TOWARDS MULTI-CT PERCEPTION AND UNDERSTANDING WITH A RADIOLOGY TOOL-AUGMENTED MULTIMODAL LARGE LANGUAGE MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Clinical CT interpretation inherently requires multi-volume reasoning, as radiologists routinely synthesize complementary evidence across anatomical regions, contrast phases, and longitudinal visits for accurate diagnostic assessment. However, existing 3D medical multimodal large language models predominantly operate on a single CT volume and fail to integrate complementary evidence across scans and longitudinal visits that radiologists routinely consider in clinical decision-making. Meanwhile, large-scale benchmarks for evaluating such multi-volume CT reasoning remain limited. To address these limitations, we formulate a challenging yet important setting, termed multi-CT perception and understanding, in which a model jointly reasons over multiple CT volumes. First, we construct a large-scale benchmark, namely CliMCT-Bench from 87,953 clinical visits involving 298,377 CT volumes, resulting in 267K visual question-answer pairs spanning six clinical capabilities from basic perception to longitudinal reasoning. We further propose M^2-FM, a multimodal large language model that integrates evidence across scans and interleaves reasoning with the adaptive invocation of radiology tools. Beyond standard contrastive pretraining and supervised fine-tuning, we introduce two innovative training strategies. Specifically, we propose multi-turn and multi-scale trajectories to enable the model to perceive both global context and salient regions, facilitating coarse-to-fine visual reasoning. Meanwhile, tool-augmented trajectories interleave reasoning with radiology tool invocation, encouraging the model to adaptively acquire additional visual evidence. Extensive experiments show that our method significantly outperforms existing state-of-the-art models by 12.41% in average accuracy, while the ablation and user studies demonstrate the effectiveness of tool invocation. Code and benchmark will be available.

1 INTRODUCTION

Computed tomography (CT) is one of the most widely used non-invasive medical imaging modalities for clinical diagnosis, disease monitoring, and post-treatment follow-up (Haydel et al., 2000). It provides volumetric anatomical information that enables clinicians to inspect organs and surrounding structures from spatial perspectives. Recently, radiologists face an increasingly heavy workload as the number of CT examinations grows significantly worldwide (Bruls & Kwee, 2020), which underscores the importance of reliable AI-assisted systems that can facilitate CT interpretation (Oren et al., 2020; Topol, 2019). Recent advances in deep learning, particularly medical multimodal large language models (MLLMs) (Jiang et al., 2025; Xu et al., 2026), have promoted the development of several CT-oriented 3D MLLMs for tasks such as visual question answering and radiology report generation (Bai et al., 2024; Wu et al., 2025). These models have the potential to improve diagnostic efficiency and alleviate the growing clinical workload.

Despite these advances, existing 3D MLLMs remain limited in supporting comprehensive CT perception and interpretation in clinical practice. First, most existing models operate on a single CT scan, whereas clinical decision-making often requires integrating evidence across scans with different anatomical coverage, contrast phases and longitudinal visits (Eisenhauer et al., 2009). For example, as shown in Fig.1, specific types of lesion may be more conspicuous in a particular con-

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.

trast phase or reconstruction setting (Laghi et al., 2003), while assessing disease progression and treatment response requires comparison with prior examinations. Reasoning over an isolated scan may overlook complementary and longitudinal evidence needed for patient-level interpretation. Second, existing models are primarily trained to directly generate responses from fixed CT inputs (Wu et al., 2025; Lin et al., 2026), with limited support for interactive visual inspection. In contrast, radiologists routinely use radiology tools to assist and refine their interpretation in clinical practice (MacMahon et al., 2017). For example, when reviewing multiple scans from the same patient on a radiology workstation, radiologists progressively focus on the series and regions most relevant to the clinical question. During this process, they use radiology tools to adjust window settings, localize anatomical structures, and quantify findings such as nodules and cysts before reaching a final conclusion. This tool-assisted workflow enables radiologists to acquire and assess evidence that may not be readily discernible from the initial visual inputs. However, existing 3D medical MLLMs do not adequately support this clinically practical workflow, as they neither jointly interpret multiple CT scans nor invoke external tools for interactive reasoning. This gap motivates a multi-CT 3D MLLM that integrates evidence across scans and adaptively invokes radiology tools to acquire additional visual evidence.

To bridge these gaps, we formulate a challenging yet clinically practical setting, termed **multi-CT perception and understanding**. To the best of our knowledge, we present the first workflow that enables medical MLLMs to analyze multiple CT scans and proactively interact with radiology tools. Specifically, we investigate this setting from both data and methodological perspectives. First, given the lack of large-scale benchmarks for evaluating multi-CT reasoning capabilities, we introduce a clinically grounded visual question answering (VQA) benchmark for multi-CT interpretation, namely **CliMCT-Bench**. The benchmark is constructed from over 87k clinical studies from a medical institution, each comprising multiple CT scans and associated radiology reports. We then derive approximately 267K VQA pairs from the clinical findings documented in these reports, spanning a comprehensive range of clinical tasks, such as abnormality characterization across multiple CT scans and progression assessment across longitudinal examinations, serving as a pioneering benchmark for multi-CT reasoning.

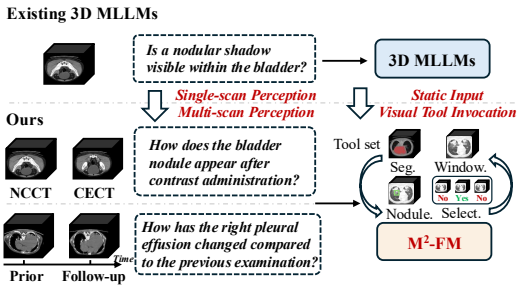


Figure 1: Comparison between our M^2-FM and existing 3D MLLMs. Our framework extends CT interpretation from single-scan to multi-scan, while enabling proactive tool invocation.

Furthermore, we propose M^2-FM, a **3D MLLM for multi-CT perception and understanding** that proactively **interacts with external radiology tools**. Specifically, M^2-FM adaptively invokes tools for scan selection, organ segmentation, window display adjustment, and lung nodule detection during visual question answering. To build M^2-FM from scratch, we develop a **progressive multi-stage training protocol**. First, we perform contrastive vision-language pretraining to establish the vision encoder's foundational visual perception capabilities. Next, we introduce two complementary training strategies beyond basic SFT. Specifically, we propose multi-turn, multi-scale SFT to develop the model's ability to perceive and interpret diverse visual regions. We further introduce tool-augmented SFT to enable the model to actively acquire necessary visual evidence by invoking radiology tools. Together, these complementary strategies enable M^2-FM to adaptively gather finer visual evidence for reasoning by invoking tools and perceive the acquired visual evidence across scales better, advancing CT interpretation from passive perception to interactive, tool-assisted visual reasoning. Extensive experiments demonstrate its effectiveness on CliMCT-Bench and the benefits of tool invocation. Our technical contributions are summarized as threefold:

- **Setting & Benchmark**: We are the first to formulate multi-CT perception and understanding as a clinically practical setting. Accordingly, we introduce CliMCT-Bench, a large-scale benchmark for multi-CT reasoning derived from real-world radiology studies.
- **Model**: We propose M^2-FM for multi-CT perception and understanding, which jointly analyzes multiple CT scans and proactively invokes appropriate tools to assist its interpretation.

- **Training Protocol**: We propose a progressive multi-stage training protocol to build M^2-FM. Beyond standard SFT, we propose multi-scale, multi-turn SFT to enhance global-to-local visual perception, and tool-augmented SFT to enable adaptive and effective radiology tool use.

2 Related Work

3D Medical Large Language Models. Recent medical MLLMs have extended multi-modal reasoning from 2D images (Wang et al., 2026b) to volumetric CT scans. RadFM (Wu et al., 2025) learns general radiological representations from large-scale 2D and 3D image-text data, while Merlin (Blankemeier et al., 2026) integrates abdominal CT with radiology reports and structured EHR data for diagnostic and prognostic tasks (Blankemeier et al., 2026). For chest CT, M3D-LaMed (Bai et al., 2024) provides a unified framework for 3D medical image-text retrieval, VQA, segmentation and so on (Bai et al., 2024). Hulu-Med (Jiang et al., 2025) further unifies medical text, 2D and 3D images, and video reasoning within a generalist VLM. More recent work has focused on improving the fidelity and efficiency of CT understanding tasks. For example, OmniCT jointly models slice-level and volume-level spatial context through spatial-consistency and organ-level alignment (Lin et al., 2026), while Photon reduces the computational cost through instruction-conditioned token scheduling (Fang et al., 2026). ***Despite these advances, existing methods predominantly operate on a single scan instead of integrating evidence across contrast phases and longitudinal examinations for clinically reliable interpretation.***

3D Medical Vision-Language Benchmarks. The growing interest in volumetric vision-language models has motivated several benchmarks for evaluating 3D medical understanding. As summarized in Table 1, M3D-Bench offers a comprehensive evaluation suite spanning image-text retrieval, VQA, segmentation and so on (Bai et al., 2024). More recently, 3D-RAD evaluates diagnostic reasoning through diverse CT-based VQA tasks, such as anomaly recognition and longitudinal temporal diagnosis (Gai et al., 2026). DeepTumorVQA further supports the evaluation of abdomen-focused questions (Chen et al., 2025). Despite their broad task coverage, these benchmarks still have several limitations: ***(i) These benchmarks primarily evaluate models on QA based on a single scan***. Even when temporal reasoning is considered, as in 3D-RAD (Gai et al., 2026), historical information is provided as textual context rather than through direct access to prior scans. ***(ii) These benchmarks may suffer from significant content redundancy and template-induced bias***, as their questions are generated from a limited set of CT volumes using fixed templates. Therefore, a clinically grounded benchmark built from real-world radiology studies and diverse multi-CT visual inputs is essential for evaluating 3D medical models in realistic clinical settings.

3 CliMCT-Bench

3.1 Data Collection

As shown in Table 1, most existing 3D medical VQA benchmarks focus on single-scan perception, with their supporting datasets often restricted to specific anatomical regions, such as DeepTumorVQA (Chen et al., 2025) and CT-RATE (Hamamci et al., 2026). These benchmarks offer limited support for jointly interpreting multiple CT scans and assessing longitudinal changes across examinations. To address these limitations, we curate CliMCT-Bench, a large-scale benchmark designed to support comprehensive multi-CT perception and understanding.

For data collection, we collaborated with a medical institution to collect CT examinations from patients across multiple clinical visits, with each examination potentially comprising multiple scans covering different anatomical regions and acquisition types. Our work is conducted according to the principles of the Declaration of Helsinki and all the data are de-identified. After filtering low-quality data, we collect 87,953 clinical visits from 44,697 patients, comprising 298,377 CT scans in total. During the data partitioning stage, we strictly followed a patient-level protocol to divide the dataset into training and test sets at a ratio of 10:1, ensuring that there is no patient-level data leakage.

Table 1: Comparison of existing 3D CT VQA benchmarks in terms of train/test pair numbers, involved scans and featured attributes.

Benchmark	Train	Test	Volumes	Coverage	Multi-scan	Temporal
M3D-VQA	481,898	13,792	∼120K	Whole body	×	×
3D-RAD	136,195	33,910	16,188	Chest	×	✓‡
DeepTumorVQA	355,962	39,650	9,262	Abdomen	×	×
Ours	260,219	7,046	298,377	Whole body	✓	✓

‡ Temporal questions use current CT volumes and textual historical abnormality labels, rather than CT scans.

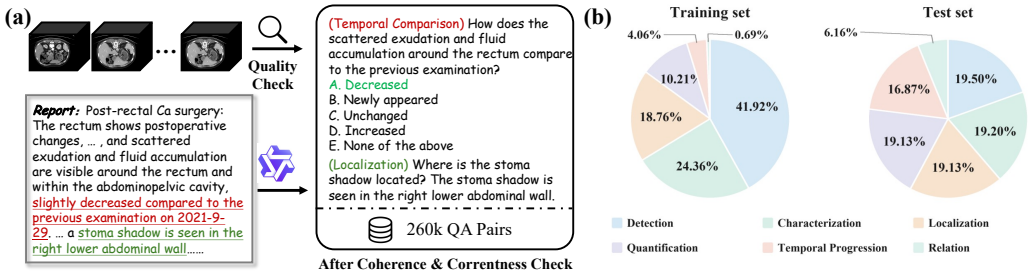


Figure 2: (a) Report-grounded VQA construction and quality-control pipeline, yielding 260K question-answer pairs for the training set. (b) Distributions of question categories.

3.2 Report-Grounded VQA Construction

Furthermore, we construct VQA samples from the radiology reports associated with each multi-CT study. The construction pipeline consists of three stages: data-quality filtering, question-answer generation, and quality control with balanced sampling. The overall pipeline is illustrated in Fig. 2-(a). First, we exclude studies that fail to meet predefined quality criteria. Second, we employ Qwen3.5-397B (Team, 2026) to extract clinically relevant statements from each report and transform them into multiple-choice or open-ended questions. To build a comprehensive corpus for multi-CT understanding, we prompt the LLM to generate questions spanning six complementary aspects, from low-level perception to longitudinal reasoning. **Detection** questions determine whether a finding or abnormality is present. **Characterization** questions examine qualitative attributes. **Localization** questions require anatomical localization at the level of laterality or sub-organ compartments. **Quantification** questions require explicit counting or measurement of anatomical structures and lesions. **Relational** questions examine relationships among findings and surrounding structures. **Temporal Comparison** questions ask how a finding has changed over time when given longitudinal studies.

Finally, the generated samples undergo LLM-assisted screening and human expert review to verify their coherence and clinical correctness. After removing ambiguous and redundant samples, we apply a controlled sampling to balance the question types, yielding approximately 260K VQA samples for the training dataset. Additionally, we use a more rigorous sampling strategy across question types, answer options, and content topics to reduce evaluation bias in the test set. Finally, we get 7,046 questions for evaluation. The distribution of question types is shown in Fig. 2-(b). Relation questions remain the rarest category because explicit relational statements are infrequent in routine radiology reports. We additionally balance the answer formats to maintain an approximately equal proportion of multiple-choice and open-ended questions in both the training and test sets. All temporal comparison questions are retained in the open-ended format, requiring models to interpret and describe longitudinal changes. More details can be found in the Appendix.

3.3 Tool-Augmented Trajectory Synthesis

Multi-turn multi-scale trajectory synthesis. Following the QA construction described in Section 3.2, we retain 150K QA pairs as single-turn data to train fundamental CT perception and question-answering capabilities, while reorganizing the remaining samples into multi-turn, multi-scale trajectories. Questions from the same patient are grouped into conversational sequences ordered from coarse to fine according to the required visual evidence. Each trajectory begins with

the original study-level input containing multiple CT series, followed by a question-relevant series displayed with an appropriate clinical window and, when available, a magnified region of interest derived from anatomical segmentation. During this process, VISTA3D (He et al., 2025) is used to produce the segmentation masks, while the windowed views use one of five clinical settings for the lung, mediastinum, abdomen, bone, or brain. We prompt Qwen3.5-397B-A17B to select the relevant series and window setting and identify the anatomical targets for ROI extraction. The bounding box enclosing these structures is then cropped and resampled to 256^3, simulating zoomed-in inspection for fine-grained perception. An example multi-turn trajectory can be seen in Fig.3. During this process, questions already assigned to the single-turn corpus are excluded to avoid duplication. This process produces 39,257 trajectories, comprising 31,139 three-turn and 8,118 two-turn dialogues. At each turn, the question is randomly formatted as either multiple-choice or open-ended.

Tool-augmented trajectory synthesis. We construct tool-augmented trajectories from failure cases of a direct-response baseline described in Section 4.3. The baseline model is evaluated on the collected training QA pairs, with errors identified through option matching for multiple-choice questions and a LLM evaluator for open-ended questions. For these challenging failure cases, we prompt Qwen3.5-397B-A17B (Team, 2026) to select the relevant CT series and invoke window adjustment, anatomical ROI extraction, or a pulmonary nodule detector when necessary. For temporal comparison, it selects relevant series from current and prior studies. Tool invocations and responses form interaction turns, with the returned information appended to the context. The generated trajectories are verified by radiologists and an LLM-based evaluator, yielding 8,145 high-quality trajectories. More details can be found in the Appendix.

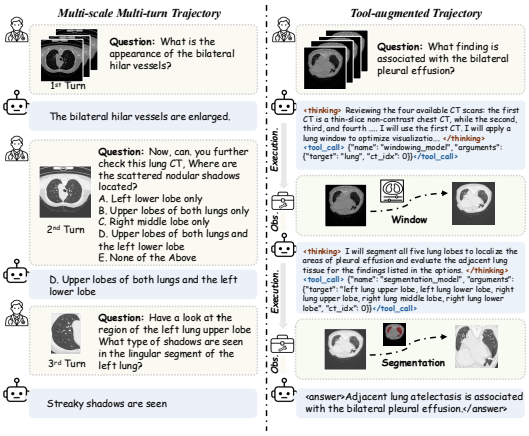


Figure 3: Examples of the synthesized multi-turn, multi-scale and tool-augmented trajectories.

Together, these two trajectory corpora provide synergistic supervision for interactive CT interpretation. As shown in Fig. 3, multi-turn, multi-scale trajectories enable the model to progressively perceive visual evidence from the complete context to localized salient regions. Tool-augmented trajectories complement this capability by teaching the model to recognize when the current visual context is insufficient and to invoke appropriate tools to acquire finer visual evidence.

4 METHODOLOGY

4.1 M^2-FM FRAMEWORK

As illustrated in Fig. 4, M^2-FM consists of a 3D vision encoder, a multimodal projector, a large language model, and a radiology toolset. Given a clinical question q and multiple CT scans $\{\mathbf{X}_i\}_{i=1}^{N}$, the model jointly analyzes all scans and interleaves the reasoning process with tools invocation.

3D vision encoder. We adopt a 3D Vision Transformer (3D ViT) (Dosovitskiy, 2020) as the vision encoder f_θ. Given a 3D CT scan $\mathbf{X}_i$, the encoder outputs patch tokens $\mathbf{V}_i = f_\theta(\mathbf{X}_i) \in \mathbb{R}^{L\times C}$, where L is the number of visual tokens and C is the feature embedding dimension.

Multi-modal projector and LLM. We employ a multimodal projector f_ϕ to align the visual features $\mathbf{V}$ into the latent embedding space of LLM. To accommodate the large number of visual tokens introduced by multi-scan CT inputs, f_ϕ incorporates a spatial pooling layer following (Bai et al., 2024), which aggregates neighboring token features while preserving their spatial organization. The pooled token sequences are then projected into the LLM's embedding space as $\mathbf{E}_i = f_\phi(\mathbf{V}_i) \in \mathbb{R}^{\tilde{L}\times d}$, where $\tilde{L}$ denotes the number of visual tokens after pooling and d is the LLM hidden size.

Finally, we adopt the LLM of Qwen3-VL-8B (Yang et al., 2025) as the language backbone. The LLM conditions on the visual embeddings of all scans, the clinical question q, and the interaction history to generate its response:

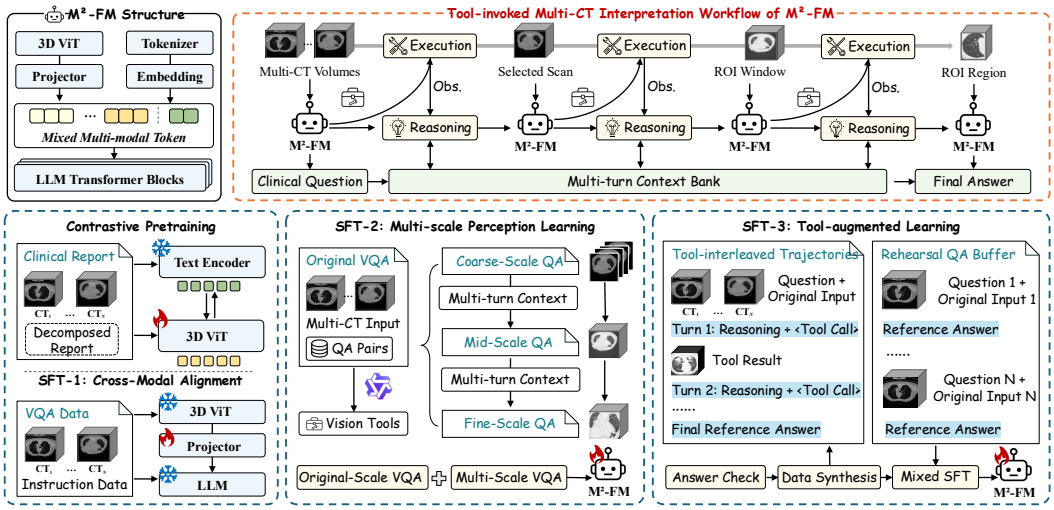


Figure 4: Illustration of the tool-invoked workflow and progressive training strategy of M^2-FM.

$$\mathbf{S} = [\mathrm{Tok}(q); \mathbf{E}_1; \mathbf{E}_2; \ldots; \mathbf{E}_N; \mathrm{Tok}(\mathcal{H})], \qquad p_\psi(y \mid q, \mathcal{X}, \mathcal{H}) = \prod_{t=1}^{|y|} p_\psi(y_t \mid \mathbf{S}, y_{<t}), \tag{1}$$

where $\mathbf{E}_i$ is the projected visual-token sequence of the i-th scan, and $\mathrm{Tok}(\cdot)$ denotes text tokenization. $\mathcal{H}$ represents the interaction history, $[\cdot\,;\cdot]$ denotes sequence concatenation, y_t is the t-th output token, and ψ denotes the parameters of the LLM.

4.2 Tool-Augmented Multi-Turn Reasoning

Radiology Toolset. To enable M^2-FM to perform fine-grained visual reasoning during multi-CT perception as radiologists, we expose several radiological operations as callable tools. The *selection* tool jointly considers the question and all available CT scans to identify those most relevant to the task. This selection process is embedded internally within the model's reasoning chain, as it requires no interaction with external tools. The *anatomy segmentation* tool uses VISTA3D (He et al., 2025) to localize the requested structures and returns a magnified regional volume. The *window adjustment* tool re-renders a selected series using one of several predefined tissue-specific display windows. Finally, the *pulmonary nodule detection* tool returns structured nodule locations and quantitative measurements. Together, these tools transform M^2-FM from a passive interpreter of fixed CT inputs into an active evidence-seeking model that can select, enhance, localize, and quantify task-relevant findings during visual reasoning.

Tool-Augmented Multi-Turn Reasoning. Inspired by tool-augmented MLLMs (Jiang et al., 2026; Yao et al., 2022; Schick et al., 2023), M^2-FM interleaves reasoning with radiology tool invocation across multiple turns. M^2-FM dynamically determines whether additional evidence is required and selects an appropriate tool during reasoning. At the t-th turn, conditioned on the clinical question q, multi-CT visual embeddings $\{\mathbf{E}_i\}_{i=1}^N$, and interaction history $\mathcal{H}_{t-1}$, the model generates an action a_t, which can be either a structured tool call or a direct answer:

$$a_t \sim p_\psi\big(\cdot \mid q, \{\mathbf{E}_i\}_{i=1}^N, \mathcal{H}_{t-1}\big), \tag{2}$$

where p_ψ denotes the output distribution of the LLM. If $a_t \in \mathcal{A}_{\text{tool}}$, the selected tool is executed to produce an observation $o_t = \mathcal{T}(a_t)$. The model then integrates the returned observation with the accumulated context and generates its subsequent reasoning:

$$r_t \sim p_\psi\big(\cdot \mid q, \{\mathbf{E}_i\}_{i=1}^N, \mathcal{H}_{t-1}, a_t, o_t\big). \tag{3}$$

The interaction state is updated according to whether the model invokes a tool or directly answers the question:

$$(\mathcal{H}_t, y) = \begin{cases} (\mathcal{H}_{t-1} \oplus (a_t, o_t, r_t), \varnothing), & a_t \in \mathcal{A}_{\text{tool}}, \\ (\mathcal{H}_{t-1}, a_t), & a_t \in \mathcal{A}_{\text{answer}}, \end{cases} \tag{4}$$

where $\mathcal{A}_{\text{tool}}$ and $\mathcal{A}_{\text{answer}}$ denote the tool-call and direct-answer action spaces, respectively, and $\oplus$ denotes context concatenation. For a tool call, visual observations are encoded through the same vision pathway, whereas textual observations are appended directly to the context. The updated history $\mathcal{H}_t$ informs the next turn, in which the model invokes another tool or produces final answers.

4.3 Progressive Training Pipeline

We train M^2-FM through progressive stages that establish foundational visual representations, cross-modal alignment, multi-scale perception and adaptive tool invocation, respectively.

Contrastive Pretraining. We first perform contrastive pretraining of our 3D ViT on multi-CT studies, where multiple CT scans share a single radiology report. Directly pairing every scan with the complete report may introduce mismatched semantic supervision because different report sections can describe distinct anatomical regions or acquisition types. Therefore, we first leverage Qwen3.5-397B-A17B to decompose each report and assign the relevant sentences to individual scans according to their semantics. Consequently, each CT scan is paired with its corresponding report sections. Then we train the 3D ViT using a standard symmetric contrastive objective (Radford et al., 2021; Fu et al., 2026). In this stage, we apply RadBERT (Chambon et al., 2022) as the text encoder. The contrastive pretraining enables the vision tower to produce powerful visual representations before instruction tuning.

Multi-scale Supervised Fine-tuning. Visual supervised fine-tuning proceeds in two steps. We first freeze the vision encoder and LLM and train only the multimodal projector on single-turn QA pairs. This warm-up stage bridges the representation gap between the pretrained 3D visual features and the LLM embedding space while preserving the capabilities of both backbones. We then train the model end-to-end on a mixture of single-turn data and the multi-turn, multi-scale trajectories introduced in Section 3.3. The multi-turn trajectories progressively present visual input from complete multi-CT visual context to localized anatomical regions. Synergistically, the single-turn samples preserve foundational CT perception and direct question-answering abilities, while the multi-turn trajectories enhance the ability of M^2-FM to perceive visual evidence from global context to fine-grained regions. Notably, to further diversify the training data, we additionally sample 76K and 80K QA pairs from 3D-RAD and DeepTumorVQA, respectively, and incorporate them into the instruction-tuning data.

Tool-Augmented Supervised Fine-tuning. Finally, we perform tool-augmented supervised fine-tuning to extend the model from coarse-to-fine visual perception to proactive tool invocation. The training data are constructed from single-turn questions incorrectly answered by the preceding baseline SFT model, which is trained solely on single-turn QA pairs. These samples expose cases in which direct inference from the initial multi-CT inputs remains unreliable. We synthesize tool-augmented trajectories from these hard cases, as described in Section 3.3. Furthermore, we observe that training solely on tool-augmented trajectories induces mode collapse toward indiscriminate tool invocation. We therefore design a **rehearsal mechanism** inspired by rehearsal-based continual learning (Luo et al., 2025). Specifically, we construct a rehearsal buffer comprising a subset of questions that the preceding model answers correctly. We then jointly train the model on the tool-augmented trajectories and rehearsal samples, preserving its direct-response capability while enabling it to invoke radiology tools only when necessary.

5 Experiments

5.1 Implementation Details

M^2-FM-8B employs the pretrained LLM backbone of Qwen3-VL-8B as the large language model component, integrating our pretrained 3D ViT as the vision encoder. We apply a standard preprocessing pipeline including coordinate reorientation, Hounsfield unit (HU) clipping, resampling, and intensity normalization. For evaluation, we apply the same evaluation protocol to all models. For multiple-choice questions, we compute accuracy by directly matching the predicted options against the ground-truth answers. For open-ended questions, we use Qwen3.5-397B-A17B as an LLM judge to assess whether each response is semantically consistent with the reference answer.

Table 2: Comparison results of our method and other state-of-the-art models on CliMCT-Bench. Results are reported as accuracy (%).

Model	Multiple-choice Question					Open-ended Question						Avg.
	Chara.	Det.	Loc.	Quan.	Rel.	Chara.	Det.	Loc.	Quan.	Rel.	Temp.	
Commercial Models												
GPT-5.5	67.50	83.48	64.49	53.94	70.91	66.57	83.19	35.26	21.63	69.63	67.62	61.62
General-purpose MLLMs												
InternVL3-8B	50.66	57.54	49.03	42.50	50.45	42.60	51.17	22.37	4.15	42.99	35.07	39.64
Qwen3-VL-8B	49.34	53.48	46.21	41.90	50.45	42.46	50.29	20.15	6.22	30.37	37.59	38.72
Qwen3.5-9B	53.77	74.20	49.48	44.13	51.82	45.41	68.42	13.63	14.22	50.00	36.25	44.29
Qwen3.6-27B	65.58	65.07	47.99	46.81	58.64	48.52	59.65	17.33	20.74	32.24	55.59	48.01
Medical MLLMs												
Hulu-Med-7B	69.72	69.28	55.72	51.56	61.82	55.33	78.51	17.48	20.89	55.61	45.58	51.65
Hulu-Med-32B	76.37	75.51	56.32	55.72	68.64	56.95	76.90	16.00	26.37	58.88	51.64	55.07
Lingshu-7B	59.23	57.68	54.23	46.21	54.09	37.28	59.65	16.00	17.04	28.04	41.63	43.03
Lingshu-32B	68.24	62.17	60.48	51.71	57.73	49.85	59.65	17.48	13.04	30.84	47.01	47.53
HealthGPT-8B	75.33	63.48	57.21	52.60	57.73	51.04	71.49	13.93	21.63	51.87	49.45	50.91
3D Medical MLLMs												
M3D-Phi-3-4B	56.57	66.09	49.33	35.51	54.09	39.05	61.84	18.07	7.26	48.60	33.05	40.93
M3D-Llama-2-7B	40.47	52.46	38.34	28.53	50.45	35.06	59.94	17.48	4.59	54.21	32.97	35.50
RadFM-14B	31.17	49.86	21.40	24.96	32.27	40.09	57.89	17.04	9.48	57.94	34.23	32.86
Merlin	29.10	43.77	27.49	12.63	35.91	30.62	53.80	15.41	6.52	22.43	29.27	27.92
OmniCT-7B	69.28	63.33	52.90	43.68	61.82	50.44	70.91	15.56	22.96	50.00	43.99	48.37
Photon-3B	59.23	61.74	51.26	53.19	54.55	38.17	65.64	10.52	5.78	36.92	37.09	42.39
M^2-FM-8B	**94.24**	**94.78**	**75.04**	**83.21**	**81.36**	**82.54**	**91.67**	**38.52**	**34.67**	**79.44**	**69.89**	**74.03**

5.2 Main Results

We compare M^2-FM with a broad range of baselines, including the commercial model GPT-5.5, general-purpose MLLMs (InternVL3 (Zhu et al., 2025) and the Qwen series (Yang et al., 2025; Team, 2026; Qwen Team, 2026)), medical MLLMs (Hulu-Med (Jiang et al., 2025), Lingshu (Xu et al., 2026), and HealthGPT (Lin et al., 2025)), and 3D medical MLLMs (M3D-LaMed (Bai et al., 2024), RadFM (Wu et al., 2025), Merlin (Blankemeier et al., 2026), OmniCT (Lin et al., 2026), and Photon (Fang et al., 2026)). Implementation details and configurations for all compared baselines are provided in the appendix. As shown in Table 2, M^2-FM achieves the highest accuracy across all question categories and answer formats on our CliMCT-Bench, surpassing the strongest baseline, GPT-5.5, by 12.41% in overall accuracy. Notably, none of the evaluated 3D medical MLLMs are designed to accommodate multiple CT volumes. This input mismatch limits their ability to integrate cross-scan evidence and may explain why they underperform some general-purpose MLLMs on our benchmark. Additionally, to assess whether multi-CT training preserves single-volume understanding, we further evaluate M^2-FM on 3D-RAD and DeepTumorVQA. Following (Wang et al., 2026a), we construct balanced evaluation subsets by reducing redundancy and answer-position bias. As shown in Table 3, our model achieves average accuracies of 60.27% and 73.15%, respectively, demonstrating competitive performance against both general-purpose and medical MLLMs. Notably, M^2-FM uses only 76K and 80K sampled QA pairs from the two benchmarks, demonstrating the competitive single-CT performance when trained on only subsets of these benchmarks.

5.3 Ablation Analysis

Ablation studies. As shown in Table 4, we progressively add the proposed training components to evaluate their individual contributions. The final training protocol improves the average accuracy to 74.03%, compared with a vanilla SFT baseline. Notably, directly introducing tool-augmented SFT without rehearsal reduces the accuracy to 68.79%, indicating that tool-use supervision alone biases the model toward excessive tool invocation and harms its direct visual perception and question-answering capabilities. Incorporating the rehearsal mechanism mitigates this issue and raises the accuracy to 74.03%. Table 5 further groups questions by the number of interaction turns used by the tool-enabled model. For the challenging questions requiring two or three turns, the direct-response baseline achieves only 27.04% and 32.21%, whereas M^2-FM improves these results to 36.05% and 40.27%. Finally, Fig. 5 shows that the model trained without rehearsal invokes tools on nearly every question, while rehearsal allows M^2-FM to answer 87.4% of questions directly and reserve effective tool use for difficult cases.

Table 3: Comparison results of our method and other state-of-the-art models on 3D-RAD and DeepTumorVQA. Results are reported as accuracy (%). For Photon-3B†, we report the task-specific performance of the two specialist models fine-tuned on 3D-RAD and DeepTumorVQA, respectively.

Model	3D-RAD							DeepTumorVQA				
	E.D.	S.T.	L.T.	I.O.	A.D.	M.C.	**Avg.**	Recog.	Meas.	Vis.R.	Med.R.	**Avg.**
Commercial Models												
GPT-5.5	57.64	53.36	75.29	45.41	39.65	17.86	53.02	71.00	32.85	45.85	36.30	49.57
General-purpose MLLMs												
InternVL3-8B	49.85	23.67	49.91	30.86	27.98	7.58	35.30	84.13	29.92	41.17	41.23	50.25
Qwen3-VL-8B	52.64	20.36	39.30	29.98	26.86	7.88	32.67	65.40	30.60	42.11	37.53	45.92
Qwen3.5-9B	51.33	62.41	59.94	34.67	29.78	18.96	47.57	47.53	29.92	41.58	41.23	41.32
Qwen3.6-27B	53.48	62.51	66.97	41.89	36.50	**22.06**	51.70	85.00	22.61	37.02	38.02	46.95
Medical MLLMs												
Hulu-Med-7B	60.33	64.50	81.45	44.63	39.91	18.96	57.40	86.00	27.56	37.02	30.86	47.45
Hulu-Med-32B	59.02	**67.80**	**81.80**	45.80	41.56	18.76	58.29	85.07	30.60	41.05	29.14	49.70
Lingshu-7B	52.23	13.71	15.32	25.49	22.36	13.17	25.18	86.53	31.27	43.11	29.88	51.32
Lingshu-32B	55.59	17.30	59.00	35.74	29.26	21.16	38.80	72.27	39.15	44.35	27.90	49.45
HealthGPT-8B	55.19	60.04	71.46	48.44	41.45	19.46	53.80	83.73	27.11	36.49	32.10	46.62
3D Medical MLLMs												
M3D-Phi-3-4B	49.85	13.30	13.89	24.51	20.52	8.18	23.44	48.80	25.98	31.72	37.53	35.53
M3D-Llama-2-7B	41.99	14.10	15.11	27.05	24.34	7.98	23.07	8.33	22.16	24.20	35.80	20.72
RadFM-14B	50.08	47.23	48.28	24.22	23.59	8.78	38.76	47.07	24.52	32.10	37.28	35.07
Merlin	49.51	12.74	13.82	21.00	22.02	7.68	23.23	53.00	22.95	30.32	35.80	35.27
OmniCT-7B	53.48	62.23	78.73	36.43	34.81	20.26	53.27	63.47	24.07	30.19	30.62	37.63
Photon-3B †	56.16	65.44	81.76	45.31	42.31	19.86	57.34	83.27	**73.34**	**67.94**	**61.23**	72.12
M^2-FM-8B	**67.72**	65.78	80.33	**48.83**	**44.34**	18.86	**60.27**	**92.40**	70.98	66.56	58.77	**73.15**

Table 4: Ablation results of multi-scale SFT (M.S.), tool-augmented SFT (T.S.), and the rehearsal mechanism (R.M.) on CliMCT-Bench.

M.S.	T.S.	R.M.	Chara.	Det.	Loc.	Quan.	Rel.	Temp.	Avg.
			85.74	91.99	51.19	57.34	74.65	66.19	70.93
✓			87.80	92.58	52.45	54.97	77.88	66.27	71.44
✓	✓		83.44	89.96	49.55	54.08	76.50	63.33	68.79
✓	✓	✓	**88.40**	**93.23**	**56.75**	**58.90**	**80.41**	**69.89**	**74.03**

Radiologist Evaluation. To evaluate the quality of the tool-invoked trajectories generated by our method, three radiologists with over five years of clinical experience independently assessed 20 randomly sampled trajectories on a five-point scale. The evaluation considered the correctness and logical coherence of the given trajectory, with higher scores indicating better quality. The radiologists achieved mean scores of 4.60, 4.60, and 4.40, respectively, yielding an overall mean of 4.53. The weighted Fleiss' kappa was 0.838, indicating strong inter-rater agreement.

Table 5: Accuracy of the base model and M^2-FM, grouped by the number of interaction turns.

Turns	**Base**	**M^2-FM**
1	77.58%	79.29%
2	27.04%	36.05%
3	32.21%	40.27%

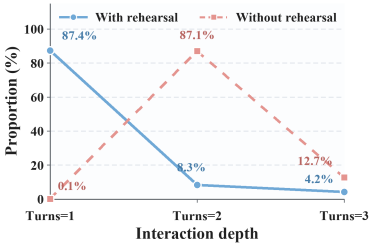


Figure 5: Distribution of interaction turns with and without rehearsal.

6 Conclusion

In this paper, we introduce the first formulation of multi-CT perception and understanding, extending 3D medical MLLMs from isolated-volume question answering to joint reasoning across scans and longitudinal visits. To support this setting, we construct CliMCT-Bench with 267K report-grounded QA pairs spanning six clinical capabilities from basic perception to temporal reasoning. We further propose M^2-FM, a unified model that jointly processes multiple CT scans and interleaves reasoning with adaptive radiology tool invocation. Its key innovation lies in two complementary trajectory designs: multi-turn, multi-scale trajectories enhance coarse-to-fine visual perception, while tool-augmented trajectories teach the model to invoke appropriate tools when necessary. M^2-FM surpasses the strongest baseline on CliMCT-Bench while retaining competitive performance on single-CT benchmarks, demonstrating the value of multi-scan integration and adaptive tool use for comprehensive CT understanding.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic data sets, implement methods, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, and polish and proofread the manuscript. We did not use generative AI tools to assist in the writing of proofs, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, interpret results. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study involves human evaluation conducted by clinical experts. All participants were clearly informed of the purpose, scope, and potential risks of the study, and their participation was entirely voluntary with informed consent obtained beforehand. Ethical approval has been obtained for all data used in the research. Our work is conducted according to the tenets of Helsinki Declaration and all the data are de-identified. We have adhered to the ICLR Code of Ethics throughout the study, ensuring fairness, transparency, and compliance with ethical and legal standards.

REPRODUCIBILITY STATEMENT

We have made extensive efforts to ensure the reproducibility of our work. All preprocessing steps, implementation details, hyperparameter settings and experiment configurations are described in detail in the main text and appendix. We will release all the source code, pretrained weight and benchmark after the double-blind review.

REFERENCES

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

Louis Blankemeier, Ashwin Kumar, Joseph Paul Cohen, Jiaming Liu, Longchao Liu, Dave Van Veen, Syed Jamal Safdar Gardezi, Hongkun Yu, Magdalini Paschali, Zhihong Chen, et al. Merlin: a computed tomography vision–language foundation model and dataset. *Nature*, 652 (8112):1318–1328, 2026.

RJM Bruls and RM Kwee. Workload for radiologists during on-call hours: dramatic increase in the past 15 years. *Insights into imaging*, 11(1):121, 2020.

M Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022.

Pierre Chambon, Tessa S. Cook, and Curtis P. Langlotz. Improved fine-tuning of in-domain transformer model for inferring covid-19 presence in multi-institutional radiology reports. *Journal of Digital Imaging*, 2022. doi: 10.1007/s10278-022-00714-8.

Yixiong Chen, Wenjie Xiao, Pedro RAS Bassi, Xinze Zhou, Sezgin Er, Ibrahim Ethem Hamamci, Zongwei Zhou, and Alan Yuille. Are vision language models ready for clinical diagnosis? a 3d medical benchmark for tumor-centric visual question answering. *arXiv preprint arXiv:2505.18915*, 2025.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Elizabeth A Eisenhauer, Patrick Therasse, Jan Bogaerts, Lawrence H Schwartz, Danielle Sargent, Robert Ford, Janet Dancey, Stephen Arbuck, Steve Gwyther, Margaret Mooney, et al. New response evaluation criteria in solid tumours: revised recist guideline (version 1.1). *European journal of cancer*, 45(2):228–247, 2009.

Chengyu Fang, Heng Guo, Zheng Jiang, Chunming He, Xiu Li, and Minfeng Xu. Photon: Speedup volume understanding with efficient multimodal large language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=xsSJw6jJBL.

Yibing Fu, Yunpeng Zhao, Zhitao Zeng, Cheng Chen, and Yueming Jin. Unleashing the power of image-tabular self-supervised learning via breaking cross-tabular barriers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4049–4057, 2026.

Xiaotang Gai, Jiaxiang Liu, Yichen Li, Zijie Meng, Jian Wu, and Zuozhu Liu. 3d-rad: A comprehensive 3d radiology med-vqa dataset with multi-temporal analysis and diverse diagnostic tasks. *Advances in Neural Information Processing Systems*, 38, 2026.

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

Micelle J Haydel, Charles A Preston, Trevor J Mills, Samuel Luber, Erick Blaudeau, and Peter MC DeBlieux. Indications for computed tomography in patients with minor head injury. *New England Journal of Medicine*, 343(2):100–105, 2000.

Yufan He, Pengfei Guo, Yucheng Tang, Andriy Myronenko, Vishwesh Nath, Ziyue Xu, Dong Yang, Can Zhao, Benjamin Simon, Mason Belue, et al. Vista3d: A unified segmentation foundation model for 3d medical imaging. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20863–20873. IEEE, 2025.

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.

Yankai Jiang, Yujie Zhang, Peng Zhang, Wenjie Li, Yichen Li, Jintai Chen, Xiaoming Shi, and Shihui Zhen. Ophiuchus: Incentivizing tool-augmented "think with images" for joint medical segmentation, understanding and reasoning. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=coJqVkqb03.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Andrea Laghi, Riccardo Iannaccone, Plinio Rossi, Iacopo Carbone, Riccardo Ferrari, Filippo Mangiapane, Italo Nofroni, and Roberto Passariello. Hepatocellular carcinoma: detection with triple-phase multi–detector row helical ct in patients with chronic hepatitis. *Radiology*, 226(2):543–549, 2003.

Tianwei Lin, Wenqiao Zhang, Sijing Li, Yuqian Yuan, Binhe Yu, Haoyuan Li, Wanggui He, Hao Jiang, Mengze Li, Xiaohui Song, et al. Healthgpt: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. *arXiv preprint arXiv:2502.09838*, 2025.

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 *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=nrZI64gTvC.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

Heber MacMahon, David P Naidich, Jin Mo Goo, Kyung Soo Lee, Ann NC Leung, John R Mayo, Atul C Mehta, Yoshiharu Ohno, Charles A Powell, Mathias Prokop, et al. Guidelines for management of incidental pulmonary nodules detected on ct images: from the fleischner society 2017. *Radiology*, 284(1):228–243, 2017.

Ohad Oren, Bernard J Gersh, and Deepak L Bhatt. Artificial intelligence in medical imaging: switching from radiographic pathological data to clinically meaningful endpoints. *The Lancet Digital Health*, 2(9):e486–e488, 2020.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, April 2026. URL `https://qwen.ai/blog?id=qwen3.6-27b`.

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

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems*, 36:68539–68551, 2023.

Qwen Team. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56, 2019.

Fakrul Islam Tushar, Avivah Wang, Lavsen Dahal, Ehsan Samei, Michael R Harowicz, Jayashree Kalpathy-Cramer, Kyle J Lafata, Tina D Tailor, Cynthia Rudin, and Joseph Y Lo. Reproducible benchmarking for lung nodule detection and malignancy classification across multiple low-dose ct datasets. *arXiv preprint arXiv:2405.04605*, 2024.

Ziyue Wang, Linghan Cai, Chang Han Low, Haofeng Liu, Junde Wu, Jingyu Wang, Rui Wang, Lei Song, Jiang Bian, Jingjing Fu, et al. 3dmedagent: Unified perception-to-understanding for 3d medical analysis. *arXiv preprint arXiv:2602.18064*, 2026a.

Ziyue Wang, Junde Wu, Linghan Cai, Chang Low, Xihong Yang, Qiaxuan Li, and Yueming Jin. Medagent-pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow. In *International Conference on Learning Representations*, volume 2026, pp. 114360–114389, 2026b.

Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5):e230024, 2023.

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

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: Generalist foundation model for unified multimodal medical understanding and reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

APPENDIX

Table of Contents

- **§A: Additional Details of CliMCT-Bench**
  - §A.1: Data Collection, Preprocessing, and Split
  - §A.2: Representative Question Examples
- **§B: Trajectory Synthesis Details**
  - §B.1: Multi-Turn Multi-Scale Trajectories
  - §B.2: Tool-Augmented Trajectories
- **§C: Radiology Toolset**
  - §C.1: Tool Definitions
  - §C.2: Tool Invocation Format
- **§D: Model and Implementation Details**
  - §D.1: Model Architecture
  - §D.2: Training Configuration
- **§E: Evaluation Details**
  - §E.1: Evaluation and System Prompt
  - §E.2: Baseline Setup
  - §E.3: External Benchmark Sampling
  - §E.4: Visualization of Tool-Augmented Trajectories

A ADDITIONAL DETAILS OF CLIMCT-BENCH

A.1 DATA COLLECTION, FILTERING, PREPROCESSING, AND SPLIT

Cohort. CliMCT-Bench is built from a retrospective, single-institution archive of whole-body CT examinations. We filter the studies in two stages. At the report level, we exclude studies with missing or unparseable reports. At the scan level, we exclude volumes containing fewer than 20 slices, thereby removing localizer images, and retain only volumetric CT scans. The resulting corpus contains 87,953 studies, 298,377 CT series, and 44,697 unique patients. Table 6 summarizes the distributions of series per study and studies per patient, while Fig. 6 shows the anatomical distribution of the included scans.

Volume preprocessing. Each CT volume is reoriented to the right-anterior-superior (RAS) coordinate system, resampled to an isotropic voxel spacing of $1 \times 1 \times 1$ mm^3, and clipped to $[-1000, 1000]$ HU using MONAI (Cardoso et al., 2022). The volume is then resized to $256 \times 256 \times 256$, linearly normalized to $[0, 1]$, and stored in `float16` format.

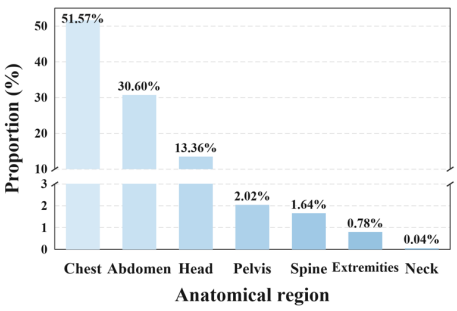


Figure 6: Distribution of anatomical regions represented in the CT dataset.

Split and question distribution. We perform patient-level stratification, yielding 40,697 patients and 80,027 studies for training and 4,000 patients and 7,926 studies for evaluation. To ensure visual diversity and question variety, we apply category-aware sampling while retaining at least one question from each study. For the evaluation set, we further balance question categories and answer-option positions to minimize potential evaluation bias. The final training and evaluation corpora contain 260,219 and 7,046 QA pairs, respectively. Table 7 reports the joint distribution across the six primary categories and two answer formats. All temporal comparison questions are retained in an

Table 6: Series-per-study and study-per-patient distributions of CliMCT-Bench. Counts of 8 or more are merged into the last bucket.

CT series per study			**Studies per patient**		
n	studies	%	n	patients	%
1	13,500	15.35	1	25,673	57.44
2	14,218	16.17	2	9,443	21.13
3	32,038	36.43	3	4,172	9.33
4	8,446	9.60	4	2,166	4.85
5	7,585	8.62	5	1,198	2.68
6	5,084	5.78	6	733	1.64
7	3,092	3.52	7	480	1.07
$\geq$8	3,990	4.54	$\geq$8	832	1.86
total	87,953	100.00	total	44,697	100.00
mean	3.39		mean	1.97	
max	34		max	42	

Table 7: Question-category and answer-format distributions of the CliMCT-Bench training corpus and evaluation set.

	Training (260,219)				**Evaluation** (7,046)			
Category	n	%	MCQ	Open	n	%	MCQ	Open
Detection	109,075	41.92	53,510	55,565	1,374	19.50	690	684
Characterization	63,396	24.36	31,287	32,109	1,353	19.20	677	676
Localization	48,826	18.76	24,409	24,417	1,348	19.13	673	675
Quantification	26,570	10.21	14,629	11,941	1,348	19.13	673	675
Temporal comparison	10,566	4.06	0	10,566	1,189	16.87	0	1,189
Relation	1,786	0.69	931	855	434	6.16	220	214
Total	**260,219**	**100.00**	**124,766**	**135,453**	**7,046**	**100.00**	**2,933**	**4,113**

open-ended format to encourage genuine understanding of longitudinal changes. Relational Reasoning does not reach the target quota because corresponding descriptions are scarce in the original reports. Question-level overlap between the multiple-choice and open-ended subsets is exactly zero. We enforce this separation by maintaining all previously selected question identifiers in an exclusion list, thereby maximizing the diversity of the training corpora.

A.2 Representative Question Examples

The following items present representative examples of the question categories in our benchmark. Image placeholders are reproduced exactly as presented to the model.

- **Detection / Presence**: What type of shadow is observed in the gallbladder? *Answer*: A nodular dense shadow is observed in the gallbladder.
- **Characterization**: What is the status of the spleen? A. Calcified B. Normal size C. Splenomegaly D. Atrophy E. None of the Above. *Answer*: C. Splenomegaly.
- **Localization**: Which specific rib exhibits bone destruction? A. Left 8th rib B. Right 6th rib C. Left 6th rib D. Left 4th rib E. None of the Above. *Answer*: C. Left 6th rib.
- **Quantification**: What is the approximate size of the largest lung nodule? *Answer*: Approximately 5$\times$7mm.
- **Relation**: Into which structure do both ureters converge? *Answer*: Both ureters are seen converging into the rectum.
- **Temporal Progression**: What is the interval change of the lung lesions compared to the prior CT? *Answer*: The lesions are increased.

B TRAJECTORY SYNTHESIS DETAILS

B.1 MULTI-TURN MULTI-SCALE TRAJECTORIES

We reorganize report-grounded questions from the training corpus into patient-level dialogues that provide progressively finer visual evidence. Questions associated with the same study are grouped to preserve visual context across turns, while those already used in the single-turn corpus are excluded to maintain question-level diversity.

Each dialogue contains two or three turns ordered by the visual evidence required to answer the questions. The first turn uses the complete study-level input containing multiple CT series. The second turn, when available, presents a question-relevant series under an appropriate clinical window. The final turn provides a magnified ROI derived from anatomical segmentation. Qwen3.5-397B-A17B selects the relevant series, window setting, and anatomical targets based on the question and series metadata. Series selection prioritizes anatomical coverage, followed by the clinically appropriate contrast phase and slice thickness. VISTA3D produces the segmentation masks; we then crop the minimal axis-aligned bounding box enclosing the selected structures and resample it to 256^3. This organization simulates progressive inspection from the complete multi-CT study to a targeted series and finally to a localized anatomical region.

The final corpus contains 39,257 dialogues and 109,653 turns. Of these, 31,139 dialogues contain three turns and 8,118 contain two turns, with the latter progressing directly from the complete study to an ROI view. Questions are independently rendered in either multiple-choice or open-ended format at each turn, yielding 54,766 multiple-choice and 54,887 open-ended turns.

B.2 TOOL-AUGMENTED TRAJECTORIES

Tool-augmented trajectories are constructed from questions answered incorrectly by a direct-response baseline trained on the single-turn corpus. Multiple-choice failures are identified through exact option-letter matching, while open-ended and temporal failures are determined by the LLM evaluator described in Section E.1. Focusing trajectory synthesis on these errors targets questions for which direct inference from the initial multi-CT input remains unreliable and additional inspection is most useful.

Trajectory generation. For each hard case, Qwen3.5-397B-A17B receives the clinical question, answer options when applicable, reference answer, supporting report sentence, and textual metadata describing the anatomical coverage, contrast phase, and slice thickness of each candidate series. Because the model cannot directly process 3D CT files during trajectory synthesis, it first selects the relevant series and then invokes CT window adjustment, anatomical ROI extraction, or pulmonary nodule detection according to the semantics of the question. Each tool call and response forms an interaction turn, and the returned visual or textual observation is appended to the context before further reasoning or answer generation. Temporal questions use a separate construction template: the model selects exactly one series from the prior study and one from the current study with compatible anatomical coverage and acquisition characteristics, then passes both indices to the visual tools in prior-to-current order.

Leakage prevention and verification. We prevent generated tool calls from encoding the correct answer through three constraints. First, ROI targets include every candidate structure mentioned in the question and all answer options, rather than only the structure associated with the reference answer. Second, findings and impressions parsed from the reports are excluded from the series-selection rationale and tool arguments.

All trajectories undergo deterministic validation of tool names, argument schemas, series indices, anatomical targets, and returned observations. To assess whether an ROI provides sufficient evidence for visual perception questions, we concatenate the contiguous key axial slices of each complete CT into a montage image and present it to Claude-Opus-4.8. This representation is necessary because Claude-Opus-4.8 cannot directly process 3D volume files. For pulmonary-nodule measurement questions, we first run the detector offline to produce a textual summary of all detections. We then prompt Qwen3.5-397B-A17 to answer each question from the detection summary in a zero-

Table 8: Clinical display settings used by the CT window-adjustment tool.

Window	Level (HU)	Width (HU)	Clip range (HU)	Primary use
Lung	-600	1500	$[-1350, 150]$	Parenchyma, airways, and nodules
Mediastinum	40	400	$[-160, 240]$	Heart, vessels, and lymph nodes
Abdomen	40	400	$[-160, 240]$	Abdominal organs and soft tissue
Bone	500	1500	$[-250, 1250]$	Vertebrae, ribs, and cortical lesions
Brain	40	80	$[0, 80]$	Brain parenchyma and haemorrhage

shot setting and retain the trajectory only when the answer is correct. Following verification and balanced sampling, the complete corpus contains 8,145 tool-augmented trajectories.

C Radiology Toolset

C.1 Tool Definitions

The radiology toolset supports internal series selection, CT window adjustment, anatomical ROI extraction, and pulmonary nodule detection. Window adjustment and ROI extraction return CT volumes, whereas the detector returns structured text. Series selection is performed internally rather than exposed as a separate tool; the selected zero-based index is passed directly to the requested operation. This design avoids an interaction turn that provides no new evidence.

C.1.1 CT Series Selection

The model selects series according to three ordered criteria. Anatomical coverage is considered first to ensure that the target structure is fully represented. Among compatible series, contrast-enhanced scans are preferred for enhancement-, vascular-, and tumour-related questions, whereas non-contrast scans are preferred for findings such as calcification, stones, and acute haemorrhage. Slice thickness serves as the final criterion, with thin-slice reconstructions preferred when the preceding criteria are satisfied. If no series adequately covers the requested anatomy, the model does not invoke a visual tool. All indices are zero-based with respect to the original series attached to the question. Windowing and ROI extraction accept either a scalar index or a list of indices, with the latter used for paired prior and current studies. The nodule detector accepts a single series index.

C.1.2 CT Window Adjustment

The windowing tool re-renders a selected CT series using one of five clinical display settings. Intensities are clipped to $[L - W/2, L + W/2]$, linearly mapped to $[0, 1]$, and resampled to 256^3. The numerical parameters are not exposed to the model; instead, the tool descriptions specify the anatomical structures and tissue properties emphasized by each setting. Based on the question, the model selects one of the five predefined settings rather than specifying L and W directly. Table 8 summarizes these settings.

C.1.3 Anatomical Segmentation

ROI extraction uses VISTA3D with the released pretrained weights. The model selects targets from the VISTA3D target-class vocabulary. When an exact label is unavailable, the nearest containing structure provides approximate spatial guidance. Given one or more labels, the tool forms the union of their masks, extracts the minimal axis-aligned bounding box without dilation, and resamples the resulting crop to 256^3. Cropping, rather than masking, retains the surrounding tissue within the box. Multiple targets produce a single merged crop, while a multi-series request returns one crop per series.

C.1.4 Pulmonary Nodule Detection

The detector follows a two-stage pipeline comprising a 3D RetinaNet nodule detector and a 3D ResNet-50 malignancy classifier. Both models are taken from a published lung-screening benchmark

(Tushar et al., 2024) and use the released weights. Lung-lobe segmentation gates the detector to identify candidates outside the lungs. Detection runs at $0.703 \times 0.703 \times 1.25$ mm spacing; the detector clips intensities to $[-1024, 300]$ HU and the classifier to $[-1000, 500]$ HU over a 64^3 cube around each candidate centre. The tool consumes a single volume with its *original* Hounsfield values and voxel spacing. Four stages run sequentially, each writing a persistent artifact so that processing can resume from an intermediate stage rather than restart:

- A whole-body organ segmentation model produces a 117-class map (Wasserthal et al., 2023), from which the five pulmonary lobes are retained.
- If the map contains no pulmonary tissue, the detector is skipped entirely and an empty report records the reason.
- The detector is applied with a sliding window of $512 \times 512 \times 208$ voxels with 0.25 overlap and Gaussian blending, non-maximum suppression at 0.22, and at most 100 detections per volume, all at the released operating values. Each surviving candidate is then cropped as a 64^3 patch and scored individually by the malignancy classifier. Because the crop is taken on the resampled grid, it spans roughly $45 \times 45 \times 80$ mm rather than an isotropic 64 mm; out-of-bounds voxels are padded with -1000 HU. Detections are written with a permissive score floor of 0.02.
- Finally, each candidate centre is resolved against the lobe map with a 2 mm tolerance, producing one of four states: inside a lobe, within tolerance of a lobe boundary, extrapulmonary, or outside the segmented volume. Boundary cases are treated as intrapulmonary. This stage writes the only file the tool reads at inference time, recording per-candidate detection score, malignancy probability, centre coordinates in millimetres, three axis extents, lobe, and lobe state.

The detector first retains candidates above 0.02 and then keeps those whose detection score or malignancy probability is at least 0.5 and whose centres lie within a lung lobe. Candidates are ranked by the higher of the two scores and limited to ten. Lung filtering is performed before truncation because 62% of routed cases contain extrapulmonary detections and 13% rank an extrapulmonary detection first. Lobe membership is determined from the segmentation map with a 2 mm tolerance.

Unlike the other tools, the detector returns a textual summary rather than a rendered volume. Each nodule is reported with its three axis extents in millimetres, the two largest of those extents, its lobe, its detection score, and its malignancy probability, with nodules ordered by confidence. Voxel coordinates are deliberately not exposed. A typical response is:

```
find_lung_nodule(lobe='right lower lobe') on series
[NSlice=318]
note:  filtered to right lower lobe
1 nodule(s), highest confidence first:
 1.  size 4.2 x 4.2 x 4.1 mm (longest 4.2 mm, second 4.2
mm); lobe=right lower lobe; det=0.99; p_cancer=0.10
```

C.2 Tool Invocation Format

Tool calls follow the Hermes convention and are represented as JSON objects enclosed by `<tool_call>` tags. A `<tool_response>` message contains one image placeholder per returned volume or the textual detector report. Representative calls are:

```
{"name":  "windowing_model", "arguments":  {"target":
"brain", "ct_idx":  0}}
{"name":  "segmentation_model", "arguments":  {"target":
"liver, spleen", "ct_idx":  1}}
{"name":  "find_lung_nodule", "arguments":  {"ct_idx":  0,
"lobe":  "right lower lobe"}}
```

The execution layer validates tool names, series indices, anatomical targets, and lobe names. Invalid calls return a readable error message, which is appended to the interaction history. Visual observations are encoded through the same 3D vision pathway as the original CT inputs, while detector responses are appended as text. Training trajectories contain at most three interactions, and inference

Table 9: Hyperparameters for contrastive pretraining and the three supervised training stages.

	Contrastive pretraining	**Projector warm-up**	**Multi-scale SFT**	**Tool-augmented SFT**
Epochs	10	1	1	1
Effective batch size	768	32	32	32
Max sequence length	512 tok.	12,000	12,000	12,000
Vision-encoder/LLM learning rate	1×10^{-4}	frozen	1×10^{-6}	1×10^{-6}
Projector learning rate	N/A	6×10^{-4}	1×10^{-6}	1×10^{-6}

is capped at five turns. We use the same response-formatting code during trajectory construction and inference to prevent train-test discrepancies.

D Model and Implementation Details

D.1 Model Architecture

We replace the native 2D vision tower with a 3D Vision Transformer that operates directly on single-channel CT volumes. Each $256 \times 256 \times 256$ input is partitioned into non-overlapping 16^3 patches, producing a $16^3 = 4{,}096$ patch grid. The encoder contains 12 transformer layers with a hidden dimension of 768, 12 attention heads, and an MLP dimension of 3,072. A class token is prepended during encoding and discarded after the final block.

To control the context length for multi-series inputs, the final 16^3 token grid undergoes spatial average pooling with a stride of two along all three axes, producing $8^3 = 512$ tokens per volume without additional learned parameters. The projector comprises LayerNorm with $\epsilon = 10^{-6}$, a $768 \to 768$ linear layer, GELU, and a $768 \to d_{\text{LLM}}$ linear layer. For the 8B language model, $d_{\text{LLM}} = 4{,}096$, and the projector contains 3.74M parameters.

We use Qwen3-VL-8B-Instruct as the language backbone. We modify its configuration only to accept single-channel volumetric inputs and disable the original DeepStack feature injection, which is defined for 2D feature maps.

D.2 Training Configuration

All stages are trained on a single node with eight NVIDIA H200 GPUs (140 GiB each), 192 CPU cores, and 2 TB of host memory. We use eight-way data parallelism, bfloat16 precision, DeepSpeed ZeRO-2 without optimizer offloading. Table 9 lists all key hyperparameters. Existing LLM inference frameworks provide extensive serving optimizations, including KV-cache management and continuous batching. To leverage these optimizations for 3D CT inference, we implement an out-of-tree plugin that integrates our 3D vision encoder and multimodal projector into the inference framework (Kwon et al., 2023) while retaining its optimized language-model runtime.

E Evaluation Details

E.1 Evaluation and System Prompt

Multiple-choice predictions are scored by exact matching after extracting the leading option letter. Open-ended answers are evaluated by Qwen3.5-397B-A17B as an LLM evaluator. The evaluator receives the question, reference answer, and prediction and returns only `YES` or `NO`. The system prompt and open-ended answer evaluation prompt are provided below.

System Prompt

```
You are a radiology assistant that answers a visual question grounded in one or more 3
    D CT volumes from the same patient. You have three tools available:
```

```
  - windowing_model    | re-render a chosen CT with an appropriate display window
  - segmentation_model | localize an anatomical target on a chosen CT
  - find_lung_nodule   | run the lung-nodule CAD detector on a chosen CT and read back
      its measurements

When multiple CT scans are provided, first review them inside <thinking> and decide
     which CT(s) are most relevant. Then call the tools on the chosen CT(s) by passing
     the 0-based index as the `ct_idx` argument. Decide which tools to call and when
     based on the question.

You MUST wrap the entire reasoning and tool interaction inside a single `<thinking>
     ... </thinking>` block, and put the final short answer inside a single `<answer>
     ... </answer>` block. `<thinking>` must be the very first thing you output.
     Nothing may appear before `<thinking>` or after `</answer>`.

TOOL SEMANTICS

`windowing_model`
- Re-renders the selected CT (identified by `ct_idx`) with a preset display window (HU
      clip + normalize) and returns the windowed full CT as an image. Contrast only;
     does not change anatomy.
- Arguments:
    - `target`: named window. MUST be exactly one of:
        - `lung`        | pulmonary parenchyma, airway lumen, pulmonary nodules,
     ground-glass opacity, emphysema, pneumothorax.
        - `mediastinum` | heart, pericardium, coronary calcification, great vessels (
     aorta, pulmonary trunk), mediastinal lymph nodes, thymus, esophagus, thyroid,
     breast, chest wall soft tissue.
        - `abdomen`     | liver, gallbladder, biliary tree, spleen, pancreas, kidneys,
      adrenal glands, stomach, bowel, bladder, prostate, uterus, ovaries, peritoneum,
     abdominal aorta, portal vein, ascites.
        - `bone`        | any vertebrae, ribs, sternum, pelvis, long bones, joint
     surfaces, cortical bone lesions, calcifications on bone.
        - `brain`       | cerebral parenchyma, ventricles, basal ganglia, midline
     structures, hemorrhage, edema, infarct.
    - `ct_idx`: 0-based index of the target CT among the N CTs shown in the user
     message. Pass a single integer to window one CT, or a list of integers (e.g. `[0,
     4]`) to window several CTs at once.

`segmentation_model`
- Localizes anatomical structures on the CT identified by `ct_idx` and returns the
     cropped 3D region of interest (ROI). Localization only | does not detect, classify
     , or confirm any pathology. Use the ROI to focus your inspection; attribute
     observations to your own reading, not to the tool.
- Arguments:
    - `target`: one or more labels taken verbatim from the LABEL VOCABULARY below,
     joined by a comma and a space. Never invent a label outside the vocabulary.
    - `ct_idx`: 0-based index of the target CT (a single integer, or a list of
     integers). When windowing_model was also called on the same case, use the same
     ct_idx.

When a parent, containing, or adjacent seg label is used because no exact label exists
     , treat the ROI as approximate spatial guidance and phrase the paragraph
     accordingly.

`find_lung_nodule`
- Runs an automated lung-nodule CAD detector on the CT identified by `ct_idx` and
     returns a TEXT report, not an image. Each detected nodule is listed with its size
     in millimetres (width x height x depth), the lung lobe it sits in, a detection
     score and a malignancy probability, ordered by confidence.
- Use it when the question asks how large a pulmonary nodule is, or how many there are
     . It supplies the measurement directly, so no windowing or segmentation call is
     needed for such a question.
- The reported millimetre sizes come from the detector and may be quoted as the
     measurement. For a question phrased as "N x N mm" or "the diameter", use the two
     largest of the three dimensions.
- The detector covers nodules inside the lung only. It is not a general lesion finder:
      do not call it for liver, kidney, mediastinal lymph node, or other non-pulmonary
     findings.
- It may return no nodule. That is an absence of detection, not proof of a normal lung
     ; fall back on your own reading of the images in that case.
- Arguments:
    - `ct_idx`: 0-based index of the target CT (a single integer).
    - `lobe`: OPTIONAL. Restrict the report to one lung lobe. MUST be exactly one of:
     `left upper lobe`, `left lower lobe`, `right upper lobe`, `right middle lobe`, `
     right lower lobe`. Pass it when the question names a lobe; omit it to receive
     every detected nodule.
```

```
LABEL VOCABULARY (use these labels verbatim in the segmentation `target`)

liver, spleen, pancreas, right kidney, aorta, inferior vena cava, right adrenal gland,
      left adrenal gland, gallbladder, esophagus, stomach, duodenum, left kidney,
     bladder, portal vein and splenic vein, small bowel, brain, lung tumor, pancreatic
     tumor, hepatic vessel, hepatic tumor, colon cancer primaries, left lung upper lobe
     , left lung lower lobe, right lung upper lobe, right lung middle lobe, right lung
     lower lobe, vertebrae L5, vertebrae L4, vertebrae L3, vertebrae L2, vertebrae L1,
     vertebrae T12, vertebrae T11, vertebrae T10, vertebrae T9, vertebrae T8, vertebrae
      T7, vertebrae T6, vertebrae T5, vertebrae T4, vertebrae T3, vertebrae T2,
     vertebrae T1, vertebrae C7, vertebrae C6, vertebrae C5, vertebrae C4, vertebrae C3
     , vertebrae C2, vertebrae C1, trachea, left iliac artery, right iliac artery, left
      iliac vena, right iliac vena, colon, left rib 1, left rib 2, left rib 3, left rib
      4, left rib 5, left rib 6, left rib 7, left rib 8, left rib 9, left rib 10, left
     rib 11, left rib 12, right rib 1, right rib 2, right rib 3, right rib 4, right rib
      5, right rib 6, right rib 7, right rib 8, right rib 9, right rib 10, right rib
     11, right rib 12, left humerus, right humerus, left scapula, right scapula, left
     clavicula, right clavicula, left femur, right femur, left hip, right hip, sacrum,
     left gluteus maximus, right gluteus maximus, left gluteus medius, right gluteus
     medius, left gluteus minimus, right gluteus minimus, left autochthon, right
     autochthon, left iliopsoas, right iliopsoas, left atrial appendage,
     brachiocephalic trunk, left brachiocephalic vein, right brachiocephalic vein, left
      common carotid artery, right common carotid artery, costal cartilages, heart,
     left kidney cyst, right kidney cyst, prostate, pulmonary vein, skull, spinal cord,
      sternum, left subclavian artery, right subclavian artery, superior vena cava,
     thyroid gland, vertebrae S1, bone lesion, airway

LABEL MAPPING

- Use the exact label when one exists. Otherwise use the smallest parent, containing,
     or directly adjacent structure that can spatially localize the target; the ROI
     then serves as approximate guidance only.
- Do not use a label merely because it is clinically associated with a condition.
     Accepted closest-label mappings include:
    - pericardial effusion / coronary artery calcification -> heart
    - pleural effusion -> adjacent involved lung-lobe labels
    - aortic stent / aortic dissection -> aorta
    - biliary tree / bile duct -> gallbladder
- Use tumor labels (lung tumor, hepatic tumor, pancreatic tumor, colon cancer
     primaries) ONLY when the imaging and question context indicate malignancy or
     neoplasm; do not map a nonspecific nodule, cyst, or inflammatory focus to a tumor
     label.

ANSWER FORMAT

- The `<answer>` MUST contain only the concise final answer to the question, with no
     additional commentary before or after.
- For multiple-choice questions, the `<answer>` MUST faithfully reproduce BOTH the
     letter AND the full option text as given in the question (e.g. "A. Diffusely
     decreased"). Do not shorten, paraphrase, or output only the letter | the full
     option content is required so that the answer can be parsed unambiguously.
- For open-set / short-answer questions, output a short natural-language answer.
```

Open-Ended Answer Evaluation Prompt

```
You are a radiology grader.  Decide whether the generated answer conveys the same
clinical fact as the reference answer.
## MEASUREMENTS (apply this first when both answers contain numbers)
Compare the numbers position by position.  Tolerance is 1 mm on EVERY number.
- Single values:  YES if they differ by at most 1 mm.
  "10mm" vs "11 mm" -> YES      "15mm" vs "14mm" -> YES
  "10mm" vs "13mm"   -> NO
- Dimension pairs:  every dimension must be within 1 mm.
  "6mm x 6mm" vs "6x5mm" -> YES
- Ranges:  compute |low1-low2| and |high1-high2|.  YES if BOTH are <= 1.
  Overlapping alone is NOT enough --- check the two bounds numerically.
  "2-5mm" vs "2-4mm" -> diffs 0,1 -> YES    "3-4mm" vs "3-6mm" -> diffs 0,2 -> NO
- A single value vs a range:  YES only if the value is inside the range or within 1 mm
of it.
Units and spacing do not matter ("11 mm" == "11mm").
## NON-NUMERIC ANSWERS
Ask yourself:  would a radiologist reading the generated answer be misled about
where the finding is, what it is, or whether it exists?  If no, answer YES.
Answer YES if:
```

```
- Same clinical fact, different wording or synonyms.
  Treat these descriptor families as equivalent:  strip-like ~ streaky ~ linear ~
band-like;
  patchy ~ flocculent ~ blurry; low-density ~ hypodense ~ cerebrospinal-fluid density;
  bilateral ~ both / on both sides.
- The generated answer adds detail INSIDE the region the reference names, without
  excluding any part of it.
  ref "in both lower lungs" vs gen "in the dorsal aspects of both lower lungs" -> YES
- The generated answer names fewer structures than the reference but everything it
  states is correct and it answers the question.
  ref "the trachea and bronchi are patent" vs gen "yes, the trachea is patent" -> YES
Answer NO if:
- Opposite polarity:  present vs absent.
- Wrong laterality:  a finding the reference puts on one side is put on the other.
  ("bilateral" == "both lungs" --- SAME, not a laterality error.)
- Different organ, different lesion type, or a different finding altogether.
- The generated answer names a region that lies partly OUTSIDE the reference's region
  (not merely a smaller part of it).
  ref "in both lungs" vs gen "in the right upper lobe and left lung" -> NO
- The generated answer is empty or answers a different question.
Reference (GT): {gt}
Generated:        {gen}
Output ONLY one word:  YES or NO.
```

E.2 Baseline Setup

All baselines are evaluated using their official inference wrappers, native preprocessing procedures, and input resolutions. Each model receives the same clinical question and complete set of CT series, converted only as required by its interface. Multiple-choice prompts request a single option letter, whereas open-ended prompts request a concise medical answer. To support multi-CT input for 2D general-purpose and medical MLLMs, we concatenate the axial slices of each CT into montage images. For 3D MLLMs, Qwen3.5-397B-A selects the most relevant CT series, which is then provided to the 3D MLLM for inference.

E.3 External Benchmark Sampling

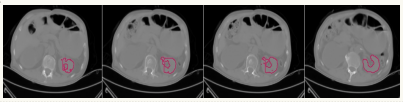
We incorporate subsets of 3D-RAD and DeepTumorVQA into supervised fine-tuning to increase anatomical and question diversity without allowing these substantially larger public corpora to dominate the training mixture. For training, we sample 76K QA pairs from 3D-RAD and 80K from DeepTumorVQA. All public CT volumes are processed using the same reorientation, resampling, intensity clipping, and 256^3 resizing pipeline as the in-house data. For evaluation, we remove redundant questions and apply balanced sampling to improve efficiency and reduce evaluation bias.

3D-RAD. The original test split contains 33,910 questions across six tasks and is dominated by existence detection, which accounts for 69% of the samples and has a majority-class accuracy of 82%. We construct a balanced evaluation subset of 13,415 questions. Existence detection is downsampled to 2,977 questions with equal numbers of positive and negative answers distributed across its 18 findings. The remaining tasks are retained in full, including 2,873 questions for each of the two temporal tasks and 4,692 open-ended questions across the other three tasks.

DeepTumorVQA. The original validation split contains 39,650 questions with substantial redundancy and template-based descriptions. We therefore construct a balanced evaluation subset of 6K questions by balancing question subtypes, templates, and answer positions while preferentially sampling less frequently used volumes. The resulting subset covers 929 studies and all 29 question subtypes, with no overlap in volumes or questions between the training and evaluation sets.

E.4 Visualization of Tool-Augmented Trajectories

Figure 7 presents two representative tool-augmented trajectories. In the first case, M^2-FM identifies the abdominal CT relevant to the renal question and invokes anatomical segmentation to obtain a localized view of the left kidney, from which it recognizes multiple cystic low-density shadows. In the second case, the model selects the thin-slice chest CT and invokes the pulmonary nodule

Questions: The CT is <video>\nWhat type of shadows are visible within the left renal parenchyma?
Answer:
<thinking>Reviewing the available CT scan: the first CT is a thick-slice non-contrast chest and abdomen CT. Since the question concerns the left renal parenchyma, which is located in the abdomen, this CT provides the necessary anatomical coverage, so I will use the first CT. The selected CT is a non-contrast CT of the chest and abdomen. I will segment the left kidney to localize the specific region of interest within the renal parenchyma. **</thinking>**
<tool_call>{\"name\": \"segmentation_model\", \"arguments\": {\"target\": \"left kidney\", \"ct_idx\": 0}}**</tool_call>**
Tool response:

<answer>Multiple cystic low-density shadows are visible within the left renal parenchyma.**</answer>**
Reference: Multiple cystic low-density shadows are visible within the left renal parenchyma.

Questions: The CT is <video> The CT is <video> The CT is <video>\nWhat is the approximate diameter of the nodule in the left lower lung lobe?
Answer:
<thinking> Reviewing the three available CT scans: the first CT is a thin-slice non-contrast chest CT, the second is a thick-slice non-contrast chest CT, and the third is also a thick-slice non-contrast chest CT. Since the question asks for the approximate diameter of a lung nodule, which requires high spatial resolution for accurate measurement, I will use the first CT. The selected CT is a non-contrast CT of the chest. I will run the lung-nodule detector on this series to obtain the nodule measurements. **</thinking>** **<tool_call>**{\"name\": \"find_lung_nodule\", \"arguments\": {\"ct_idx\": 0}}**</tool_call>**
Tool response:
1. size 15.1 x 15.6 x 15.9 mm; lobe=left upper lobe; det=0.99; p_cancer=0.31
2. size 10.5 x 10.5 x 10.5 mm; lobe=right lower lobe; det=0.97; p_cancer=0.28
3. size 13.7 x 14.6 x 14.8 mm; lobe=right middle lobe; det=0.97; p_cancer=0.39
4. size 10.8 x 10.7 x 11.6 mm ;lobe=**left lower lobe**; det=0.94; p_cancer=0.14
5. …..
<answer> The nodule in the left lower lung lobe measures approximately 10mm in diameter.**</answer>**
Reference: The approximate diameter is 10 mm

Figure 7: Representative tool-augmented reasoning trajectories of M^2-FM. In the left example, the model selects the relevant abdominal CT and invokes anatomical segmentation to inspect the left kidney. In the right example, it selects a thin-slice chest CT and invokes the pulmonary nodule detector to obtain quantitative measurements. Both trajectories demonstrate adaptive scan selection, tool invocation, and evidence-grounded answer generation.

detector to retrieve quantitative measurements, enabling it to determine the approximate diameter of the nodule in the left lower lobe. These examples illustrate how M^2-FM adaptively selects the relevant scans and tools based on the required evidence and grounds its final answer in the returned visual or textual observations.