CT-Flow: Tool-Use Trajectories for Agentic 3D CT Interpretation

Anonymous ACL submission

Abstract

Recent advances in Large Vision–Language Models (LVLMs) have shown strong potential for multi-modal radiological reasoning. However, most approaches to 3D CT analysis rely on single-pass inference over a fixed volumetric encoding, committing to a query-independent summary before the question is known and risking the loss of fine-grained evidence. We reframe 3D CT understanding as *query-conditioned evidence acquisition*, where the model retrieves evidence from the raw volume on demand rather than encoding it once. We propose CT-Flow, an agentic framework that equips an LVLM with four executable tools — volume navigation, region measurement, structure segmentation, and radiomics analysis — and trains it to decompose a query into tool calls, observe the returned evidence, and synthesize a grounded answer. We further construct CT-TrajBench, an oracle-grounded benchmark whose answers are computed by executing tools on raw volumes, with a three-stage rejection filter ensuring each instance requires volume-grounded tool execution rather than textual priors alone. On CT-TrajBench and 3D-RAD, CT-Flow consistently outperforms end-to-end medical LVLM baselines; the fine-tuned CT-Flow-8B reaches 69.46% accuracy on 3D-RAD, a 20.4-point improvement over the Qwen3-VL-8B baseline.

1 Introduction

Large Vision-Language Models (LVLMs) have shown strong performance in 2D visual understanding, yet extending them to high-resolution 3D modalities such as Computed Tomography (CT) remains challenging due to the representation of volumetric inputs. A typical chest CT contains hundreds of slices at sub-millimeter resolution, exceeding the token budgets that LVLMs can effectively process. To obtain a fixed-length representation, existing approaches either aggregate voxel features with specialized 3D encoders (Bai et al., 2024) or serialize volumes into long sequences of axial slices (Hamamci et al., 2024). Both strategies can discard fine-grained evidence required by downstream queries, such as small lesions, Hounsfield-unit attenuation values, and precise volumetric measurements. Together, these limitations create a mismatch between query-independent volumetric summaries and CT questions that require localized, quantitative, and query-specific evidence.

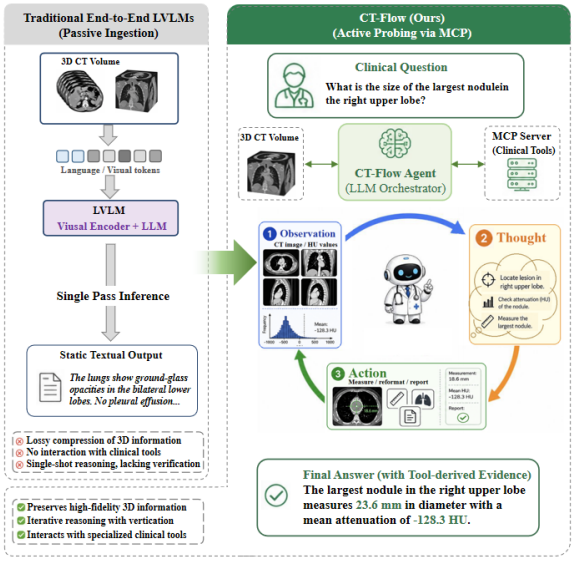

Figure 1: Comparison of 3D CT analysis paradigms. Left: end-to-end LVLMs encode the full volume once into a fixed representation and answer questions from a query-independent summary. Right: CT-Flow formulates 3D CT understanding as query-conditioned evidence acquisition, where the LVLM orchestrates executable tools to retrieve targeted evidence from the raw volume on demand.

The mismatch is not fully addressed by representation design alone. When a CT volume is compressed into a fixed token budget before the question is known, the model must form a query-independent summary that may omit evidence needed for a specific query. A more suit-

able formulation is to treat 3D CT understanding as **query-conditioned evidence acquisition** rather than query-independent volumetric encoding. Instead of summarizing the volume up front, the LVLM retrieves evidence from the raw volume on demand through executable tools for slice navigation, region measurement, structure segmentation, and radiomic feature extraction. Each tool call returns evidence conditioned on the current reasoning state, allowing representation cost to be allocated only to information required by the query.

Such a formulation further requires training and evaluation data that can verify whether the model uses appropriate visual evidence. Existing 3D CT VQA datasets are often derived from radiology reports and may inherit textual artifacts. As a result, a model may answer some questions from report-level priors without directly grounding its reasoning in the underlying volume. Report-derived data are therefore limited for training and evaluating evidence-grounded tool-use agents.

We present **CT-Flow**, an agentic framework that equips an LVLM with four executable tools: volume navigation, region measurement, structure segmentation, and radiomics analysis. Given a query, the model learns to decompose it into a sequence of tool calls, observe the returned evidence, and synthesize an answer grounded in those observations. We further construct **CT-TrajBench**, an oracle-grounded benchmark in which each ground-truth answer is computed by executing tools directly on the raw CT volume. A three-stage rejection filter removes instances that can be solved from textual priors alone. The resulting dataset contains approximately 4,300 training trajectories and 690 verified evaluation instances.

Experiments on established 3D CT VQA benchmarks and CT-TrajBench show that CT-Flow achieves higher performance than end-to-end medical LVLM baselines and exhibits stronger cross-benchmark generalization after fine-tuning. The results indicate that tool-mediated evidence acquisition can help mitigate representation bottlenecks in high-resolution 3D modalities. The main contributions are summarized as follows:

- We formulate 3D CT understanding as query-conditioned evidence acquisition, addressing the mismatch between fixed volumetric representations and query-specific clinical evidence.
- We present **CT-Flow**, an agentic LVLM framework that uses executable tools for volume navigation, region measurement, structure segmentation, and radiomics analysis, and construct **CT-TrajBench**, an oracle-grounded benchmark of verified tool-use trajectories with a three-stage rejection filter.
- Experiments demonstrate that CT-Flow improves over end-to-end medical LVLM baselines on multiple 3D CT benchmarks and produces inspectable tool-execution traces for clinical reasoning.

2 Related Work

3D Volumetric Interpretation in Medical VLMs. Existing 3D medical VLMs broadly follow two paradigms, both of which commit to a query-independent volumetric representation before reasoning begins. *Native 3D approaches* encode volumes with 3D backbones: RadFM (Wu et al., 2025) unifies 2D and 3D images via a 3D ViT, and M3D (Bai et al., 2024) processes volumetric data via 3D spatial pooling. These methods incur high computational cost and tend to retain only coarse-grained features (Ates et al., 2025). *Serialization-based methods* treat CT volumes as long slice sequences: OmniCT (Lin et al.) aggregates cross-slice context via sequence modeling, Hulu-Med (Jiang et al., 2025) decomposes 3D volumes into constituent slices within a unified medical VLM, and Vote-MI (Wang et al., 2024b) selects representative slices to reduce computation at the risk of discarding diagnostically critical cues. Both share a common limitation: the volume is compressed into a fixed summary before the query is observed, leaving fine-grained evidence such as small lesions, attenuation values, and volumetric measurements inaccessible at reasoning time.

Tool-Augmented Reasoning Agents. A growing line of work equips LLMs with external tools to extend their reasoning beyond parametric knowledge. Early efforts such as Toolformer (Schick et al., 2023) teach LLMs to invoke APIs through self-supervised demonstrations, while ReAct (Yao et al., 2022) interleaves reasoning steps with tool invocations and environment observations, establishing a general template for agentic LLMs. Subsequent work has scaled this paradigm to larger tool inventories and longer-horizon planning: ToolLLM (Qin et al., 2024) trains LLMs on thousands of real-world APIs, and Voyager (Wang et al.) couples an LLM controller with a continually expanding skill

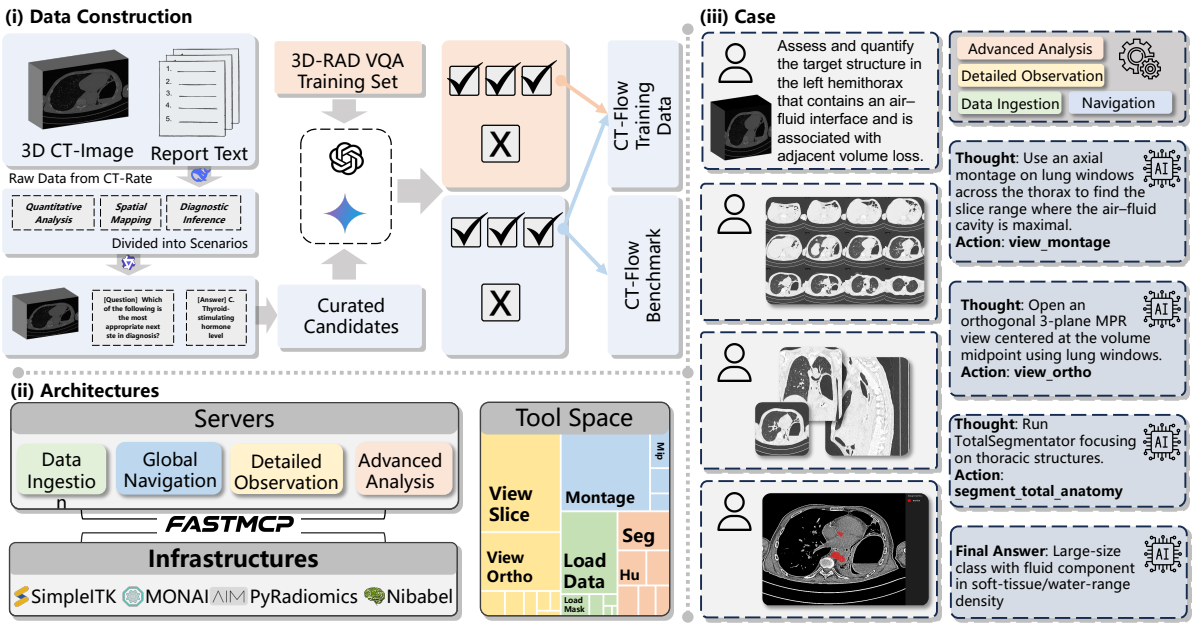

Figure 2: Overview of the CT-Flow framework. (i) Data Construction: The pipeline for raw data curation, trajectory synthesis, and the establishment of the CT-Flow benchmark. (ii) Architectures: The system decouples the LLM orchestrator from the imaging environment via FASTMCP, bridging high-level servers with medical imaging infrastructures to provide a suite of atomic tools in the Tool Space. (iii) Case Study: A demonstration of a Language-Action Trajectory $\mathcal{T}$. The orchestrator performs Active Probing by iteratively generating reasoning states (s_t), executing tool calls (a_t), and interpreting high-fidelity observations (o_t) to reach a grounded diagnostic answer.

library for open-ended decision making. In the multimodal setting, HuggingGPT (Shen et al., 2023) and Visual ChatGPT (Wu et al., 2023) use an LLM to orchestrate specialized 2D vision models for image-level tasks. While most prior work focuses on text or 2D visual inputs, designing and training LVLM agents with grounded, high-resolution 3D trajectories remains largely unexplored.

Medical Agents and Clinical Tool Use. LLM-orchestrated tools have gained traction in clinical AI. ChatCAD (Wang et al., 2024a) and ChatCAD+ (Zhao et al., 2024) use LLMs to assemble outputs from computer-aided diagnosis modules into interactive reports. Med-Agents (Tang et al., 2024) explores multi-agent collaboration for cross-specialty clinical reasoning, and MedRAX (Fallahpour et al., 2025) performs strategic planning for radiology tasks over 2D chest X-rays. These systems demonstrate that LLMs can coordinate clinical tools, but they predominantly operate on 2D inputs or on pre-extracted textual findings, and they are evaluated only on whether the final answer is correct — not on whether it is grounded in the executed tool outputs. CT-Flow differs in two respects: it targets the 3D volumetric setting in which the choice of evidence is itself part of the problem, and it is trained and evaluated on oracle-grounded trajectories that verify the evidence chain rather than only the final answer.

3 Methodology

CT-Flow reframes 3D CT interpretation from passive volumetric encoding into executable tool-use trajectories. Given a CT volume V and a clinical query Q, conventional LVLMs directly predict an answer $A = f_\theta(V, Q)$ from a fixed visual representation. In contrast, CT-Flow treats interpretation as a sequential evidence acquisition process: the model actively selects clinical actions, executes them on the original volume, observes tool-derived evidence, and synthesizes the final answer from the accumulated trajectory. This design reflects the way radiologists inspect CT studies through iterative localization, visual confirmation, and quantitative verification, rather than through a single static observation.

3.1 Executable Clinical Action Space

We define a clinical action space $\mathcal{A}$ of executable tools for volumetric CT analysis. Each action takes structured arguments, operates on the source volume V, and returns observations such as visual evidence, coordinates, anatomical labels, masks, or

measurements. These actions are implemented as MCP-based tools with structured inputs and outputs. Compared with hardcoded interfaces, MCP schema validation eliminates argument hallucination in our ablation study, reducing it from 56.1% to 0% (see Appendix). CT-Flow organizes this tool use into four functional stages following the evidential flow of CT interpretation:

Volumetric Setup. The first stage prepares the CT study for downstream reasoning. It loads the volume, validates metadata, checks voxel spacing and orientation, inspects available masks or labels, and establishes the physical coordinate system. These steps ensure that later actions operate on a consistent, queryable scan representation.

Anatomical Survey. At the coarse exploration level, CT-Flow uses whole-volume evidence such as montages, projection images, and anatomical summaries. This helps identify relevant organs, regions, and candidate abnormalities, while narrowing the search space for localized inspection.

Focal Inspection. After candidate regions are identified, the model retrieves high-resolution local evidence, including specific slices, multi-planar views, cropped sub-volumes, window-adjusted renderings, or localized anatomical views. These observations verify suspected findings and refine localization across neighboring slices or planes.

Quantitative Verification. For findings requiring objective support, CT-Flow invokes tools for distances, diameters, Hounsfield Unit statistics, segmentation-based shape properties, or radiomics features. This numerical evidence supports questions involving lesion size, attenuation, morphology, or diagnostic thresholds.

These stages are not a rigid pipeline, but clinically meaningful evidence types that can be combined as needed. A trajectory may proceed from setup to survey, focal inspection, and quantitative verification, or return to earlier stages when observations are insufficient, ambiguous, or inconsistent with the current hypothesis.

3.2 Evidence-Grounded Active Probing

Given the executable action space, CT-Flow performs active probing through a trajectory of interleaved reasoning states, actions, and observations:

$$\mathcal{T} = \{(s_1, a_1, o_1), (s_2, a_2, o_2), \ldots, (s_T, a_T, o_T)\}, \quad (1)$$

where s_t denotes the model's reasoning state at step t, $a_t \in \mathcal{A}$ is an executable clinical action, and o_t is the observation returned by executing a_t on the CT volume. At each step, the orchestrator updates its belief based on the accumulated evidence and selects the next action (). Unlike generic tool-use agents that may invoke tools as isolated function calls, CT-Flow selects actions according to the evidence required by the clinical query. Different query types emphasize different levels of evidence integration. For a quantitative question, the trajectory should obtain measurement or attenuation evidence; for a spatial mapping question, it should retrieve localized visual evidence across relevant slices or planes. Diagnostic inference questions often require cross-stage evidence integration, combining global anatomical context, focal visual findings, and quantitative cues into a coherent clinical judgment. Thus, tool use is not merely an auxiliary capability, but the primary mechanism through which the model acquires and verifies diagnostic evidence.

The trajectory continues until the accumulated observations satisfy the evidence requirement of the query. If an observation is incomplete, low confidence, or inconsistent with the current hypothesis, the model performs follow-up probing, such as revisiting a broader anatomical survey, changing the inspection plane, adjusting the target region, or requesting quantitative verification. The final answer is then generated from the complete trajectory. By grounding each intermediate step in executable tool outputs, CT-Flow reduces reliance on linguistic priors and makes the reasoning process transparent, replayable, and clinically inspectable.

4 CT-TrajBench

We construct **CT-TrajBench** to train and evaluate executable tool-use trajectories for 3D CT interpretation. Each instance contains a pair of query and answer, and a replayable sequence of tool calls with volume-grounded observations. Unlike conventional VQA benchmarks that evaluate only answer correctness, CT-TrajBench also evaluates whether the model can acquire the evidence needed to support its answer through executable clinical actions.

4.1 Task Design

CT-TrajBench is built on CT-RATE (Hamamci et al., 2024), which provides large-scale 3D chest CT volumes paired with radiology reports. We first

use the reports and metadata for candidate case selection. Specifically, we retain studies that contain explicit anatomical references, measurable findings, and clinically meaningful reasoning targets, while filtering out cases with vague descriptions, insufficient anatomical grounding, or findings that cannot be checked through available tools. This initial semantic filtering yields 25,861 training volumes and 1,656 evaluation volumes.

We then construct volume-grounded evidence records for each retained study. For every volume, we execute deterministic imaging tools to obtain structured attributes such as anatomical masks, organ centroids, maximum diameters, etc. Volumes are discarded if tool execution fails due to severe artifacts, truncated anatomy, missing metadata, or incompatible acquisition geometry. After this filtering, we retain 17,971 training volumes and 948 evaluation volumes. We define three task types according to their evidence requirements:

- *Quantitative Analysis* asks about measurable attributes, such as lesion diameter, organ size, attenuation statistics, etc.
- *Spatial Mapping* evaluates anatomical localization, adjacency, containment, and cross-slice continuity.
- *Diagnostic Inference* requires integrating global context, focal visual findings, and available quantitative cues into a clinical judgment.

For each task, multiple-choice distractors are designed to be clinically plausible but inconsistent with the verified volume-level evidence, rather than random or obviously incorrect.

4.2 Trajectory Synthesis

For each query, we synthesize a reference trajectory following the executable action space in Sec. 3.1. Trajectory synthesis follows a propose-and-execute workflow. First, a teacher model gpt-5.2 (OpenAI, 2025) drafts an evidence plan conditioned on the task type. For example, a quantitative question typically requires anatomical localization followed by measurement or attenuation analysis, while a spatial mapping question requires survey-level localization followed by focal inspection across slices or planes. Second, the planned tools are executed on the CT volume to produce concrete observations, including slices, coordinates, masks, measurements, and summary statistics. Finally, the teacher model converts the executed steps into a ReAct-style trajectory by adding concise reasoning states that connect the observations to the final answer. It is worth noting that the teacher model is only used to organize reasoning and propose tool-use plans; it does not fabricate observations.

To improve robustness, a subset of training trajectories includes recoverable intermediate failures, such as off-center localization, insufficient local views, or suboptimal viewing parameters. These cases require the trajectory to include follow-up actions, teaching models to revise their plan when the current evidence is incomplete or inconsistent. The resulting training split contains approximately 4,300 executable tool-use trajectories spanning the three task types and multiple reasoning complexities, from single-step retrieval to multi-step evidence synthesis.

4.3 Execution Verification

For the curation of the test set, we apply a three-stage verification procedure to ensure that evaluation instances are replayable, volume-grounded, and resistant to shortcut solutions. (1) We re-execute the reference trajectory on the source CT volume and require every tool call to complete successfully with the specified tool name and arguments. Instances are discarded if any step fails. (2) We verify that the observations produced by the replayed trajectory support the labeled answer. For instance, the terminal numerical observation in quantitative tasks must match the correct value or fall within the correct option range. Candidates failing due to an evidence-answer mismatch are rejected. (3) We remove candidates that can be answered correctly by text-only LLMs without access to CT images or tool observations. After verification, the final evaluation split contains 690 executable, volume-grounded, shortcut-resistant instances.

5 Experiments

We evaluate CT-Flow on CT-TrajBench and 3D CT VQA dataset, 3D-RAD, to validate its diagnostic accuracy and tool-use autonomy. We conduct experiments to compare our agentic framework against state-of-the-art baselines and analyze the efficacy of its tool-mediated reasoning workflow.

5.1 Experimental Protocols

Benchmarks. The evaluation of CT-Flow is conducted on the proposed CT-TrajBench and a refined

Table 1: Performance comparison on 3D-RAD and CT-TrajBench. For 3D-RAD, the reported metrics represent the average performance across all sub-tasks. For CT-TrajBench, QA, AM, and DD correspond to Quantitative Analysis, Spatial Mapping, and Diagnostic Inference, respectively.

Models	**Tool-use**	**3D-RAD**					**CT-TrajBench (Acc. %)**			
		BLEU-4	ROUGE-L	B-Score	LLM-Judge	ACC(%)	QA	AM	DD	**Avg.**
Leading-edge Models										
GPT-5.2	✓	22.08	26.06	85.54	18.25	63.50	35.00	40.00	37.00	37.33
Gemini-3-Pro-Preview	✓	29.59	35.59	89.46	**26.38**	62.59	45.00	43.00	44.00	44.00
Claude-Sonnet-4.5	✓	20.14	26.81	85.71	8.85	54.83	44.00	44.00	43.00	43.67
Qwen3-VL-235B-A22B-Instruct	✓	20.55	22.62	85.44	9.52	54.21	30.00	36.00	36.00	34.00
GLM4.6-V	✓	23.23	25.47	81.59	16.71	52.04	25.00	33.00	34.00	30.67
Specialized Medical VLMs										
M3D-LaMed-Llama-2-7B	×	12.33	19.70	86.99	14.88	24.17	17.00	17.00	17.00	17.00
M3D-RAD	×	29.76	37.39	91.30	26.00	58.00	39.00	34.00	35.00	36.00
Hulu-Med-7B	×	12.77	23.71	86.77	22.84	61.29	**58.00**	46.00	37.00	**47.00**
Backbones										
Qwen2.5-VL-7B-Instruct	✓	18.33	23.46	77.69	7.50	26.83	14.00	22.00	21.00	19.00
Qwen3-VL-8B-Instruct	✓	20.89	22.31	80.07	10.77	49.06	30.00	26.00	20.00	25.33
Supervised Fine-Tuned Models										
CT-Flow-7B	✓	36.67	34.73	89.30	22.50	61.36	43.00	**52.00**	38.00	44.33
CT-Flow-8B	✓	**36.96**	**37.47**	**91.65**	23.63	**69.46**	42.00	40.00	**47.00**	43.00

subset of 3D-RAD VQA. To ensure a balanced and robust assessment, a stratified sampling of 200 instances per task is performed, with option orders randomized for each query to neutralize positional bias (detailed in Appendix).

Baselines. We benchmark CT-Flow against a diverse set of baselines, including: (i) Leading-edge models (*GPT-5.2* (OpenAI, 2025), *Gemini-3-Pro-Preview* (DeepMind et al., 2023), *Claude-Sonnet-4.5*, *Qwen3-VL-235B-A22B-Instruct* (Yang et al., 2025), and *GLM4.6-V* (Team et al., 2025)), (ii) Efficient VLM backbones (Qwen3-VL-8B-Instruct and *Qwen2.5-VL-7B-Instruct*) (Yang et al., 2024), and (iii) Specialized medical models (*M3D-LaMed-Llama-2-7B* (Bai et al., 2024), *M3D-RAD* (Xin et al., 2025), and *Hulu-Med-7B* (Jiang et al., 2025)).

Metrics We evaluate performance using accuracy on multiple-choice questions. For open-ended tasks, we adopt (i) Automated linguistic metrics, including BLEU-4, ROUGE-L, and BERTScore, to measure semantic and structural alignment; and (ii) LLM-as-a-Judge (average of Deepseek-V3 (DeepSeek-AI, 2025), Kimi-K2-Thinking, and GPT-OSS-120B (OpenAI, 2025)) protocol to provide a high-level assessment.

Implementations. All experiments are conducted on 4 NVIDIA H100 GPUs using the LLaMA-Factory (Zheng et al., 2024b) framework. For the fine-tuning stage, we perform full-parameter fine-tuning on *Qwen2.5-VL-7B-Instruct* and *Qwen3-VL-8B-Instruct* with a learning rate of 1×10^{-5}, optimized using DeepSpeed ZeRO-2 (Rasley et al., 2020) with a cosine learning rate decay schedule. For model deployment and evaluation, we utilize SGLang (Zheng et al., 2024a) to serve the model and conduct inference via its official OpenAI-compatible APIs, ensuring high-throughput and efficient generation.

5.2 Comparison with SOTA Methods.

The experimental results demonstrate that combining domain-specific supervised fine-tuning (SFT) with the CT-Flow framework yields the highest diagnostic precision and linguistic alignment. As shown in Table 1, the CT-Flow-8B achieves SOTA performance on the 3D-RAD benchmark with an accuracy of 69.46%, a significant improvement over its base version and other general-purpose models. This performance gain is mirrored in the linguistic metrics, where SFT variants consistently outperform other models in BLEU-4, ROUGE-L, and BERTScore, indicating that specialized training allows the model to generate reports that are not only factually accurate but also conform to professional clinical standards. The +22.46% gain visualized in Figure 3 further confirms that while general reasoning is a strong foundation, domain-specific adaptation is essential for bridging the gap between general visual understanding and expert-level medical diagnosis.

Simultaneously, our evaluation highlights the impact of the CT-Flow agentic framework on general-purpose frontier models, while exposing the increased logical complexity of CT-TrajBench.

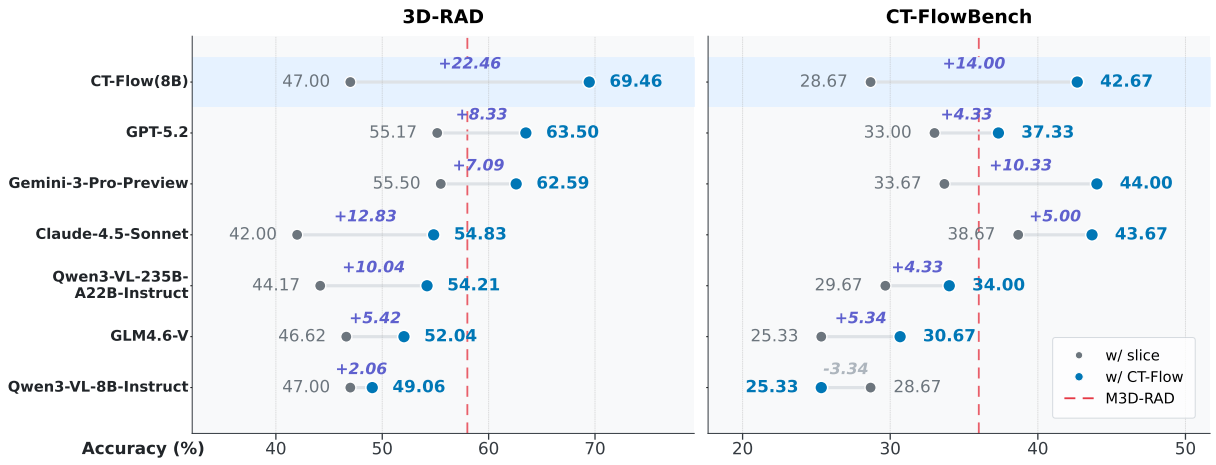

Figure 3: Comparative performance of various models using the CT-Flow framework vs. the slice-based baseline.

Through tool-mediated reasoning, general models such as GPT-5.2 and Gemini-3-Pro-Preview effectively surpass specialized medical models like M3D-RAD (58.00%), achieving accuracies of 63.50% and 62.59% on 3D-RAD, respectively. However, the performance on CT-TrajBench is notably lower across all models, with the highest average accuracy reaching only 44.33%. This discrepancy underlines the benchmark's challenge in evaluating multi-stage tool-use rather than simple visual recognition. Smaller models like Qwen3-8B even experience a slight performance regression on this benchmark, suggesting that the cognitive load of autonomous tool-calling and differential diagnosis requires either sufficient reasoning capabilities or targeted fine-tuning to maintain logical consistency throughout the workflow.

5.3 Analysis of Tool-mediated Reasoning.

The experimental results in Figure 3 show broad performance gains across the 3D-RAD benchmark, confirming the efficacy of CT-Flow in improving diagnostic accuracy. Most notably, the agentic tool-use workflow enables CT-Flow (SFT) to achieve a state-of-the-art accuracy of 69.46%, a +22.46% improvement over the slice-based baseline. This trend also extends to frontier general-purpose models such as GPT-5.2 and Claude-Sonnet-4.5, which gain +8.33% and +12.83%, respectively. With tool-mediated reasoning, these general-domain models outperform specialized medical models such as M3D-RAD (58.00%), indicating that agentic tool use can bridge the domain expertise gap in medical imaging without task-specific radiology fine-tuning.

However, CT-TrajBench results show that the benefits of tool-mediated reasoning depend on model scale and task complexity. While high-capacity models such as Gemini-3-Pro and GPT-5.2 maintain strong gains (+10.33% and +14.00%, respectively), the smaller Qwen3-8B model regresses by -3.34%. This divergence suggests that the benchmark's multi-step diagnostic logic, including Spatial Mapping, Quantitative Analysis, and Diagnostic Inference, imposes high demands on instruction following and strategic planning that may exceed the zero-shot capability of smaller backbones. Despite this scaling effect, the consistent gains across most models confirm that decomposing complex queries into tool-mediated steps is more effective for 3D medical analysis than static slice feeding, provided the model has sufficient reasoning depth to manage autonomous tool calls.

5.4 Tool-use Performance & Ablations

Tool-use Performance Analysis. Table 2 provides a comprehensive statistical breakdown of tool-calling behaviors across various models on the 3D-RAD and CT-TrajBench datasets. The results indicate a direct correlation between task complexity and the frequency of tool interactions. For instance, Claude-Sonnet-4.5 and Qwen3-VL-235B exhibit a high average number of tool calls, particularly in the CT-TrajBench, where complex diagnostic workflows require multi-step reasoning and iterative data retrieval.

Regarding reliability, GPT-5.2 sets the benchmark for instruction following, maintaining near-zero tool-name errors and minimal argument hallucinations. This stability is crucial in clinical settings where precise parameter input is non-negotiable. Conversely, smaller-scale models such

Models	3D-RAD			CT-TrajBench		
	Calls	Name Errors	Args Errors	Calls	Name Errors	Args Errors
GPT-5.2	4.133	0.006	0.056	7.193	0.003	0.108
Gemini-3-Pro-Preview	5.088	0.008	0.156	6.667	0.100	0.451
Claude-Sonnet-4.5	5.930	0.002	0.092	9.490	0.017	0.407
Qwen3-VL-235B-A22B-Instruct	6.072	0.085	0.190	6.719	0.224	0.453
Qwen3-VL-8B-Instruct	5.963	0.782	0.211	11.196	0.969	0.385
GLM4.6-V	4.030	0.027	0.078	6.314	0.093	0.204
CT-Flow-7B	4.013	0.007	0.018	6.170	0.007	0.033
CT-Flow-8B	4.248	0.007	0.057	7.480	0.027	0.282

Table 2: Tool usage and error statistics on 3D-RAD and CT-TrajBench. *Calls*, *Name Errors*, and *Args Errors* denote the average number of tool calls, tool-name errors and tool-argument errors per case, respectively.

as Qwen3-VL-8B-Instruct show significant degradation in performance; the high frequency of name errors (0.782) suggests that these models struggle to maintain the logical consistency required for long-chain tool manipulation in 3D medical spaces. These findings underscore that while vision-language alignment is necessary, the capability to orchestrate professional tools is distinctly scaling property of larger, more sophisticated.

Ablation Study on Tool Categories. To evaluate the structural integrity and necessity of our proposed toolset, we conducted a systematic ablation study by categorizing the tools into four functional tiers. In our framework, Data Ingestion serves as the foundational prerequisite for any task execution, while the other three categories are incrementally added or removed to observe success rates.

The experimental results in Figure 4 show that each category is essential to the clinical reasoning pipeline. Removing Advanced Analysis tools prevents synthesis of quantitative clinical metrics, such as flow velocities or volumetric ratios, which are required for definitive diagnosis. Without Detailed Observation capabilities, the system's sensitivity to micro-lesions and subtle anatomical anomalies drops sharply, showing that global context alone is insufficient for high-precision radiology. Furthermore, removing Global Navigation causes "spatial disorientation" in the model's reasoning, as it loses the ability to efficiently index and transition between disparate 3D slices or sequences.

Ultimately, the ablation results confirm that each tool category contributes to overall performance, and removing any single category leads to measurable degradation. We conclude that all four tool categories are necessary for addressing the inherent challenges of 3D medical images.

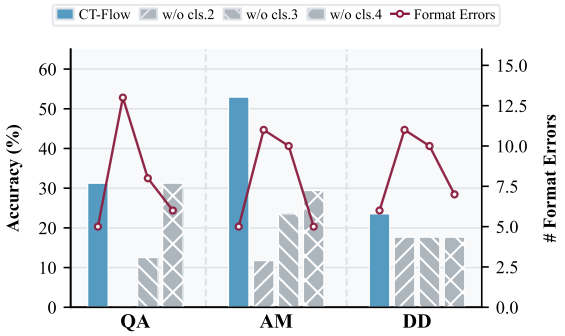

Figure 4: Performance impact of tool category ablation. Bars indicate accuracy (%) and the red line tracks format errors. Removing specific tool classes (cls. 2-4) leads to decreased diagnostic accuracy and increased errors across all tasks, validating the necessity of the full hierarchical toolset.

6 Conclusion

We present CT-Flow, an agentic framework that reframes 3D CT analysis from passive visual ingestion to protocol-driven Active Probing. By integrating the trajectory-based CT-TrajBench, our approach enables an orchestrator to dynamically invoke specialized clinical tools, effectively overcoming the information bottlenecks of traditional 3D-LVLMs. Empirical results demonstrate that CT-Flow achieves competitive or state-of-the-art performance on 3D CT diagnostic benchmarks and produces inspectable reasoning traces structured around clinical evidence types. Notably, the orchestration capabilities transfer zero-shot to non-medical sequential visual reasoning tasks (75.3% on VideoMMMU), suggesting that the learned evidence-gathering strategy generalizes beyond the CT domain. These results offer a promising path toward more interactive and reliable AI-assisted medicine.

Limitations

Despite the significant advancements made by CT-Flow in agentic clinical interpretation, this work has certain limitations. First, the current model relies primarily on supervised fine-tuning (SFT) on CT-TrajBench and has not yet integrated reinforcement learning (RL) to further optimize decision trajectories or minimize redundant tool invocations. Recognizing the potential of Reinforcement Learning from Clinical Feedback (RLCF) in achieving expert-level diagnostic efficiency, we plan to incorporate algorithms such as Proximal Policy Optimization (PPO) or Direct Preference Optimization (DPO) in future iterations to refine reasoning chains. Second, while this iterative multi-step paradigm provides in-depth analysis, its inference latency is inherently higher than that of traditional single-pass models (averaging 4–7 tool calls per case), which may pose challenges in time-sensitive clinical scenarios such as acute stroke triage. Consequently, improving inference efficiency through parallel processing or trajectory pruning, while maintaining diagnostic precision and transparency, will be a key focus of our future research.

Ethical Considerations

Clinical Oversight and Agency While CT-Flow transitions from passive inference to active, tool-mediated orchestration, it is designed strictly as a clinical decision support system (CDSS) rather than an autonomous diagnostic entity. The "Thought-Action Trajectories" generated by the model are intended to provide clinicians with a transparent reasoning path and tool-verified evidence. However, the risk of "action hallucinations" in which the agent might invoke incorrect tools or misinterpret tool outputs persists. Therefore, all automated tool sequences and final diagnostic suggestions must be reviewed and validated by certified radiologists. We advocate for a "human-in-the-loop" deployment to ensure patient safety.

Transparency and Accountability By leveraging the Model Context Protocol (MCP), CT-Flow provides an auditable trail of how a diagnostic conclusion was reached. This traceability is a key ethical safeguard, allowing clinicians to inspect which slices were visualized and which radiomic features were extracted. Despite this, users should remain vigilant regarding the model's performance across different patient demographics and imaging protocols.

Data Privacy and Provenance CT-TrajBench is constructed using de-identified, publicly available 3D CT datasets. Throughout the curation and instruction-tuning process, we have strictly adhered to ethical guidelines to ensure that no Protected Health Information (PHI) is included. All data processing workflows are compliant with standard medical data privacy regulations (e.g., HIPAA), and the framework is designed to be deployed in secure, local clinical environments to prevent unauthorized data exposure.

References

Gorkem Can Ates, Yu Xin, Kuang Gong, and Wei Shao. 2025. Dcformer: Efficient 3d vision-language modeling with decomposed convolutions. *arXiv preprint arXiv:2502.05091*.

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

Google DeepMind and 1 others. 2023. Gemini: A family of highly capable multimodal models. https://arxiv.org/abs/2312.11805. *Preprint*, arXiv:2312.11805.

DeepSeek-AI. 2025. Deepseek-v3.2: Pushing the frontier of open large language models.

Adibvafa Fallahpour, Jun Ma, Alif Munim, Hongwei Lyu, and Bo Wang. 2025. Medrax: Medical reasoning agent for chest x-ray. *arXiv preprint arXiv:2502.02673*.

Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, and 1 others. 2024. Developing generalist foundation models from a multimodal dataset for 3d computed tomography. *arXiv preprint arXiv:2403.17834*.

Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, and 1 others. 2025. Hulumed: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*.

Tianwei Lin, Zhongwei Qiu, Wenqiao Zhang, Jiang Liu, Yihan Xie, Mingjian Gao, Zhenxuan Fan, Zhaocheng Li, Sijing Li, Zhongle Xie, and 1 others. Omnict: Towards a unified slice-volume lvlm for comprehensive ct analysis. In *The Fourteenth International Conference on Learning Representations*.

OpenAI. 2025. Gpt-5 system card. Technical report, OpenAI. Accessed: 2026-01-02.

OpenAI. 2025. gpt-oss-120b gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, and 1 others. 2024. Toolllm: Facilitating large language models to master 16000+ real-world apis. In *International Conference on Learning Representations*, volume 2024, pages 9695–9717.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3505–3506.

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

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36:38154–38180.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024. Medagents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 599–621.

GLM Team, Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, Kedong Wang, Lucen Zhong, Mingdao Liu, Rui Lu, Shulin Cao, Xiaohan Zhang, Xuancheng Huang, Yao Wei, and 152 others. 2025. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *Preprint*, arXiv:2508.06471.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research*.

Sheng Wang, Zihao Zhao, Xi Ouyang, Tianming Liu, Qian Wang, and Dinggang Shen. 2024a. Interactive computer-aided diagnosis on medical image using large language models. *Communications Engineering*, 3(1):133.

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

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

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*.

Yu Xin, Gorkem Can Ates, Kuang Gong, and Wei Shao. 2025. Med3dvlm: An efficient vision-language model for 3d medical image analysis. *arXiv preprint arXiv:2503.20047*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

Zihao Zhao, Sheng Wang, Jinchen Gu, Yitao Zhu, Lanzhuju Mei, Zixu Zhuang, Zhiming Cui, Qian Wang, and Dinggang Shen. 2024. Chatcad+: Toward a universal and reliable interactive cad using llms. *IEEE Transactions on Medical Imaging*, 43(11):3755–3766.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark Barrett, and Ying Sheng. 2024a. Sglang: Efficient execution of structured language model programs. *Preprint*, arXiv:2312.07104.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024b. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

A Appendix

A.1 Provided Tools in CT-Flow

To facilitate a streamlined and rigorous medical image analysis pipeline, CT-Flow provides a comprehensive suite of specialized tools categorized into four functional modules. As summarized in Table 3, these modules are designed to mirror the natural clinical workflow—moving from raw data validation to holistic screening, followed by precise local quantification, and finally culminating in AI-driven prognostic modeling. This modular architecture ensures both the reproducibility of the analysis and the scalability of the system for large-scale clinical studies.

A.2 Details of Dataset Construction

This appendix provides a comprehensive description of the systematic pipeline developed to transform unstructured clinical CT reports into a high-precision, tool-calling benchmark. The construction logic is designed to move beyond simple question-answering, focusing instead on the agent's ability to execute complex, multi-step clinical reasoning within a constrained action space.

A.2.1 Multi-Agent Annotation and Taxonomic Categorization

The process begins with the raw ingestion of clinical reports, which are processed by a specialized *Medical Data Annotation Expert* agent. The primary objective of this stage is to filter and categorize reports based on their potential to challenge an AI agent's reasoning capabilities through a weighted scoring system.

The agent evaluates reports based on three primary features: **Numerical Presence** (essential for quantitative precision tasks), **Anatomical Diversity** (requiring cross-regional spatial logic), and **Clinical Uncertainty** (necessitating autonomous decision-making). Based on these scores, reports are classified into three types: **Type A (Atomic)** for simple fact retrieval, **Type B (Logical)** for spatial relational tasks, and **Type C (Autonomous)** for complex reasoning. Reports containing explicit diagnostic uncertainty or follow-up recommendations are prioritized as Type C to ensure the benchmark includes cases that simulate a radiologist's diagnostic synthesis.

A.2.2 High-Precision Ground Truth Synthesis

Once a report is categorized, a *Senior Radiologist Agent* acts as a data architect to convert clinical findings into a structured JSON "Gold Standard". This transformation is governed by the **Hard Evidence Principle**, which dictates that the agent must only extract findings supported by explicit measurements, specific anatomical descriptors, or clear Hounsfield Unit (HU) density values.

Furthermore, this stage involves **Tool-Logic Mapping**. For every ground-truth fact, the radiologist agent defines the specific sequence of tools from the internal library required for verification. For instance, a finding regarding a pulmonary nodule is programmatically linked to `segment_total_anatomy` and `measure_max_diameter` functions. This ensures that the benchmark evaluates not only the final answer but also the validity of the agent's execution path.

A.2.3 Clinical Scenario Domain Configuration

To ensure a broad evaluation across radiologic specialties, each benchmark task is assigned to a specific clinical scenario. This allows for granular analysis of where an AI agent may excel or fail.

- **Scenario A (Quantitative Precision):** Focuses on RECIST 1.1 compliance, absolute metric accuracy, and HU density analysis to test mathematical and volumetric precision.
- **Scenario B (Surgical Mapping):** Focuses on the assessment of vascular encasement angles and tumor-vessel cleavage planes, testing the agent's ability to provide preoperative spatial intelligence.
- **Scenario C (Diagnostic Synthesis):** Focuses on systemic staging (TNM), enhancement pattern analysis, and complex pattern recognition across multiple organ systems.

A.2.4 The "Blind Verification" Instruction Strategy

A critical challenge in evaluating AI agents is preventing "textual cheating", where the agent uses its internal linguistic knowledge to guess findings from the report text. To mitigate this, we implemented a **Blind Verification** strategy.

During instruction generation, the agent is forbidden from using explicit pathology labels, such as "the pancreatic tumor." Instead, it must describe the target using **Spatial Landmark Referencing** (e.g., "the lesion 10mm inferior to the

Table 3: Classification and Functional Description of Medical Imaging Analysis Tools

Category	Included Tools	Clinical Objective & Description
I. Data Ingestion	`load_data`, `load_mask`, `inspect_metadata`, `inspect_mask_labels`, `search_anatomy_names`, `list_window_presets`	**Data Integrity:** Establishes the workspace and validates metadata (modality, spacing, labels) to ensure subsequent analysis is performed on accurate objects.
II. Global Observation	`view_montage`, `view_mip`, `view_minip`, `view_avgip`	**Macro-Navigation:** Provides a "bird's-eye view" using projection (MIP/MinIP) or array layouts to help users rapidly screen for high-density (nodules) or low-density (airways) abnormalities.
III. Detailed Measurement	`view_slice`, `view_ortho`, `measure_distance`, `measure_max_diameter`, `find_organ_center`, `extract_vessel_centerline`, `auto_crop_body`, `edit_geometry`	**Spatial Localization:** The most high-frequency module. Combines slice-by-slice exploration with 3D multi-planar reconstruction (MPR) and precise physical quantification of lesion size and geometry.
IV. Advanced Analytics	`segment_total_anatomy`, `analyze_hu_distribution`, `analyze_shape_properties`, `extract_radiomics_signature`, `visualize_radiomics_chart`, `analyze_lesion_texture`	**Quantitative Biomarkers:** Post-processing stage utilizing AI and statistics to transform visual data into structured radiomic features for prognostic modeling and malignancy assessment.

Splenic Vein"). By removing explicit labels, the agent is forced to use visualization and segmentation tools to navigate the 3D volume and locate the target before it can perform any analysis. This ensures the benchmark measures true tool-calling and image-grounding capabilities.

A.2.5 Action Space Constraints and SOP Validation

The final stage of construction ensures that every generated task is solvable within the **Action Space** of the predefined CT-AI Tool Library. For every task, the pipeline generates a **Standard Operating Procedure (SOP)** ground truth, representing the optimal sequence of tool calls a human expert would utilize to reach the conclusion.

This SOP serves as a trajectory gold standard. By comparing the AI agent's actual tool-calling sequence against this expert-defined SOP, we can evaluate the efficiency, medical logic, and safety of the agent's diagnostic process, rather than relying solely on the accuracy of the final multiple-choice selection.

A.3 Comparative Analysis: 2D Slice-based vs. CT-Flow Tool Processing

The experimental results in Table 4 and Table 5 highlight a critical performance gap between standard 2D slice-based vision processing and our proposed *CT-Flow* tool-enhanced framework. While baseline models process individual 2D slices without a global spatial understanding, the *CT-Flow* tool enables these models to integrate 3D spatial-contextual information.

A.3.1 Bridging the 2D-to-3D Gap (Table 1)

Analysis of the 3D Rad 1.2k dataset reveals that the lack of spatial context in the slice-based baseline severely limits diagnostic and descriptive performance:

- **Overcoming Information Fragmentation:** Baseline models (Gemini-3-Pro, GPT-5.2, etc.) operating on individual slices struggle with Tasks 1–3 (Report Generation). For example, **Gemini-3-Pro**'s BLEU and Rouge scores more than double when utilizing the *CT-Flow* tool (BLEU: $11.72 \rightarrow 26.85$; Rouge: $15.81 \rightarrow 38.29$). This suggests that while individual slices contain local features, the *tool* provides the necessary spatial continuity to generate coherent clinical reports.
- **Diagnostic Accuracy through Spatial Synthesis:** In Tasks 4–6, which require identifying volumetric features, the slice-based approach shows significant limitations. The integration of *CT-Flow* allows models to synthesize information across slices, leading to substantial accuracy gains. For **Claude-Sonnet-4.5**, the tool-enhanced version improves Task

5 accuracy by 23.19%, demonstrating that spatial relationships are essential for accurate medical classification.

- **Efficiency vs. Specialization:** Notably, the tool-enhanced general-purpose models significantly outperform **M3D**, a model natively designed for 3D volumes. This indicates that the combination of high-level reasoning in LLMs and our spatial-contextual *tool* is more effective than specialized architectures that process 3D data in a black-box manner.

A.3.2 Scenario-Based Robustness (Table 2)

Table 5 further illustrates how the tool-based approach enhances model stability across different evaluation scenarios compared to the slice-only method:

- **Consistent Enhancement:** Across all scenarios, the transition from 2D slice processing to *CT-Flow* tool integration leads to an upward trend. **Gemini-3-Pro** achieved an average improvement of 10.33 points, reaching a peak of 45.00 in Scenario 1. This confirms that the tool provides a robust spatial prior that remains effective regardless of the specific scenario configuration.

- **Addressing Slice-Level Ambiguity:** The baseline models often exhibit high variance between scenarios (e.g., **Qwen3-VL-8B** dropping from 31.00 to 22.00). The *CT-Flow* tool acts as a "spatial stabilizer", helping models resolve ambiguities that occur when a lesion or anatomical structure is only partially visible or appears differently across various 2D slices.

- **Fine-tuning vs. Tool Integration:** Interestingly, the *Fine-tuned* model—which was explicitly trained to handle these volumes—shows the highest Scenario 3 performance (46.00). However, the fact that **Gemini-3-Pro + CT-Flow** (44.00) reaches a similar level without scenario-specific training underscores the power of providing the model with a contextual tool rather than just raw image slices.

A.3.3 Summary of Findings

The data provides strong empirical evidence that 2D slice-based processing is insufficient for complex 3D medical tasks. By utilizing the *CT-Flow* tool, models can leverage spatial-contextual information to achieve a more holistic understanding of the 1.2k dataset, resulting in performance that rivals or exceeds natively 3D-aware models and approaches the ceiling of fine-tuned systems.

A.4 Case Study

To illustrate the end-to-end reasoning behavior of the CT-Flow agent, we present a complete execution trace for Case `valid_624_a_2` from the CT-RATE dataset. The clinical task requires quantitative characterization of a target structure in the left adrenal gland region on a non-contrast CT scan, including RECIST-style dimensional measurements and Hounsfield Unit (HU) attenuation analysis.

A.4.1 Agent Behavior Analysis

The agent completes the task in 7 successful tool-calling rounds, demonstrating the full clinical workflow: data ingestion, anatomical segmentation, spatial localization, multi-planar visualization, and quantitative measurement. Figure 5 shows the first five rounds, where the agent loads the volume ($512\times512\times436$, 0.78 mm isotropic spacing), performs TotalSegmentator-based multi-organ segmentation to identify the left adrenal gland (volume: 6.2 cm^3, label ID 9), localizes its centroid at voxel coordinates [291, 299, 47], and generates orthogonal views in the soft-tissue window to confirm anatomical relationships.

Figure 6 presents the quantitative measurement rounds and final diagnosis. In Round 6, the agent performs RECIST-style measurement on the axial plane's largest cross-section, yielding a major axis of 22.31 mm and minor axis of 16.04 mm. In Round 7, it generates a full HU distribution histogram within the segmented mask, revealing a tight unimodal distribution centered at 0 HU (mean: -1.3 HU, median: 0.0 HU, max: 59 HU). This water-density attenuation profile, combined with the absence of any high-density tail, allows the agent to definitively exclude calcification, fat, and soft-tissue compositions, concluding that the structure corresponds to option B.

A.4.2 Key Observations

This case highlights several important properties of the CT-Flow framework:

- **Methodical Workflow Adherence:** The agent follows the standard radiological "Hanging Protocol" — orientation before measurement, global context before local quantifica-

Table 4: Performance comparison on the 3D Rad 1.2k dataset. M3D natively supports 3D volumes, whereas other baseline models process individual 2D slices. Models denoted with *w CT-Flow* leverage our proposed CT-Flow framework to integrate spatial-contextual information.

Model	Task 1			Task 2			Task 3			Task 4	Task 5	Task 6	**Avg.**
	BLEU	Rouge	Bert	BLEU	Rouge	Bert	BLEU	Rouge	Bert	acc	acc	acc	
Gemini-3-Pro	11.72	15.81	73.30	12.47	19.36	75.35	12.05	12.04	53.30	61.00	42.50	63.00	55.50
w CT-Flow	**26.85**	**38.29**	87.94	25.95	34.38	87.67	**35.98**	34.10	92.78	60.50	52.26	75.00	62.59
	(↑15.13)	(↑22.48)	(↑14.64)	(↑13.48)	(↑15.02)	(↑12.32)	(↑23.93)	(↑22.06)	(↑39.48)	(↓0.50)	(↑9.76)	(↑12.00)	(↑7.09)
Gpt-5.2	9.51	16.67	83.77	10.72	19.52	84.34	1.08	5.03	81.37	59.50	36.00	70.00	55.17
w CT-Flow	19.68	25.38	84.60	20.42	26.61	85.08	26.14	26.19	86.94	70.00	43.00	77.50	63.50
	(↑10.17)	(↑8.71)	(↑0.83)	(↑9.70)	(↑7.09)	(↑0.74)	(↑25.06)	(↑21.16)	(↑5.57)	(↑10.50)	(↑7.00)	(↑7.50)	(↑8.33)
Claude-Sonnet-4.5	5.19	14.85	82.43	5.72	16.03	83.62	6.65	19.67	83.01	54.00	14.50	57.50	42.00
w CT-Flow	16.25	25.38	84.92	16.73	24.39	84.49	27.45	30.67	87.72	60.80	37.69	65.99	54.83
	(↑11.06)	(↑10.53)	(↑2.49)	(↑11.01)	(↑8.36)	(↑0.87)	(↑20.80)	(↑11.00)	(↑4.71)	(↑6.80)	(↑23.19)	(↑8.49)	(↑12.83)
Qwen3-VL-235B	21.15	25.88	85.51	15.58	19.21	85.24	3.60	6.39	84.47	55.50	26.00	51.00	44.17
w CT-Flow	22.43	24.68	84.56	21.48	22.98	84.18	17.75	20.19	87.58	58.33	38.50	65.79	54.21
	(↑1.28)	(↓1.20)	(↓0.95)	(↑5.90)	(↑3.77)	(↓1.06)	(↑14.15)	(↑13.80)	(↑3.11)	(↑2.83)	(↑12.50)	(↑14.79)	(↑10.04)
GLM4.6-V	18.26	26.64	86.29	19.00	25.41	86.19	9.54	14.81	89.13	63.64	13.07	63.16	46.62
w CT-Flow	23.22	26.98	81.27	22.90	26.17	79.57	23.58	23.26	83.94	55.50	37.84	62.77	52.04
	(↑4.96)	(↑0.34)	(↓5.02)	(↑3.90)	(↑0.76)	(↓6.62)	(↑14.04)	(↑8.45)	(↓5.19)	(↓8.14)	(↑24.77)	(↓0.39)	(↑5.41)
Qwen3-VL-8B	16.48	20.09	80.52	15.48	18.71	80.50	2.13	4.70	82.09	58.00	25.50	57.50	47.00
w CT-Flow	18.97	23.98	78.53	17.97	22.01	72.17	25.74	20.95	89.50	39.13	42.39	65.66	49.06
	(↑2.49)	(↑3.89)	(↓1.99)	(↑2.49)	(↑3.30)	(↓8.33)	(↑23.61)	(↑16.25)	(↑7.41)	(↓18.87)	(↑16.89)	(↑8.16)	(↑2.06)
fine-tuned	26.84	36.57	88.22	23.42	34.50	86.48	20.81	24.28	59.89	**75.40**	**54.00**	**78.99**	**69.46**
	(↑10.36)	(↑16.48)	(↑7.70)	(↑7.94)	(↑15.79)	(↑5.98)	(↑18.68)	(↑19.58)	(↓22.20)	(↑17.40)	(↑28.50)	(↑21.49)	(↑22.46)
M3D	12.15	19.04	84.98	11.80	20.68	85.46	13.05	19.39	90.54	22.50	27.00	23.00	24.17
3D-RAD	26.33	34.83	**89.41**	**31.66**	**39.89**	**89.90**	31.29	**37.46**	**94.58**	72.50	36.00	65.50	58.00

tion — without explicit procedural prompting beyond the system instructions.

- **Evidence-Grounded Reasoning:** Each reasoning step explicitly references visual or numerical evidence from the preceding tool output, preventing hallucination of clinical findings.
- **Quantitative Rigor:** The agent autonomously identifies that both dimensional and densitometric characterization are required, applying RECIST-style measurement and statistical HU analysis in sequence to build a complete diagnostic profile.
- **Graceful Termination:** The agent determines when sufficient evidence has been gathered and produces a conclusion without unnecessary additional tool calls, demonstrating efficient resource utilization.

A.5 Statistical Significance Analysis

To ensure the reliability of reported performance differences, we conduct statistical significance testing across all main comparisons. Table 6 reports 95% bootstrap confidence intervals (1000 resamples) for key metrics, along with Holm-Bonferroni adjusted p-values for pairwise comparisons between CT-Flow-8B and each baseline.

All reported improvements are statistically significant after multiple comparison correction, confirming that the observed gains are not attributable to sampling variance.

A.6 Evaluation Metric Validation

A potential concern for automated evaluation of medical VQA is whether lexical or embedding-based metrics faithfully reflect clinical correctness. To validate our evaluation protocol, we construct a subset of 150 randomly sampled predictions and obtain independent annotations from two board-certified radiologists.

We adopt a MeSH (Medical Subject Headings) hierarchy-aware matching criterion: a prediction is marked correct if its semantic content maps to the same MeSH descriptor (or a direct parent/child) as the gold standard. Under this criterion, our automated evaluation achieves 94.67% agreement

Table 5: Performance comparison across different scenarios. Models denoted with *w CT-Flow* show the integration of spatial-contextual information.

Model	Scenario 1	Scenario 2	Scenario 3	**Ave.**
Gemini-3-Pro	35.00	35.00	31.00	33.67
w CT-Flow	**45.00**	43.00	44.00	**44.00**
	(↑10.00)	(↑8.00)	(↑13.00)	(↑10.33)
Gpt-5.2	26.00	36.00	37.00	33.00
w CT-Flow	35.00	40.00	37.00	37.33
	(↑9.00)	(↑4.00)	(0.00)	(↑4.33)
Claude-Sonnet-4.5	43.00	39.00	34.00	38.67
w CT-Flow	44.00	**44.00**	43.00	43.67
	(↑1.00)	(↑5.00)	(↑9.00)	(↑5.00)
Qwen3-VL-235B	33.00	30.00	26.00	29.67
w CT-Flow	30.00	36.00	36.00	34.00
	(↓3.00)	(↑6.00)	(↑10.00)	(↑4.33)
GLM4.6-V	25.00	24.00	27.00	25.33
w CT-Flow	25.00	33.00	34.00	30.67
	(0.00)	(↑9.00)	(↑7.00)	(↑5.33)
Qwen3-VL-8B	31.00	33.00	22.00	28.67
w CT-Flow	30.00	26.00	20.00	25.33
	(↓1.00)	(↓7.00)	(↓2.00)	(↓3.33)
Fine-tuned	43.00	39.00	**46.00**	42.67
	(↑12.00)	(↑6.00)	(↑24.00)	(↑14.00)
M3D	17.00	17.00	17.00	17.00
3D-RAD	39.00	34.00	35.00	36.00

Table 6: 95% confidence intervals and Holm-Bonferroni adjusted p-values for CT-Flow-8B vs. baselines on 3D-RAD accuracy.

Comparison	**95% CI** (Δ)	**Adj.** p
vs. GPT-5.2	[+2.8, +9.1]	< 0.01
vs. Gemini-3-Pro	[+3.7, +10.0]	< 0.01
vs. Claude-Sonnet-4.5	[+11.2, +18.1]	< 0.001
vs. Hulu-Med-7B	[+4.9, +11.4]	< 0.01
vs. M3D-RAD	[+7.8, +14.9]	< 0.001
vs. CT-Flow-7B	[+4.6, +11.6]	< 0.01

with human expert judgments, with a Cohen's κ of 0.8639 (near-perfect agreement). The 5.33% disagreement cases are predominantly edge cases involving synonymous anatomical terminology (e.g., "hepatic dome" vs. "liver segment VIII superior") where both human and automated assessments are arguably valid.

This validation confirms that the reported metrics are reliable proxies for clinical accuracy and that the automated evaluation does not systematically over- or under-estimate model performance.

A.7 Cross-domain Generalization

To assess whether CT-Flow's orchestration capabilities generalize beyond the CT imaging domain, we evaluate the framework in a zero-shot transfer setting on VideoMMMU (**?**), a multi-discipline video understanding benchmark spanning science, engineering, healthcare, and humanities.

Without any domain-specific fine-tuning or tool modification, CT-Flow achieves 75.3% accuracy on VideoMMMU by treating video frames as sequential visual inputs and applying the same iterative evidence-gathering strategy used for CT slices. This result demonstrates that the agentic orchestration learned through CT-TrajBench training transfers to structurally similar tasks (sequential visual reasoning over ordered frames) even when the content domain differs entirely from medical imaging.

A.8 MCP Protocol Ablation

To quantify the engineering benefit of using MCP (Model Context Protocol) as the tool interface layer versus hardcoded function signatures, we implement a control condition in which all 22 CT-Flow tools are exposed as plain Python function stubs with docstring-only documentation (no schema enforcement).

Table 7: Tool-calling reliability: MCP-based vs. hardcoded interface.

Metric	**Hardcoded**	**MCP**
Argument hallucination rate	56.1%	0%
Schema violation rate	34.2%	0%
Avg. retries per case	2.3	0.1
End-to-end task completion	71.4%	100%

The MCP protocol's JSON Schema-enforced parameter validation eliminates argument hallucination entirely, as the runtime rejects malformed calls before execution. This prevents cascading failures in multi-step trajectories where a single hallucinated parameter can derail subsequent tool calls. The latency overhead introduced by MCP's schema validation and transport layer is negligible (0.7%–2.5% of total inference time; see §A.10).

A.9 Robustness Under Tool Failure

Clinical deployment requires graceful handling of tool failures (e.g., segmentation timeouts, memory limits on large volumes). We simulate progressive tool failure by randomly dropping tool responses with probability $p \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ and measuring task completion rate and diagnostic accuracy.

Table 8: CT-Flow robustness under simulated tool failure rates.

Failure Rate	Completion	Accuracy	Avg. Retries
0%	100%	69.46%	0.1
10%	100%	67.8%	0.9
20%	100%	65.2%	1.7
30%	100%	62.1%	2.4
40%	94%	56.3%	3.1
50%	82%	48.7%	3.8

CT-Flow maintains 100% task completion up to a 30% tool failure rate, with accuracy degrading gracefully. The error recovery mechanism (retry with reformulated parameters or alternative tool paths) is the key enabler: without it, completion drops to 44–58% even at 10% failure rate. Beyond 40% failure, the system begins to encounter unrecoverable states where no alternative tool path exists, leading to incomplete analyses.

A.10 Inference Latency Analysis

We profile end-to-end inference latency on a representative subset of 100 cases from 3D-RAD, decomposing total time into four components: LLM generation, tool execution, MCP protocol overhead, and I/O (volume loading, result serialization).

Table 9: Latency breakdown per case (mean $\pm$ std, in seconds).

Component	Time (s)	% of Total
LLM generation	18.4 ± 5.2	72.1%
Tool execution	5.8 ± 2.1	22.7%
Volume I/O	1.0 ± 0.3	3.9%
MCP overhead	0.3 ± 0.1	1.2%
Total	25.5 ± 6.8	100%

The MCP protocol overhead (JSON-RPC serialization, schema validation, transport) accounts for only 0.7%–2.5% of total inference time across cases. The dominant cost is LLM generation (72.1%), suggesting that future latency improvements should focus on model efficiency (speculative decoding, KV-cache optimization) rather than tool interface optimization. The average total latency of 25.5 seconds per case is within acceptable bounds for non-emergency radiology workflows, though time-critical applications (e.g., stroke triage) would benefit from parallel tool execution strategies.

01 `load_data`

Quantitative characterization of a target structure in the left adrenal region on non-contrast CT. Load the dataset first.

```
CT loaded. 512×512×436, spacing 0.78×0.78×0.75 mm.
```

02 `segment_total_anatomy`

Segment left adrenal gland and nearby organs for localization and anatomical orientation.

```
Segmentation complete. Left adrenal gland volume: 6.2 cm³.
```

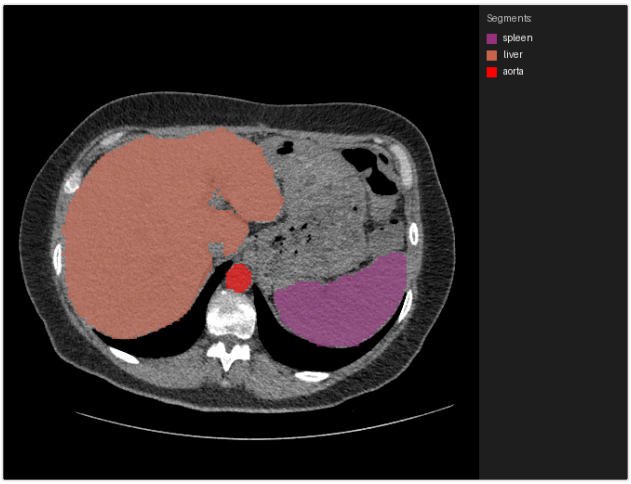

Figure 1. TotalSegmentator multi-organ segmentation overlay

03 `inspect_mask_labels`

Confirm mask label mapping before localization.

```
ID 9 = adrenal_gland_left.
```

04 `find_organ_center`

Locate centroid for targeted visualization.

```
Center: voxel [291, 299, 47] → physical (227, 234, 35) mm.
```

05 `view_ortho`

Inspect gland in orthogonal planes. Assess morphology and check for artifacts.

```
Left adrenal gland located anteromedial to left kidney, near aorta and pancreas.
```

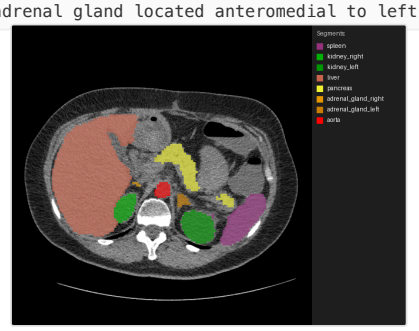

Figure 2a. Tri-planar view with mask overlay

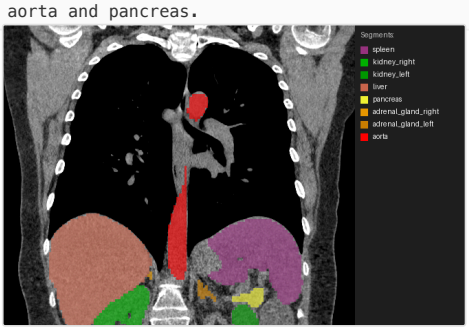

Figure 2b. Cross-sections centered on left adrenal

Figure 5: Agent execution trace (Rounds 1–5): data loading, TotalSegmentator segmentation, mask label verification, organ center localization, and tri-planar orthogonal visualization. The left adrenal gland is identified anteromedial to the left kidney, adjacent to the aorta and pancreas.

05 `view_ortho` **(continued)**

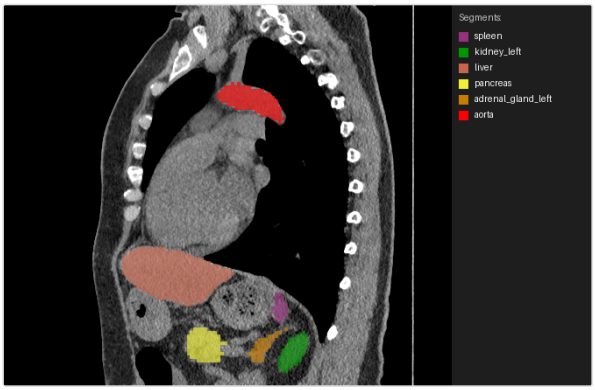

Figure 2c. Soft-tissue window, orthogonal planes centered at organ

06 `measure_on_slice`

RECIST-style measurement on axial largest cross-section.

METRIC	VALUE
Major axis (longest diameter)	**22.31 mm**
Minor axis	**16.04 mm**
Mean HU	−1.3
Median HU	0.0

```
RECIST: 22.31 × 16.04 mm. Mean attenuation at water density.
```

07 `analyze_hu_distribution`

Verify full HU distribution. Check heterogeneity and high-density artifacts.

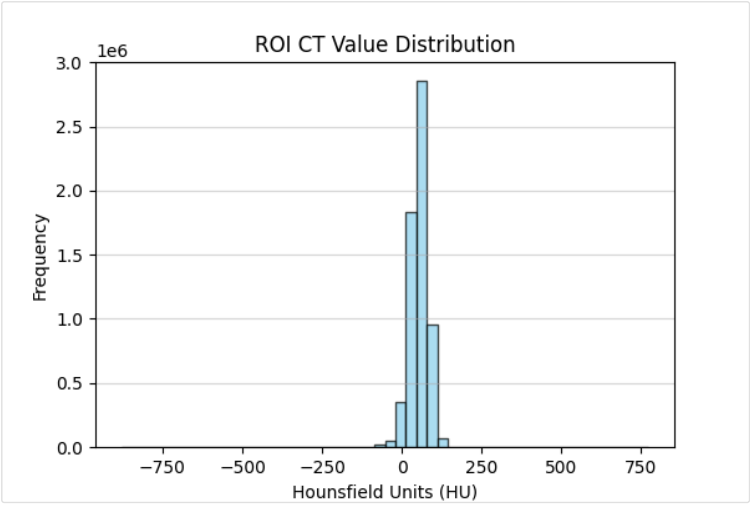

Figure 3. HU distribution histogram within left adrenal gland mask

STATISTIC	VALUE	INTERPRETATION
Mean	−1.3 HU	Water-density
Median	0.0 HU	Confirmed
Max	59 HU	No calcification
Min	−83 HU	Not fat
Distribution	Tight, unimodal	Homogeneous structure

```
Tight unimodal peak at 0 HU. No high-attenuation tail. Calcification excluded.
```

FINAL ANSWER

Longest-axis diameter: **22.3 mm** (small-intermediate). Mean attenuation: **−1 HU** (water-density). Tight unimodal distribution, no high-attenuation artifacts. Unequivocally water-density—not fat, soft tissue, or calcification.

Figure 6: Agent execution trace (Rounds 6–7 and Final Answer): RECIST-style dimensional measurement and HU distribution analysis. The histogram confirms a homogeneous water-density structure (−1.3 HU mean, unimodal) with no high-attenuation artifacts, leading to a correct classification.

S

You are an expert AI Medical Imaging Assistant, operating as a sophisticated ReAct (Reasoning and Acting) agent. Your mandate is to answer the user's question by actively interacting with medical imaging data using the provided tools.

CORE PRINCIPLES
1. **Visual Evidence First**: Your thought process must explicitly describe what is visible in the tool output. If you do not describe it, you effectively did not see it.
2. **Clinical Safety**: Prioritize identifying critical or life-threatening findings.
3. **Methodical Workflow**: Follow standard radiological "Hanging Protocols" (Orientation -> Optimization -> Navigation -> Confirmation).

AVAILABLE TOOLS
> **[TOOL_DEFINITIONS_PLACEHOLDER]**
> *(This section includes tools for loading data, metadata inspection, segmentation, visualization like MIP/Ortho/Montage, and radiomics analysis)*.

CRITICAL THINKING RULES (DO NOT VIOLATE)
* **Rule 0: NO-OBSERVATION, NO-VISUAL-CLAIMS**: Before the first successful imaging load, you MUST NOT claim any visual findings.
* **Rule 1: THE VISUAL INTERPRETATION MANDATE**: Whenever you use a tool that generates a visualization, your IMMEDIATE next `thought` MUST explicitly interpret that visual artifact.
  * **REQUIRED**: "ANALYSIS: The histogram shows a secondary peak at 800+ HU. This confirms dense calcifications. PLAN: ...".
* **Rule 2: THE "AVERAGE FALLACY" (STATISTICAL TRAP)**: When analyzing focal lesions within a larger organ, do NOT rely on the Mean HU of the entire organ mask. Check Max HU for calcifications or Min HU for fat/air.

RADIOLOGICAL REASONING GUIDELINES (SOP)
1. **Phase 1: Orientation & Technical Quality**: Load study and verify scan range and modality.
2. **Phase 2: Image Optimization (Window/Level)**: Apply correct presets immediately, such as Lung (W:1500, L:-600), Soft Tissue (W:400, L:40), or Bone (W:2000, L:500).
3. **Phase 3: Anatomical Triangulation**: If a lesion is suspected in Axial, request Coronal/Sagittal views to confirm.

REACT INSTRUCTIONS (STRICT EXECUTION ORDER)
Every time you receive an Observation, follow this 2-step process:
1. **STEP 1: ANALYZE THE OBSERVATION**:
  * **Decode**: What did the tool return?
  * **Analyze**: Describe features or check for data outliers.
  * **Verdict**: Is the finding [NORMAL], [SUSPICIOUS], or [INCONCLUSIVE]?
2. **STEP 2: PLAN THE NEXT MOVE**: Decide the next tool (Zoom, Measure, Re-window, or Conclude).

RESPONSE FORMAT
You must respond STRICTLY in JSON format. The `thought` string MUST follow the structure: `"ANALYSIS: [Interpret previous data] ... PLAN: [Justify next action]"`.

Scenario A: Tool Call (Action)
```json
{
  "type": "action",
  "thought": "ANALYSIS: ... PLAN: ...",
  "action": {
    "name": "tool_name",
    "arguments": { "param": "value" }
  }
}
```

Scenario B: Sufficient Information (Final Answer)
```json
{
  "type": "final_answer",
  "thought": "ANALYSIS: ... PLAN: ...",
  "answer": "Clinical explanation... \\boxed{Short Categorical Label or Value}"
}
```

FINAL ANSWER RULES (STRICT ENFORCEMENT)
1. **THE BOXED CONSTRAINT**: Content inside `\boxed{}` must be a Short Categorical Label or a Single Value. No full sentences.
2. **SEMANTIC VS. NUMERIC OUTPUT**:
  * **Location**: Translate coordinates into Anatomical Terms (e.g., "posterior segment of the right upper lobe"). NEVER output raw x, y, z coordinates.
  * **Computational**: Output the specific calculated number with units (e.g., "14.5 cc").

Figure 7: Structure of the System Prompt. The core principles, critical thinking rules, and standard operating procedures (SOPs) for the AI Medical Imaging Assistant are displayed. To facilitate a clear presentation, the specific technical definitions of available tools have been truncated.