anatomical confusion (e.g., mistaking mediastinal abnormalities as lung-related or confusing the heart with the pleura) due to the lack of region-guided control. In contrast, the CT-Agent architecture guides the model's attention to the correct region via a structured planning mechanism, reducing such structural-level errors. These results demonstrate the advantage and interpretability of agent-based planning in 3D medical QA.

5.3.2 *Effectiveness of hierarchical token compression*

Ablation studies were conducted to evaluate the efficacy of the token compression module in the CT-Agent framework by systematically replacing its components with four alternative token processing strategies. First, the original module was substituted with a simple truncation strategy that discards excess tokens. Second, slice-level random sampling was conducted to reduce sequence length. Third, fixed slice sampling was tested by selecting slices at predetermined intervals. Fourth, our proposed compression module was deployed without global tokens, and the module was finally integrated with global tokens. Table 3 shows that truncation significantly degraded performance in lung-related tasks, with precision and recall in lung presence detection decreasing to 0.857 and 0.319, respectively, due to the loss of slice information. Random sampling caused instability, yielding only 0.505 F1 in lung presence detection, while fixed-interval sampling provided better consistency, but was limited to anatomical variations. CT-Agent without global tokens slightly decreased performance, from 0.566 to 0.559 in lung abnormality F1. The full CT-Agent achieved the best results, with 0.941 recall and 0.952 F1 in lung presence detection, showing that adaptive compression and global tokens enhance local feature retention and cross-slice reasoning. The performance gap emphasizes our method's value for 3D CT VQA. Similar trends were observed in heart-related metrics, further confirming the effectiveness of our design.

Moreover, the local token budget and top-k retention ratio in LTS are ablated in Table 4. Increasing the total retained tokens from 32 to 64 yields clear gains on both lung and heart tasks. For a fixed budget, a higher retention ratio (85% vs. 50%) consistently improves recall and F1, indicating that overly aggressive pruning discards

Table 3 Comparison of token compression strategies in CT-Agent, including truncation, random/fixed sampling, and variants with or without global tokens.

Method	Lung						Heart					
	Presence			Abnormality			Presence			Abnormality		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
No compression (trunc.)	0.857	0.319	0.465	0.835	0.141	0.242	0.810	0.354	0.493	0.786	0.289	0.418
Random sampling	0.894	0.352	0.505	0.882	0.251	0.391	0.850	0.556	0.672	0.821	0.460	0.590
Fixed-interval sampling	0.862	0.375	0.523	0.818	0.282	0.419	0.841	0.713	0.772	0.802	0.650	0.718
CT-Agent (w/o Global)	**0.970**	0.841	0.901	**0.932**	0.399	0.559	**0.876**	0.835	0.855	0.834	0.726	0.776
CT-Agent (w/ Global)	0.962	**0.941**	**0.952**	0.917	**0.410**	**0.566**	0.868	**0.894**	**0.880**	**0.842**	**0.728**	**0.781**

Table 4 Ablation of LTS hyperparameters—effect of the local token budget ($K + M$) and the top-k retention ratio on lung/heart presence and abnormality metrics.

Setting	Lung						Heart					
	Presence			Abnormality			Presence			Abnormality		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
$K + M = 32$, top-$k = 50\%$	0.450	0.128	0.199	0.311	0.092	0.142	0.406	0.179	0.248	0.382	0.112	0.173
$K + M = 32$, top-$k = 85\%$	0.479	0.154	0.233	0.385	0.126	0.190	0.463	0.272	0.343	0.411	0.235	0.299
$K + M = 48$, top-$k = 50\%$	0.616	0.437	0.511	0.598	0.271	0.376	0.605	0.391	0.475	0.578	0.355	0.442
$K + M = 48$, top-$k = 85\%$	0.755	0.519	0.615	0.702	0.288	0.408	0.719	0.503	0.592	0.699	0.484	0.572
$K + M = 64$, top-$k = 50\%$	0.891	0.822	0.855	0.844	0.342	0.487	0.832	0.774	0.802	0.795	0.703	0.746
$K + M = 64$, top-$k = 85\%$	0.962	**0.941**	**0.952**	**0.917**	**0.410**	**0.566**	0.868	**0.894**	**0.880**	**0.842**	**0.728**	**0.781**
$K + M = 128$, top-$k = 50\%$	0.924	0.876	0.899	0.893	0.367	0.520	0.809	0.814	0.811	0.769	0.690	0.727
$K + M = 128$, top-$k = 85\%$	**0.965**	0.905	0.933	0.912	0.397	0.553	**0.869**	0.841	0.854	0.828	0.721	0.771

Table 5 Comparison of different exemplar retrieval strategies on radiology report generation. We report standard NLG metrics: BLEU-1 to BLUE-4 (BL-1 to BL-4), ROUGE-L (RL), and METEOR (M). CT-Agent (w/ Zero) uses no exemplars (zero-shot), CT-Agent (w/ Stat) uses statically selected exemplars, and CT-Agent (w/ Retr) uses prediction-guided exemplar retrieval.

Method	BL-1	BL-2	BL-3	BL-4	RL	M
CT-Agent (w/ Zero)	0.423	0.274	0.188	0.136	0.346	0.335
CT-Agent (w/ Stat)	0.484	0.349	0.265	0.210	0.440	0.399
CT-Agent (w/ Retr)	**0.502**	**0.374**	**0.290**	**0.231**	**0.490**	**0.425**

informative context. Notably, $K + M = 64$ with top-$k = 85\%$ offers the best trade-off, matching the full model's performance while keeping the sequence compact; further enlarging the budget to $K + M = 128$ shows diminishing returns. Overall, a moderate local budget paired with a permissive retention setting preserves lesion cues while maintaining efficiency.

5.3.3 *Effectiveness of prediction-guided exemplar retrieval*

To evaluate the effectiveness of the prediction-guided exemplar retrieval module, it is compared with two alternative strategies: a zero-shot baseline, where no exemplar is provided and the model solely relies on its internal knowledge to generate reports, and a static few-shot baseline, where a fixed set of exemplars is used for all inputs without semantic matching.

Table 5 reveals that replacing prediction-guided retrieval with static few-shot decreases BLEU-4 and METEOR by 0.021 and 0.026, respectively, while zero-shot substitution amplifies these decreases to 0.095 and 0.090. The sharp decreases in BLEU-3/4 indicate degraded fluency and semantic coherence, and the synchronized METEOR decrease reveals reduced lexical diversity and semantic precision.

Notably, static few-shot lacks input-specific adaptability, and zero-shot strategies over-rely on model generalization without contextual guidance. In contrast, prediction-guided retrieval dynamically selects semantically aligned examples via intermediate predictions, resolving these limitations. This mechanism significantly improves report quality, underscoring its necessity for robust, context-aware NLG in CT-Agent.

Moreover, a detailed qualitative analysis revealed that static, few-shot prompting often leads to the use of irrelevant or suboptimal exemplars, negatively affecting report coherence and accuracy. In contrast, our dynamic retrieval method aligns exemplar context with the specific query, enhancing the overall quality and clinical relevance of the generated reports.

6 Discussion

Scope. This study presents CT-Agent. It aims to identify common anatomical regions and their associated abnormalities and to produce structured textual descriptions that enhance clinical entity coverage and anatomical consistency. The scope of this work is confined to image-level findings and question answering; it does not incorporate cross-modal information such as patient history or laboratory results, nor does it aim to provide full diagnostic reasoning. Accordingly, CT-Agent is positioned as an assistive tool for radiological image interpretation rather than a system for definitive clinical decision-making.

Applications. CT-Agent has several potential applications in practical clinical settings. It can automatically draft preliminary reports, helping radiologists reduce reporting time and minimize omissions, thereby improving overall efficiency. Furthermore, its interactive, region-level question answering allows clinicians to quickly verify or retrieve information about specific anatomical areas while reading scans. In addition, the standardized, structured outputs produced by the system can serve as useful educational resources for medical students and junior radiologists, supporting the training of consistent reporting practices.

Scalability. Our focus on non-contrast chest CT is driven by the availability of CT-RATE and RadGenome-Chest CT, which provide large-scale, structured region annotations and QA pairs that enable comprehensive training and evaluation. The approach is not limited to thoracic anatomy, the LoRA-based plugin mechanism is a general PEFT scheme that can be instantiated for any body region given appropriate region-level labels and training data. In principle, if similarly annotated datasets become available for brain CT or other anatomical sites, site-specific LoRA tools can be trained so that the agent acquires dedicated recognition and description capability for those regions. Thus, cross-site extension is not a trivial "drop-in": anatomical differences, broader disease spectra, and heterogeneous acquisition/contrast protocols may require redesign of question templates and infusion of domain knowledge to handle cross-organ relationships.

Limitations. Several limitations should be noted. On the methodological side, the global-local token reduction strategy improves efficiency by preserving salient image features; however, in rare cases, it may reduce sensitivity to subtle lesions or uncommon abnormalities. All outputs require human oversight; safety and reliability depend on transparency, interpretability, and rigorous validation within clinical workflows. On the evaluation side, although our CE metric moves toward clinically grounded assessment, the current design relies on 18 "common abnormalities" and an improved exact-match procedure. This scope does not fully capture the richness of comprehensive radiology reports—including lesion attributes (size, morphology, and, when available, contrast enhancement), relationships to adjacent structures, differential diagnoses, and follow-up recommendations. Consequently, the reported "clinical accuracy" may underrepresent these nuanced but clinically meaningful dimensions, limiting generalizability to more complex reporting scenarios. Typical failure cases are summarized in Appendix A.

Future directions. We will collaborate with hospitals and radiology departments to conduct large-scale blinded evaluations of the system, assessing accuracy, safety (particularly the detection of rare but critical findings and the control of potential hallucinations), and practical utility within routine workflows; this will be a near-term priority. Moreover, we plan to expand supervision by incorporating contrast-enhanced CT phases and adding expert fine-grained labels, complementing automatic metrics with structured human review by board-certified radiologists using guideline-based rubrics, and extending evaluation to differential diagnosis generation and longitudinal follow-up across serial examinations. Finally, we will refine the CE metric toward a structured, lesion-centric scheme that scores attributes, spatial relations, and management recommendations rather than relying on exact string matches, aiming to better align with real clinical reasoning.

7 Conclusion

This study proposed CT-Agent, a multimodal-LLM agent designed for 3D CTQA. CT-Agent addresses two fundamental challenges of 3D CT analysis through an anatomy-aware and a global token compression strategy: anatomical complexity and spatial relationships. CT-Agent enables interpretable, region-specific inference while preserving token efficiency and semantic integrity. Extensive experiments on the CT-RATE and RadGenome-Chest CT datasets demonstrate that CT-Agent outperforms state-of-the-art methods across NLG and CE. Our ablation studies confirm the effectiveness of each core component. Notwithstanding the demonstrated effectiveness of CT-Agent in both NLG and CE on 3D chest CT tasks, several limitations remain. Future studies should integrate multimodal clinical evidence, such as prior radiology reports, to enhance contextual reasoning. Moreover, we plan to incorporate longitudinal scan analysis and real-time physician feedback to support interactive and temporally aware diagnostic workflows.

Acknowledgements This work was supported in part by National Natural Science Foundation of China (Grant Nos. 62025206, 62302436, U23A20296) and Zhejiang Province's "Lingyan" R&D Project (Grant No. 2024C01259).

References

1 McCollough C H, Leng S, Yu L, et al. CT dose index and patient dose: they are not the same thing. Radiology, 2011, 259: 311–316
2 Schober O, Kiessling F, Debus J. Molecular Imaging in Oncology. 2nd ed. Cham: Springer, 2020. 31–110
3 Khlaut J, Ferreres E, Tordjman D, et al. RadSAM: segmenting 3D radiological images with a 2D promptable model. 2025. ArXiv:2504.20837
4 Cohen J P, Blankemeier L, Chaudhari A. Explaining 3D computed tomography classifiers with counterfactuals. 2025. ArXiv:2502.07156
5 Yu X, Yang Q, Liu H, et al. Enhancing single-slice segmentation with 3D-to-2D unpaired scan distillation. 2024. ArXiv:2406.12254
6 Li C, Wong C, Zhang S, et al. LLaVA-Med: training a large language-and-vision assistant for biomedicine in one day. In: Proceedings of Conference on Neural Information Processing Systems, 2023
7 Zhang X, Wu C, Zhao Z, et al. PMC-VQA: visual instruction tuning for medical visual question answering. 2023. ArXiv:2305.10415
8 Wang Z, Liu L, Wang L, et al. R2GenGPT: radiology report generation with frozen LLMs. Meta-Radiol, 2023, 1: 100033
9 Chen Z, Song Y, Chang T H, et al. Generating radiology reports via memory-driven transformer. 2020. ArXiv:2010.16056
10 Liu S, Zhang X, Zhou X, et al. BPI-MVQA: a bi-branch model for medical visual question answering. BMC Med Imag, 2022, 22: 79
11 Hyland S L, Bannur S, Bouzid K, et al. Maira-1: a specialised large multimodal model for radiology report generation. 2023. ArXiv:2311.13668
12 Thawkar O, Shaker A, Mullappilly S S, et al. XrayGPT: chest radiographs summarization using medical vision-language models. 2023. ArXiv:2306.07971
13 Hamamci I E, Er S, Menze B. CT2Rep: automated radiology report generation for 3D medical imaging. In: Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention, 2024. 476–486
14 Chen H, Zhao W, Li Y, et al. 3D-CT-GPT: generating 3D radiology reports through integration of large vision-language models. 2024. ArXiv:2409.19330
15 Li S, Xu B, Luo Y, et al. ViT3D alignment of LLaMA3: 3D medical image report generation. 2024. ArXiv:2410.08588
16 Di Piazza T, Lazarus C, Nempont O, et al. CT-AGRG: automated abnormality-guided report generation from 3D chest CT volumes. 2024. ArXiv:2408.11965
17 Liu C, Wan Z, Wang Y, et al. Benchmarking and boosting radiology report generation for 3D high-resolution medical images. 2024. ArXiv:2406.07146
18 Lee C, Park S, Shin C I, et al. Read like a radiologist: efficient vision-language model for 3D medical imaging interpretation. 2024. ArXiv:2412.13558
19 Bai F, Du Y, Huang T, et al. M3D: advancing 3D medical image analysis with multi-modal large language models. 2024. ArXiv:2404.00578
20 Wang W, Ma Z, Wang Z, et al. A survey of LLM-based agents in medicine: how far are we from Baymax? 2025. ArXiv:2502.11211
21 Ghezloo F, Seyfioglu M S, Soraki R, et al. PathFinder: a multi-modal multi-agent system for medical diagnostic decision-making applied to histopathology. 2025. ArXiv:2502.08916
22 Wang Z, Wu J, Low C H, et al. MedAgent-Pro: towards multi-modal evidence-based medical diagnosis via reasoning agentic workflow. 2025. ArXiv:2503.18968
23 Feng J, Zheng Q, Wu C, et al. M^3Builder: a multi-agent system for automated machine learning in medical imaging. 2025. ArXiv:2502.20301
24 Liu G, He J, Li P, et al. PeFoMed: parameter efficient fine-tuning of multimodal large language models for medical imaging. 2024. ArXiv:2401.02797
25 Yang J, Shi R, Ni B. MedMNIST classification decathlon: a lightweight autoML benchmark for medical image analysis. In: Proceedings of IEEE International Symposium on Biomedical Imaging, 2021. 191–195
26 Kuş Z, Aydin M. MedSegBench: a comprehensive benchmark for medical image segmentation in diverse data modalities. Sci Data, 2024, 11: 1283
27 Hamamci I E, Er S, Almas F, et al. Developing generalist foundation models from a multimodal dataset for 3D computed tomography. 2024. ArXiv:2403.17834
28 Zhang X, Wu C, Zhao Z, et al. RadGenome-Chest CT: a grounded vision-language dataset for chest CT analysis. 2024. ArXiv:2404.16754
29 Yang J, Shi R, Wei D, et al. MedMNIST v2-a large-scale lightweight benchmark for 2D and 3D biomedical image classification. Sci Data, 2023, 10: 41
30 Li J N, Zhou Z W, Yang J C, et al. MedShapeNet–a large-scale dataset of 3D medical shapes for computer vision. Biomed Eng-Biomed Te, 2024, 70: 71–90
31 Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision. 2021. ArXiv:2103.00020
32 Hu E J, Shen Y, Wallis P, et al. LoRA: low-rank adaptation of large language models. In: Proceedings of International Conference on Learning Representations, 2022
33 Mao Y R, Ge Y H, Fan Y J, et al. A survey on LoRA of large language models. Front Comput Sci, 2025, 19: 197605
34 Yang S, Chen Y, Tian Z, et al. VisionZip: longer is better but not necessary in vision language models. 2024. ArXiv:2412.04467
35 Papineni K, Roukos S, Ward T, et al. Bleu: a method for automatic evaluation of machine translation. In: Proceedings of Annual Meeting of the Association for Computational Linguistics, 2002. 311–318
36 Lin C Y. ROUGE: a package for automatic evaluation of summaries. In: Proceedings of Text Summarization Branches Out, 2004. 74–81
37 Banerjee S, Lavie A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 2005. 65–72

Appendix A Qualitative error analysis

Note. The following catalog summarizes error types and occasional hallucinations of LLMs during training and testing, intended to guide manual quality control and prioritize model improvements; they do not imply that these errors occur in the majority of cases.

(1) Trachea & bronchi.

- [Omission]
  - Tracheal deviation. Deviated airway evident on imaging but not mentioned.
  - Bronchiectasis. Actual bronchial dilatation present on CT was overlooked or not reported.
- [Hallucination]
  - Bronchiectasis. Unwarranted bronchial dilatation described despite no such finding in the reference.
  - Tracheal deviation. Falsely indicated shift of the trachea where none existed.
  - Mucus plugs/secretions. Mentioned without visible airway obstruction or filling.

(2) Thyroid.

- [Omission]

– Thyroid nodule. Missed subcentimeter hypodense/heterogeneous focus that warrants mention.
• [Hallucination]
– Thyroid nodules. Spurious thyroid nodules or gland enlargement reported without support in the ground truth.

(3) Lung.

• [Omission]
– Ground-glass opacity. Subtle focal ground-glass opacities present on CT were not mentioned in the report.
– Consolidation. Areas of pulmonary consolidation seen on imaging were omitted.
– Pulmonary nodule. Small pulmonary nodule(s) noted in the ground truth were unreported by the model.
– Atelectasis. Minor subsegmental collapse/linear atelectasis evident on CT was not noted in the report.
– Fibrotic scarring. Sequelae fibrotic changes (old scars) visible in the lungs were overlooked and not mentioned.
– Emphysema. Basilar/apical emphysematous changes present in the reference were not acknowledged in the report.
– Pneumothorax. A small apical pneumothorax present on the scan was not recognized or mentioned by the report.
• [Hallucination]
– Emphysema. Emphysematous changes were described by the model despite no such finding in the ground truth.
– Pulmonary nodules. The model reported extra or non-existent lung nodules (overstating the nodule burden not in reference).
– Consolidation. Patchy infiltrates/consolidation were noted by the model with no imaging evidence in the reference.
– Atelectasis. The model falsely identified subsegmental atelectasis that was not actually present.
– Fibrosis. Fibrotic changes were incorrectly inferred by the model where the reference showed none.
– Pulmonary mass. A suspicious lung mass was noted in the report despite no corresponding mass in the ground truth.

(4) Heart.

• [Omission]
– Cardiomegaly. An enlarged cardiac silhouette seen in the reference was not acknowledged in the model's report.
– Pericardial effusion. A small pericardial fluid collection present on CT was left unreported by the model.
• [Hallucination]
– Pericardial effusion. The model falsely reported a pericardial effusion that was not actually present in the ground truth.
– Coronary artery plaque. Calcified coronary atherosclerosis was described by the model despite no mention of it in the reference report.
– Valve calcification. The report noted cardiac valve calcifications that were not reported in the ground truth.

(5) Mediastinum.

• [Omission]
– Mediastinal lymphadenopathy. Pathologically enlarged mediastinal lymph nodes documented in the ground truth were not described by the report.
– Aortic atherosclerotic plaque. Tiny mural plaques in the arch/ascending aorta not mentioned or quantified.
• [Hallucination]
– Mediastinal lymph nodes. The model mentioned mediastinal lymph nodes (e.g., calcified or "millimetric" nodes) that were not noted in the reference report.
– Aortic atherosclerosis. Calcific plaques or dilation of the thoracic aorta were reported by the model without any mention in the ground truth.

(6) Pleura.

• [Omission]
– Pleural effusion. Trace subpulmonic or posterior costophrenic collections not reported.
– Pleural thickening. Subtle focal thickening or pleural reaction overlooked.
• [Hallucination]
– Pleural thickening. The model described pleural thickening/calcifications that were not actually noted in the ground truth.
– Pleural effusion. A pleural fluid collection was reported by the model despite no effusion being present in the reference.

(7) Esophagus.

• [Omission]
– Hiatal hernia. A sliding hiatal hernia seen on CT was not acknowledged.
• [Hallucination]
– Hiatal hernia. The reported hiatal hernia was not substantiated by the ground truth findings.

(8) Abdomen.

• [Omission]
– Hepatic steatosis. Diffuse fatty liver changes (low liver density) noted by the radiologist were not reported.
– Gallstones. Gallbladder calculi visible on imaging were present in the ground truth but went unmentioned in the report.
– Renal cysts. Incidental simple kidney cysts or anatomical variants observed on CT were not reported.
• [Hallucination]
– Gallstones. The model erroneously reported gallbladder stones that were not noted in the reference report.

(9) Bones.

• [Omission]
– Vertebral metastasis. An aggressive bone lesion (metastatic destruction in a vertebra or rib) seen on CT was not included in the model's findings.
– Compression fracture. A vertebral compression deformity present in the ground truth was not mentioned by the model.
• [Hallucination]
– Old fracture. The model incorrectly identified a healed or old fracture (e.g., vertebral or rib) that the ground truth did not report.
– Hemangioma. A benign vertebral hemangioma was noted in the generated report despite no such lesion in the reference.

– Osteopenia. Diffuse low bone density was asserted by the model with no basis in the ground truth (no mention of osteopenia).

– Degenerative changes. The model added exaggerated degenerative changes (osteophytes, etc.) beyond what was noted in the actual report.

(10) Breast.

• [Omission]

– Post-surgical collection. A loculated fluid collection at a post-mastectomy site (following breast cancer surgery) seen on CT was not described by the model.

– Breast prosthesis. The presence of bilateral breast implants (prosthetic material visible on imaging) was not acknowledged in the model report.

– Breast lesion. An abnormal breast density (e.g., a left breast mass vs. a lymph node) apparent on the scan was not reported by the model.

– Chest wall mass. A soft-tissue mass in the chest wall or paravertebral muscle (e.g., metastatic lesion) noted in the ground truth was omitted in the model output.

• [Hallucination]

– Breast calcifications. The model noted the presence of calcifications within the breast tissue, which were not described in the reference report.

– Axillary adenopathy. The generated report described prominent or enlarged axillary lymph nodes that were not noted in the reference.

Appendix B Prompt template

```
Task classification template:
You are a medical-domain assistant who classifies the user's intent and extracts anatomical focus from the question if applicable.

Given a natural language query related to a chest CT scan, your task is to:
- Determine whether the query is a radiology report generation request or a region-guided question (QA task).
- If it is a QA task, identify the anatomical region(s) mentioned or implied in the question.
   Choose from the following predefined regions:
   ["Trachea and Bronchi", "Thyroid", "Lung", "Heart", "Mediastinum", "Pleura", "Esophagus", "Abdomen", "Bone", "Breast"]
- If it is a report generation task, leave the 'target_region' as an empty list.

Return your results in the following JSON format:
'''json
{
  "task_type": "QA" or "Report",
  "target_region": [list of anatomical regions or empty list]
}

User query: {{user_question}}
```

```
Query rewriting template:

You are a clinical assistant helping standardize diagnostic questions for chest CT interpretation. Given:
- A user question written in free-form natural language.
- A predefined set of clinical question templates.
- The anatomical region identified by the system.

Your task:
- Identify the clinical intent of the user question (e.g., presence detection, abnormality localization, or size estimation).
- Choose the most appropriate predefined template.
- Fill in the placeholders (e.g., {region}, {abnormality}) using information inferred from the user's question.
- If the abnormality type is not explicitly mentioned, use a general placeholder like "abnormality".

Predefined question templates:
1. What are the abnormalities in the {region}?
2. What is the approximate size of the {abnormality} in the {region}?
3. Where is the {abnormality} located in the image?
4. Can {abnormality} be identified in the {region}?

Input:
User question: {{user_question}}
Target anatomical region: {{region}}

Output format:
Rewritten clinical query: {{generated_question}}
```

```
Answer generation template:

You are a medical-domain assistant. Your task is to answer the user's original question by referencing the provided professional
     version of the question and its corresponding answer. Assume the reference question is a clinically accurate reformulation of the
      user's intent. Use its answer as the basis for your response. If the reference answer does not fully cover the user's question,
     you may cautiously infer missing details based on established medical knowledge, and explain your reasoning briefly.

User question: {{user_question}}

Reference question: {{reference_question}}

Reference answer: {{reference_answer}}

Output format: answer: {{generated answer}}
```

```
Report generation template:

You are a board-certified radiologist. Given structured findings for each anatomical region of a chest CT scan, generate a clinically
     styled radiology report that matches expert-written radiology report style and language patterns.

Follow these detailed instructions:

Anatomical regions:
Report in the following strict order:
Trachea and Bronchi > Thyroid > Lung > Heart > Mediastinum > Pleura > Esophagus > Abdomen > Bone > Breast

Language style and formatting:
- Use passive voice, objective tone, and formal clinical phrasing.
- For each region, begin with a normal finding sentence if applicable.
- For abnormal findings, follow the 4-part structure:
  [Anatomical location] + [Image finding] + [Measurement if applicable] + [Interpretation]


Standard sentence templates:

Normal description template (use exactly when applicable):
- "Trachea and both main bronchi are open".
- "No occlusive pathology was detected in the trachea and both main bronchi".
- "Heart contour and size are normal".
- "Pericardial effusion-thickening was not observed".
- "Thoracic esophagus calibration was normal and no significant wall thickening was detected".
- "Mediastinal main vascular structures, heart contour, size are normal".
- "No enlarged lymph nodes in pathological dimensions were detected".
- "Pleural effusion-thickening was not detected".
- "No space-occupying lesion was detected in the liver that entered the cross-sectional area".
- "Bone structures in the study area are natural".
- "Vertebral corpus heights are preserved".

Abnormal description patterns (apply when findings exist):
Use phrases like:
- "Ground-glass opacities are observed in the [lung region], especially in the [peripheral areas]".
- "Millimetric nodules are observed in both lungs, the largest measuring [X] mm in [location]".
- "Subsegmental atelectasis areas are noted in the [segment]".
- "There is a pleural effusion with loculation measuring [X] cm at its thickest point".
- "A [X] mm diameter calculus was observed in the gallbladder lumen".
- "Diffuse degenerative changes and osteophytic taperings are noted in the thoracic vertebrae".

Interpretation language:
- "Findings are evaluated in favor of [diagnosis]".
- "Findings appear stable".
- "It is recommended to be evaluated together with clinical and laboratory results".
- "No mass lesion with distinguishable borders was detected".
- "As far as can be observed$\ldots$"

Final output:
- Combine all findings into a fluent, multi-paragraph report, organized in the specified order.
- Avoid subjective judgments or treatment suggestions.
- Do not summarize or conclude; only report imaging findings.
- Use radiological English close to examples above to improve quality and consistency.
- Try to use longer and more coherent sentences as much as possible.

Input structured finding: {{inputs}}

Examples: {{examples}}
```

Appendix C User interface for clinical evaluation of CT-Agent-generated reports

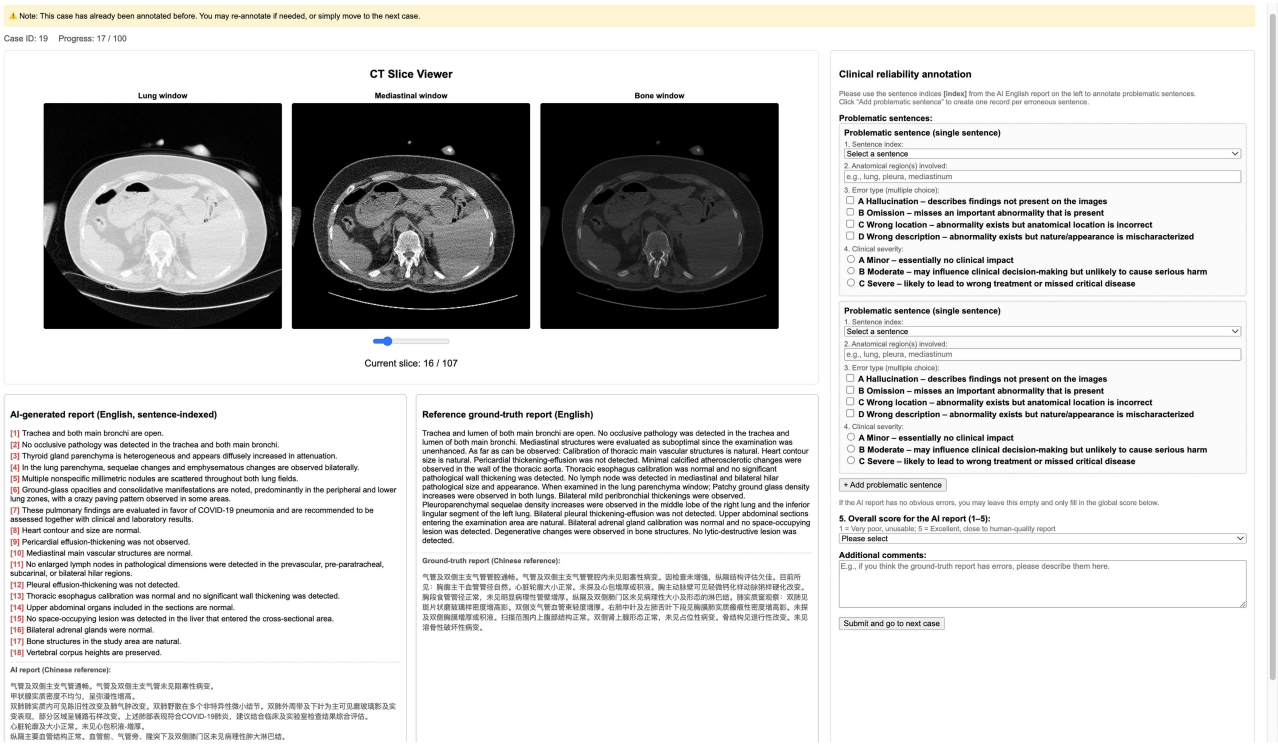


Figure C1 (Color online) Screenshot of the web-based interface used by radiologists to perform clinical reliability evaluation of CT-Agent-generated CT reports.