MED3DNAVI: AGENTIC VISUAL EXPLORATION OF 3D CT VOLUMES VIA A STRUCTURED VOLUMETRIC READING PROTOCOL

Anonymous authors
Paper under double-blind review

ABSTRACT

Three-dimensional computed tomography (CT) is central to diagnosing abdominal, thoracic, and neurological conditions, yet vision-language models (VLMs) natively perceive only 2D images. Prior work bridges this gap in two ways, each with a characteristic failure mode. Fixed 2D conversions, whether uniformly sampled slices or a trained 3D encoder, commit what the model sees before it knows what it needs, yielding *vision without agency*. Text-based medical agents instead reason over tool reports without ever inspecting the pixels, yielding *agency without vision*. We argue that both limitations trace to a single root cause: the VLM is never placed in a closed-loop visual dialogue with the volume, unlike a radiologist who actively navigates and cross-checks across planes and windows before concluding. Consequently, we present Med3DNavi, a training-free agent that couples a programmable CT viewer, which renders on-demand multi-planar slices, grids, HU windowing, ROI zoom, and segmentation-mask overlays, with a Structured Volumetric Reading Protocol (SVRP). SVRP follows an orient–investigate–adjudicate workflow, routing each case either to an audit path for verifying an available tool prior or to an autonomous path for proactive and independent visual exploration. Experiments on DeepTumorVQA and 3D-RAD show that Med3DNavi improves over competitive baselines, with benefits that generalize across backbones. Our code is publicly available at `https://anonymous.4open.science/r/Med3DNavi-5530`.

1 INTRODUCTION

Three-dimensional computed tomography (CT) is the dominant imaging modality for diagnosing abdominal, thoracic, and neurological conditions, providing rich volumetric information about tissue density, organ morphology, and pathological changes across hundreds of contiguous slices. While recent advances in vision-language models (VLMs) have demonstrated remarkable capabilities in interpreting 2D medical images (Nori et al., 2023; Li et al., 2023; Sellergren et al., 2025; Team et al., 2025), extending these capabilities to 3D CT volumes remains a fundamental challenge: VLMs natively perceive only 2D images, yet the diagnostic evidence in a CT study is distributed throughout a 3D spatial volume. Bridging this dimensional gap is essential for building AI systems that can autonomously participate in radiological diagnosis.

Existing efforts to bridge this gap fall into two broad paradigms, each with a characteristic failure mode. The first paradigm converts the 3D volume into a fixed 2D representation before feeding it to the VLM, either by uniformly sampling slices along the axial axis, as in the default evaluation protocols of recent 3D CT benchmarks (Chen et al., 2026; Gai et al., 2026), or by compressing the volume into visual tokens with a trained 3D encoder (Bai et al., 2024; Wu et al., 2025). Because this conversion is open-loop, the slices, viewing plane, and HU window, or the encoder's compressed tokens, are all committed before the model knows what it will need to see. However, no fixed conversion fashion suits all cases, because slice relevance depends on the question being asked. The trained encoder faces the same dilemma in another form: its representation is optimized for average tasks rather than for the question at hand. In short, this paradigm gives the VLM **vision without agency**: it sees the frames it is handed but cannot act to obtain the ones it needs.

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.

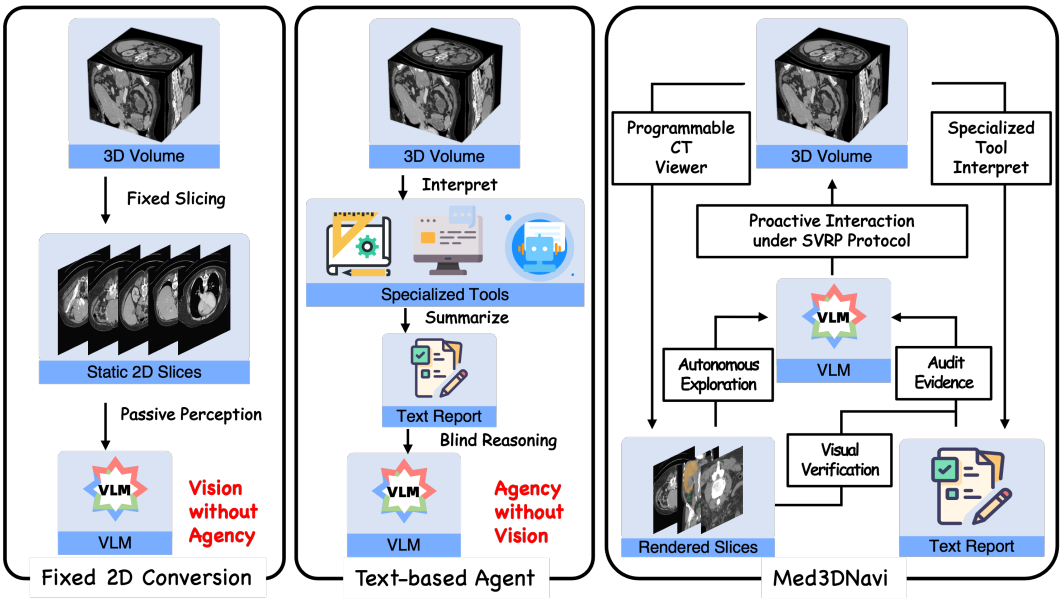

Figure 1: Three paradigms for VLM-based 3D CT diagnosis.

The second, more prevalent paradigm deploys VLM-based agents that interact with 3D data through specialized tools. Nevertheless, in most existing medical agent frameworks (Tang et al., 2024; Wang et al., 2025; Yu et al., 2025), the VLM only acts as a text processor. It receives structured reports or measurement outputs produced by upstream expert models (e.g., segmentation networks (Isensee et al., 2021; Wasserthal et al., 2023; He et al., 2025; Ma et al., 2024; Du et al., 2024) or detection pipelines) and performs reasoning purely over this textual intermediate. The VLM is thereby prevented from verifying whether those findings are visually supported in the original images. When the upstream model produces an erroneous segmentation boundary, misses a subtle lesion, or mischaracterizes a finding, the text-level agent has no mechanism to detect or correct the error. Hence the agent suffers from **agency without vision**, reasoning blindly over text: it can act and reason in a loop, yet never actively looks at the pixels.

Crucially, these two failure modes are two faces of a single root cause: **the VLM is never placed in a closed-loop, direct visual dialogue with the 3D volume**. One paradigm has eyes but no hands; the other has hands but no eyes. In neither case can the model decide what to look at next from what it has just seen. Yet this closed loop is exactly how a radiologist reads a CT study: actively navigating the volume and cross-checking across planes and windows before committing to a conclusion. Because standard VLMs are not natively equipped to perceive or interact with 3D CT volumes, reproducing this behavior requires both programmable access to the volumetric data and a structured protocol for active visual investigation.

To this end, we propose **Med3DNavi**, a training-free agent in which the backbone VLM actively and directly explores the target 3D CT volume. Med3DNavi couples a programmable CT viewer (rendering on-demand multi-planar slices, grids, HU windowing, ROI zoom, and segmentation-mask overlays) with a structured volumetric reading protocol (**SVRP**) that governs how the volume is explored (Fig. 1). The protocol follows a shared skeleton of orienting on the volume, investigating the target region, and adjudicating from the collected evidence. Its entry phase, Orient, localizes the target region and routes each case according to whether reliable prior evidence is available. When an upstream tool has already produced prior, the **audit** route re-attaches the visual grounding and verifies the claim against the pixels (for example, by scrolling a segmentation boundary under zoom) instead of trusting the number; when no reliable prior exists, the **autonomous** route explores the volume from scratch through the full loop of planning independent check paths, verifying them in isolated contexts, and adjudicating the collected evidence into a final verdict. The framework maintains per-phase tool-call budgets, restricted view modes, and per-path context isolation.

Our contributions are threefold: (1) We reframe 3D CT VLM reasoning as a *vision-agency* dichotomy, passive perception versus blind reasoning, traced to a common root: the absence of closed-loop direct visual exploration. We then present Med3DNavi, a training-free, backbone-agnostic agent that closes this loop. (2) We design SVRP, a structured volumetric reading protocol whose shared orient phase routes each case into an audit or an autonomous path, with multi-path independent verification and multimodal adjudication that restore visual verification of upstream outputs. (3) We evaluate Med3DNavi on DeepTumorVQA and 3D-RAD, analyzing routing behavior, per-route gains, and viewer-capability ablations. The method delivers strong results on both benchmarks, and the analyses attribute the gains to the closed-loop visual interaction.

2 Related Work

Vision-language models for medicine. Medical VLMs and vision-language foundation models (Li et al., 2023; Tu et al., 2024; Sellergren et al., 2025; Chen et al., 2024c;a; Lu et al., 2024; Yuan et al., 2026; Team et al., 2025) have been developed through biomedical instruction tuning and radiology-scale pretraining, improving performance on medical tasks such as visual question answering and report generation (Lau et al., 2018; Liu et al., 2021; He et al., 2020). However, these models remain largely native to 2D image inputs, thus handling a full 3D volume typically requires either additional volumetric training or reducing the volume to 2D surrogates. At the same time, modern general-purpose VLMs have become strong visual reasoners, exhibit competitive zero-shot competence on medical imagery, and can invoke external tools within agentic systems (Liu et al., 2023; Nori et al., 2023; Hurst et al., 2024; Chen et al., 2024b; OpenAI, 2026). This makes them an attractive foundation for 3D medical image navigation. Therefore, rather than training a new 3D-specialized model, we keep the backbone off-the-shelf and supply 3D capability through a programmable viewer and a structured reading protocol. This design keeps Med3DNavi training-free, backbone-agnostic, and easy to transfer across VLM backbones.

From 2D proxies to 3D volumetric diagnosis. The most direct route applies a 2D VLM to a volume by feeding a fixed set of sampled slices, which serves as a default evaluation protocol of recent 3D CT benchmarks (Chen et al., 2026; Gai et al., 2026), and it fixes what the model sees before reasoning begins. A second route trains 3D encoders or volumetric tokenizers so that an MLLM can consume the volume directly, as in M3D and RadFM (Bai et al., 2024; Wu et al., 2025); these models require large-scale 3D training data, and their compressed representation remains fixed before reasoning, leaving perception open-loop. A third route wraps LLMs or VLMs in agent frameworks that call specialized tools, yet in most medical agent systems the model reasons over textual reports and tool outputs without inspecting the images itself (Tang et al., 2024; Wang et al., 2025). Closest to our work, 3DMedAgent (Wang et al., 2026) closes a perception-reasoning loop over CT volumes, but its visual steps remain constrained: candidate slices are ranked by pretrained encoders rather than chosen from what the model has just seen, each verification step inspects a single slice with one tool operation, and tool-derived statistics enter as facts while visual grounding is never re-examined. Our method, Med3DNavi, instead lets the backbone VLM proactively compose multi-planar views, HU windows, and zooms under SVRP, checks hypotheses through context-isolated independent paths, and adjudicates them multimodally.

3 Method

3.1 Overview

Med3DNavi takes as input a CT volume $V \in \mathbb{R}^{H\times W\times D}$ of Hounsfield values, a clinical query q, and optionally a prior P. The prior may come from upstream reports or from outputs of external expert tools invoked during inference, such as segmentation maps, quantitative measurements, or other structured analyses. For simplicity, we precompute and unify all such priors into a textual report that is provided as part of the model input. The system couples three components: a frozen backbone VLM π, a programmable CT viewer $\mathcal{T}$ that renders the volume into 2D images on demand, and a protocol controller that executes SVRP and enforces its constraints on every tool call. Concretely, SVRP organizes diagnosis into a sequence of bounded tool-using phases; within a phase ϕ, the model alternately emits viewer actions and reads the returned images,

$$a_t = \pi(h_t), \quad o_t = \mathcal{T}(V, a_t), \quad h_{t+1} = h_t \cup \{(a_t, o_t)\}, \quad n_\phi \le B_\phi, \tag{1}$$

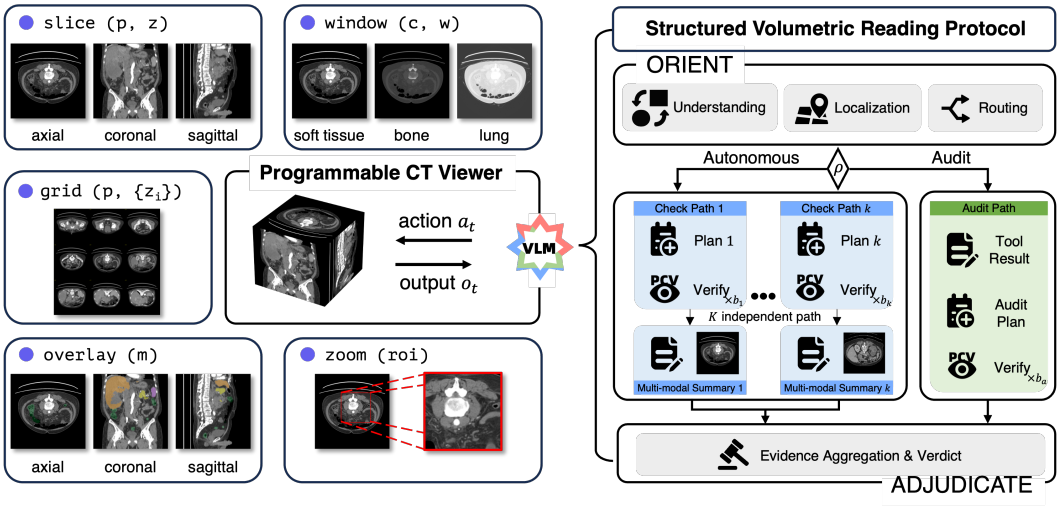

Figure 2: Med3DNavi Overview.

where a_t specifies a rendering request (Sec. 3.2), h_t is the phase context, o_t is the viewer output, n_ϕ counts the viewer calls made so far in phase ϕ, and the loop ends when the model emits a done signal or exhausts its budget B_ϕ. The SVRP protocol follows the skeleton of orienting, investigating, and adjudicating: a shared Orient phase localizes the target and selects a route $\rho \in \{\text{audit}, \text{autonomous}\}$, after which the two routes instantiate the investigation stage in different ways and finally converge on the same adjudication stage to fuse the collected summary (Fig. 2).

3.2 Programmable CT Viewer

The viewer abstracts the core functions of a radiologist's workstation into a compact action space, while remaining easy for a VLM to operate. Each action $a = (\text{mode}, \ell, z, (u, w), \text{roi}, \text{overlay})$ selects a rendering mode $\in \{\text{info}, \text{slice}, \text{grid}, \text{mip}\}$, an anatomical plane ℓ, a slice index or range z, an HU window (u, w), and optional region-of-interest zoom-in box and mask-overlay flag.

Fine-grained inspection with slice mode. The primitive operation of the viewer is a single windowed slice. Unlike natural images, CT is a volumetric modality where findings extend across three dimensions, and voxel intensities span thousands of Hounsfield units. These properties make direct display of the raw volume both inefficient and lossy, and motivate an interactive rendering interface.

The slice view is first rendered as $v = S_{\ell,z}(V)$, where $S_{\ell,z}$ extracts index z from plane $\ell \in \{\text{axial}, \text{coronal}, \text{sagittal}\}$. As clinically meaningful findings extend across neighboring slices and vary across orientations, slice mode and the associated rendering operation are available in all three planes, allowing the model to track a lesion spatially rather than rely on a single cross-section.

A second key control is windowing. Previous work often adopts a fixed window (Chen et al., 2026), but this is inherently biased toward particular organs or tissue types and may discard diagnostically relevant information elsewhere. Our viewer therefore allows the model to choose the observation window dynamically:

$$I_{u,w}(v) = \text{clip}\big((v - (u - w/2))/w,\ 0,\ 1\big), \tag{2}$$

where (u, w) is drawn from a set of clinical presets $\mathcal{W}$ (soft tissue, lung, bone, and others) or set freely by the model. This flexibility allows the VLM to expose subtle density differences and abnormalities that would otherwise be suppressed under a fixed display setting.

Two optional post-compositions further enhance the rendered evidence. (1) The viewer supports cropping and zooming, motivated by prior findings (Jiang et al., 2026) that VLMs perform better at interpreting fine-grained abnormalities. (2) The mask overlay projects a segmentation map onto the rendered slice, which serves two purposes: it allows the model to use segmentation, a fundamental tool in medical image inspection, as part of its own visual reasoning; and it enables the model to

assess whether the mask is anatomically accurate, which in turn lays the foundation for downstream auditing of tool-generated results.

Coarse browsing with grid mode. While slice mode supports precise local inspection, many findings must be assessed over a longer spatial extent. The grid mode addresses this need by tiling multiple slices into a single canvas: $\mathrm{tile}\big(\{I_{u,w}(S_{\ell,z_i}(V))\}_{i=1}^{n}\big)$, trading per-slice detail for broader coverage. This mode lets the model rapidly survey a range of slices in one glance before zooming into suspicious intervals. It serves the same purpose as scrolling through a CT stack—the way radiologists observe the emergence, persistence, or disappearance of a finding across depth—but achieves it by simultaneous presentation, so that depth-wise trends can be read off in a single view.

Info and MIP modes. The remaining two modes serve auxiliary roles. The info mode returns metadata μ, such as volume shape, voxel spacing, and intensity range, and slice overviews from three anatomical planes. These cues help the model build an initial spatial understanding of the volume when the tool prior does not provide sufficient guidance. The mip mode renders a maximum-intensity projection over a specified range, providing a compact summary that is particularly useful for highlighting high-attenuation structures or quickly screening extended regions.

3.3 Structured Volumetric Reading Protocol

The programmable CT viewer specifies the action space for volumetric inspection. On top of this interface, we define a structured volumetric reading protocol that governs how these actions are deployed during diagnosis. SVRP organizes diagnosis the way a radiologist organizes a read: orient first to establish where to look, investigate second to examine the question at hand, and adjudicate last to consolidate the evidence into an answer.

Orient Phase. The Orient phase establishes a basic spatial understanding of the volume before any diagnostic commitment is made. Its purpose is to first determine where to look, and only later reason about what the finding means. This separation reduces the risk that an early semantic impression biases subsequent verification, while the resulting context c can be reused by later phases without repeating the same localization steps. Operationally, the phase resembles the initial scout pass of a radiologist. With the help of CT viewer info mode or the tool prior, VLM is tasked only with geometric localization rather than diagnosis. The output is a structured context c paired with a routing decision ρ:

$$(c, \rho) = \mathrm{Orient}_{\pi,\mathcal{T}}(V, q, P), \quad c = (g,\ \mu), \quad \rho \in \{\text{audit}, \text{autonomous}\}, \tag{3}$$

where g is a three-dimensional bounding box for the target region and μ contains volume metadata. The route is set to audit when a reliable prior is available, that is, when the task falls within the capability of an upstream tool, and to autonomous otherwise.

Investigation Phase: (1) Autonomous route. When $\rho =$ autonomous, diagnosis proceeds without external tool prior. A single reasoning call first decomposes the question into K independent check strategies i_k with per-path budgets b_k:

$$\{(i_k, b_k)\}_{k=1}^{K} = \pi(q, c), \quad K \le K_{\max}, \quad \sum_k b_k \le B_V, \quad b_k \le b_{\mathrm{cap}}, \tag{4}$$

Here, each strategy defines a distinct line of inspection, such as checking different imaging signs, examining complementary anatomical views, or testing alternative explanations for the same finding. The budget vector thus acts as the model's own attention allocation over the sub-questions it considers worth checking. Each strategy then instantiates the Verify loop of Eq. 1 with its phase context initialized to the path-isolated context $\mathcal{M}_k = (q, c, i_k)$, i.e. $h_0^{(k)} = \mathcal{M}_k$, and with phase budget $B_\phi = b_k$:

$$(s_k, x_k) = \mathrm{Verify}_{\pi,\mathcal{T}}(V, \mathcal{M}_k) \tag{5}$$

where the model actively issues a sequence of viewer actions a_t to inspect the volume according to the plan, refining its search over slices, planes, windows, and zoom as evidence accumulates. Each path outputs a textual summary s_k and a decisive key image x_k. Isolation is enforced by construction rather than by prompting. Since no path can access another's observations, agreement across paths reflects genuine cross-checking rather than shared confirmation bias.

(2) Audit route. When $\rho =$ audit, the Orient phase has determined that a reliable prior produced by an upstream tool is available, so the task can in principle be addressed by that tool. In this case,

the role of the route shifts from evidence discovery to judging whether the returned tool result is trustworthy. This design addresses a key limitation of agents that reason purely over text: they can only passively accept tool outputs, with no mechanism to verify their correctness. Consider volume measurement as an example. A text-only agent has no choice but to trust the numerical result computed from a segmentation mask. However, in the clinically natural workflow, clinicians inspect the segmentation mask visually, scroll through its boundaries, and only then sign off on the derived volume. The audit route is designed to emulate this process. Rather than accepting the tool output at face value, it requires the VLM to validate the result by restoring the visual grounding discarded by standard tool interfaces and rendering it back onto the volume. Formally, given the tool result supplied by the prior P, the route runs as one focused verification phase that returns:

$$(s_a, x_a) = \mathrm{Audit}_{\pi,\mathcal{T}}(V, P, q, c) \tag{6}$$

where s_a records whether the tool result agrees with the volume or requires revision, together with the supporting reasoning, and x_a is the decisive overlay image. This single summary-image pair then enters the Adjudication phase as the degenerate case $K = 1$.

Adjudication Phase. The adjudication phase consolidates the evidence collected during investigation into a final answer. In the autonomous route, it aggregates the K independent paths through a single multimodal call:

$$y = \mathrm{Adjudicate}_{\pi}\big(q, c, \{(s_k, x_k)\}_{k=1}^{K}\big), \tag{7}$$

where the adjudicator reads each path summary s_k together with its key image x_k. This design preserves the visual grounding of each claim, allowing the model to compare supporting evidence across paths, identify agreement or conflict, and determine which conclusion is best supported by the images. In the audit route the input set collapses to the single pair (s_a, x_a), so cross-path comparison degenerates to reading one summary against its overlay image.

4 Experiments

4.1 Experimental Setup

Benchmarks and metrics. We evaluate Med3DNavi on two public 3D CT benchmarks. DeepTumorVQA (Chen et al., 2026) contains four subtasks, all evaluated by accuracy. 3D-RAD (Gai et al., 2026) includes three generation subtasks, which are evaluated using BERTScore (Zhang et al., 2020), and three classification subtasks, which are evaluated by accuracy.

Baselines. We compare against three families. (i) 2D VLMs: Lingshu-32B (Team et al., 2025), MedGemma-27B (Sellergren et al., 2025), Qwen3.6-27B (Qwen Team, Alibaba Group, 2026), and Gemini-3.5-Flash (Google DeepMind, 2026), following the official DeepTumorVQA protocol of feeding nine uniformly sampled slices. (ii) 3D-encoder VLMs: RadFM (Wu et al., 2025) and M3D (Bai et al., 2024) (based on Llama2-7B), evaluated in zero-shot setting. (iii) Agentic baseline: 3DMedAgent (Wang et al., 2026) (backbone Qwen3.6-27B).

Med3DNavi settings. Med3DNavi is training free and backbone agnostic. We instantiate it with Qwen3.6-27B and Gemini-3.5-Flash respectively under the same viewer and SVRP configuration. The Orient phase uses at most 3 viewer calls, each Verify path uses at most 6 viewer calls, with a total Verify budget of 12 over at most 3 paths. Plan and Adjudication each use a single tool free reasoning call. All runs use temperature 0 for deterministic decoding, so that performance differences reflect the method itself rather than stochastic variation in generation. Following the 3DMedAgent implementation, we use the structured report derived from the VISTA3D (He et al., 2025) segmentation mask as the tool prior.

4.2 Main Results

Table 1 reports the main comparison results. We highlight three observations.

Observation 1. Closed-loop exploration outperforms fixed perception. Across both DeepTumorVQA and 3D-RAD, Med3DNavi improves over 2D VLM baselines built on the same backbone, including both Qwen3.6-27B and Gemini-3.5-Flash. Replacing the standard nine-slice input with viewer-based exploration yields consistent gains across all subtasks. Med3DNavi also substantially

Table 1: Main results on DeepTumorVQA and 3D-RAD. Qwen means Qwen3.6-27B, Gemini means Gemini-3.5-Flash. Best results are shown in bold while second-bests are underlined.

	2D VLM				3D Encoder		Agent		
Task / Model	Lingshu 32B	MedGemma 27B	Qwen 27B	Gemini –	RadFM 13B	M3D 7B	3DMedAgent† (Qwen)	Med3DNavi (Qwen)	Med3DNavi (Gemini)
DeepTumorVQA									
Measurement	32.4	21.7	31.1	32.4	26.9	17.1	62.0	64.0	**68.2**
Med. Reasoning	26.2	29.7	48.3	58.6	17.9	30.3	60.6	**68.4**	68.2
Recognition	60.2	59.2	62.7	63.0	51.4	38.0	51.0	67.6	**69.7**
Vis. Reasoning	41.4	37.2	43.8	41.2	23.1	26.3	50.6	**59.6**	57.6
Avg.	40.0	37.0	46.5	48.8	29.8	27.9	56.0	64.9	**65.9**
3D-RAD									
Exist. Detection	81.6	75.6	65.8	73.1	29.2	18.0	**82.8**	82.1	75.6
Long. Temporal	35.9	39.0	57.1	67.3	44.1	25.5	54.0	72.8	**76.0**
Static Temporal	24.7	20.5	38.5	57.6	43.0	24.1	36.8	55.2	**63.5**
Classifi. Avg.	47.4	45.0	53.8	66.0	38.8	22.5	57.9	70.0	**71.7**
Img. Observation	89.8	87.3	90.0	90.2	86.8	86.1	–	**91.6**	90.3
Anom. Detection	89.2	87.5	89.6	88.8	87.2	86.6	–	**90.4**	90.3
Med. Computation	87.5	91.3	93.3	90.0	86.9	91.5	–	**94.3**	93.8
Generation Avg.	88.8	88.7	91.0	89.6	86.9	88.1	–	**92.1**	91.5

†3DMedAgent's code only supports multiple-choice questions.

exceeds 3D encoder VLMs. This pattern suggests that the key advantage comes from closed loop volumetric exploration, which allows the model to gather evidence iteratively under its own control instead of relying on a fixed visual representation.

Observation 2. Model-directed inspection outperforms tool-constrained slice verification. Under the same Qwen3.6-27B backbone, Med3DNavi outperforms 3DMedAgent on all DeepTumorVQA subtasks and on most 3D-RAD subtasks. The difference lies in how the two systems acquire and validate visual evidence. In 3DMedAgent, visual exploration remains largely tool guided and each verification step is limited to a single tool operation. By contrast, Med3DNavi lets the model actively control the viewer based on what it has just seen, compose multi-step inspection paths, and verify sub-questions in isolated contexts before adjudicating across them. This comparison suggests that active, visually proactive exploration is more effective than passive verification over tool proposed slices.

Observation 3. The gain transfers across backbones. The same qualitative trend is preserved when Med3DNavi is instantiated with Gemini-3.5-Flash instead of Qwen3.6-27B. In both cases, it consistently improves over the corresponding 2D baseline. This cross-backbone consistency indicates that the benefit mainly comes from the generalizable programmable viewer and the structured volumetric reading protocol, rather than from any model-specific adaptation.

4.3 Route Analysis

Routing behavior. Figure 3 shows the routing pattern across DeepTumorVQA question subtypes. Quantitative and directly measurable subtypes, such as organ volume measurement and HU statistics, are routed almost entirely to the audit route, since upstream tools can provide structured priors by parsing the segmentation mask. By contrast, more open-ended diagnostic subtypes, such as lesion counting and staging, are routed to the autonomous route, where no comparable tool prior is available in our setting. The per-subtype accuracy markers further show that audit-dominant quantitative subtypes are solved with high accuracy, whereas the more open-ended autonomous subtypes remain more challenging, consistent with the lack of reliable priors. Between these two anchors, a band of

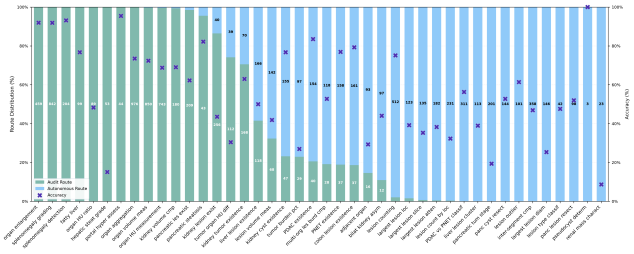

Figure 3: Routing behavior on DeepTumorVQA.

partially measurable subtypes carries both routes, and it can be traced to the case-level availability and reliability (App. C.4). Route assignments are also consistent across backbones (detailed in App. C.3), with question-level agreement of 80.7%, indicating that routing is driven by task structure rather than by the backbone.

Per-route behavior. Table 2 further analyzes the two routes on the task subsets assigned exclusively to each route by ablating the programmable CT viewer and SVRP. Disabling the viewer restricts the model to making decisions solely from tool priors, while disabling SVRP leaves the viewer available but removes the structured control protocol, so the model explores in an unconstrained ReAct-style (Yao et al., 2023) manner. For fairness, we prompt the model to match the average number of tool calls in the SVRP condition. **On the audit subset**, the viewer is the dominant factor. Without it, the model can only read the tool prior as text and cannot verify whether it is visually grounded in the volume, leading to a sharp performance drop. Enforcing the full protocol yields a further gain. Importantly, this gain originates from SVRP exclusively, as under the same average tool call number, the structured protocol performs better. Concretely, among the audit cases that the prior-only condition answers incorrectly, 63% are recovered once the prior is re-grounded and audited against the volume. These support the motivation of the audit route: a tool prior becomes useful only when it is re-grounded and verified against the image. **On the autonomous subset**, both the viewer and the protocol are necessary. Viewer access alone yields mixed changes, improving medical and visual reasoning but degrading recognition; consistent gains emerge only when it is paired with SVRP, which structures exploration to avoid wasting the interaction budget or drifting from relevant evidence.

Table 2: Route-level analysis on DeepTumorVQA.

Viewer	SVRP	Measurement	Med. Reasoning	Recognition	Vis. Reasoning	# Calls
Autonomous route						
–	–	–	47.1	60.4	54.6	–
✓	–	–	49.6	58.6	58.1	12.6$_{\pm 4.3}$
✓	✓	–	**55.8**	**64.5**	**60.2**	12.3$_{\pm 1.5}$
Audit route						
–	–	61.5	67.2	83.0	55.4	–
✓	–	63.5	88.8	92.7	73.9	4.2$_{\pm 1.1}$
✓	Auto†	65.8	82.0	85.7	71.0	11.8$_{\pm 1.5}$
✓	✓	**69.4**	**91.3**	**93.1**	**77.8**	4.6$_{\pm 1.5}$

†cases of the audit subset reprocessed by the autonomous route.

We also evaluate what happens when routing is intentionally mismatched. Reprocessing the audit subset with the autonomous route leads to lower accuracy and higher interaction cost, because brute-force exploration often ignores a prior that the audit route would verify directly. This comparison confirms that the two routes are not interchangeable and more interaction alone does not improve accuracy. The benefit comes from how viewer calls are structured, not from how many are made.

4.4 Ablation Study of the Programmable CT Viewer

Capability ablation. Table 3 ablates the viewer capabilities by disabling them one at a time from the full configuration. Removing all four advanced capabilities simultaneously causes the largest performance drop on most subtasks. This confirms that the full capability set is important for effective volumetric reading. The leave one out results further clarify the role of each component. Removing mask overlay causes the largest degradation on measurement, indicating that overlay is critical for grounding quantitative priors in visible anatomy. Removing HU windowing most strongly hurts medical reasoning, consistent with the fact that window adjustment reveals density contrasts that may be suppressed under a fixed display setting. Removing ROI zoom leads to moderate but consistent drops across tasks, suggesting that zoom mainly supports fine grained inspection. Removing multi-planar views causes the smallest degradation on average, suggesting that axial views are often sufficient for many cases, which is consistent with routine clinical practice, although complementary planes still provide additional useful evidence.

View-mode ablation. Restricting the viewer to a single view mode also reduces performance. It reduces performance on all four subtasks, most notably visual reasoning, which drops from 60.3 to 54.9 when only the slice mode is enabled, and to 57.1 when only the grid mode is enabled. The best performance is obtained when both modes are available. The two modes thus provide complementary evidence: slice mode supports detailed local inspection, while grid mode provides broader cross-depth coverage to track how findings appear and persist across the volume.

Table 3: Ablation of the programmable CT viewer on a randomly sampled 1,000-case subset of DeepTumorVQA (accuracy, %). ✓ marks an enabled capability or view mode; disabling multi-planar viewing leaves axial views only, and disabling HU windowing uses a fixed (40, 400) window. Best in bold.

Multi-planar	Window	Mask	Zoom	Slice	Grid	Measurement	Med. Reasoning	Recognition	Vis. Reasoning
✓	✓	✓	✓	✓	✓	**67.6**	**70.2**	**64.9**	**60.3**
Capability ablation (slice and grid enabled)									
–	✓	✓	✓	✓	✓	67.3	69.2	63.5	59.5
✓	–	✓	✓	✓	✓	66.5	62.1	63.0	58.2
✓	✓	–	✓	✓	✓	60.4	65.7	64.4	59.5
✓	✓	✓	–	✓	✓	65.6	68.7	61.9	57.6
–	–	–	–	✓	✓	55.4	61.1	61.4	53.5
View-mode ablation (all capabilities enabled)									
✓	✓	✓	✓	✓	–	66.3	68.2	64.4	54.9
✓	✓	✓	✓	–	✓	67.4	67.0	64.8	57.1

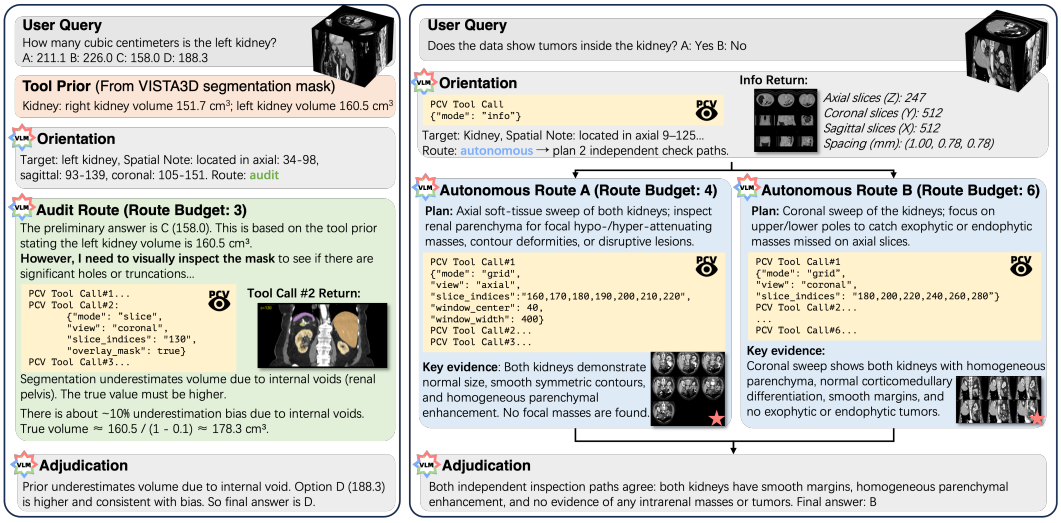

Figure 4: Two representative Med3DNavi runs, one per route.

4.5 Case Study

Figure 4 shows two runs where closed-loop dialogue turns a raw prior or unguided evidence into a verifiable conclusion. In the audit route (left), the VISTA3D prior reports a left-kidney volume of 160.5 cm^3 (option C); mask-overlay re-grounding reveals mask-unfilled internal voids, i.e. about 10% underestimation; correcting it shifts the answer to D ($\approx$ 178.3 cm^3). In the autonomous route (right), lacking a usable prior, the agent plans axial and coronal grid sweeps; both find homogeneous parenchyma and no focal mass, adjudicating B (No).

5 Conclusion

We presented **Med3DNavi**, a training-free, backbone-agnostic agent that lets a general VLM actively explore 3D CT volumes through a programmable viewer and a Structured Volumetric Reading Protocol (SVRP). Combining on-demand volumetric access with audit/autonomous routing, it closes the visual investigation loop missing from direct 3D VLMs and prior tool-mediated agents. Without task-specific training, it outperforms strong 2D VLMs, 3D encoder VLMs, and an agentic baseline on DeepTumorVQA and 3D-RAD across open and proprietary backbones, while yielding auditable traces exposing the evidence behind each answer. We hope this work offers a practical recipe for grounding VLMs in volumetric medical imaging and for verifiable multi-path reasoning.

AI USE STATEMENT

Vision-language models are used in this work as the backbone agents of the proposed Med3DNavi framework: each VLM serves as the reasoning core that drives the programmable CT viewer and executes the SVRP protocol. These models are the object of our study and all experimental results are obtained by running them within our framework. In addition, we used a large language model to assist with language editing and polishing of the text. All scientific content, claims, and conclusions were produced, reviewed, and verified by the authors, who take full responsibility for the integrity of the work.

ETHICS STATEMENT

This work uses only publicly available research benchmarks (DeepTumorVQA and 3D-RAD), whose patient data have already been de-identified and anonymized by the original curators. We did not collect any new human-subject data, access private medical records, or conduct experiments involving patients, and all datasets and models are used in accordance with their respective licenses and terms of use. We therefore foresee no ethical concerns beyond those already addressed by the dataset providers.

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.

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

Yixiong Chen, Wenjie Xiao, Pedro RAS Bassi, Boyan Wang, Liang He, Xinze Zhou, Sezgin Er, Ibrahim Ethem Hamamci, Zongwei Zhou, and Alan Yuille. DeepTumorVQA: A hierarchical 3D CT benchmark for stage-wise evaluation of medical VLMs and tool-augmented agents. *arXiv preprint arXiv:2605.09679*, 2026.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024b.

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, et al. CheXagent: Towards a foundation model for chest X-ray interpretation. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*, 2024c.

Yuxin Du, Fan Bai, Tiejun Huang, and Bo Zhao. SegVol: Universal and interactive volumetric medical image segmentation. *Advances in Neural Information Processing Systems*, 37:110746–110783, 2024.

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.

Google DeepMind. Gemini 3.5 Flash. `https://deepmind.google/models/model-cards/gemini-3-5-flash/`, 2026. Model card, May 19, 2026.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. PathVQA: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

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.

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

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

Zheng Jiang, Heng Guo, Chengyu Fang, Changchen Xiao, Xinyang Hu, Lifeng Sun, and Minfeng Xu. MedVR: Annotation-free medical visual reasoning via agentic reinforcement learning. *arXiv preprint arXiv:2604.08203*, 2026.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):180251, 2018.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in neural information processing systems*, 36: 28541–28564, 2023.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pp. 1650–1654. IEEE, 2021.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature medicine*, 30(3):863–874, 2024.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature communications*, 15(1):654, 2024.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*, 2023.

OpenAI. GPT-6 Astra. https://openai.com/index/gpt-6-astra/, 2026.

Qwen Team, Alibaba Group. Qwen3.6-27B. https://qwen.ai/blog?id=qwen3.6-27b, 2026. Model release, April 21, 2026.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. MedGemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

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

LASA Team, Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, Yu Sun, Junao Shen, Chaojun Wang, Jie Tan, Deli Zhao, Tingyang Xu, Hao Zhang, and Yu Rong. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *CoRR*, abs/2506.07044, 2025. doi: 10.48550/ARXIV.2506.07044. URL https://doi.org/10.48550/arXiv.2506.07044.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical AI. *Nejm Ai*, 1(3):AIoa2300138, 2024.

Wenxuan Wang, Zizhan Ma, Zheng Wang, Chenghan Wu, Jiaming Ji, Wenting Chen, Xiang Li, and Yixuan Yuan. A survey of LLM-based agents in medicine: How far are we from Baymax? *Findings of the Association for Computational Linguistics: ACL 2025*, 2025.

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*, 2026.

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, Aug 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-62385-7. URL https://doi.org/10.1038/s41467-025-62385-7.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Yongrui Yu, Zhongzhen Huang, Linjie Mu, Shaoting Zhang, and Xiaofan Zhang. Radiologist Copilot: An agentic framework orchestrating specialized tools for reliable radiology reporting. *arXiv preprint arXiv:2512.02814*, 2025.

Hangjie Yuan, Yichen Qian, Zhiwei Tang, Xianzhe Xu, Lirong Wu, Sicheng Yang, Jinwang Wang, Pengju Wang, Zhitao Zeng, Yizeng Han, et al. Clinfusion: A vision-centric multimodal llm system for holistic medical understanding. *arXiv preprint arXiv:2607.24743*, 2026.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020.

A PSEUDOCODE OF SVRP

Algorithm 1 summarizes one Med3DNavi run. In our default configuration the Orient budget is 3 calls, each Verify path is capped at 6 calls with a total verify budget of 12 across at most 3 paths, and the Plan and Adjudicate phases make exactly one tool-free call each. Every viewer call inside

Algorithm 1 SVRP execution in Med3DNavi

Require: volume V, question q, optional prior P

$c, \rho \leftarrow \textsc{Orient}_{\pi,\mathcal{T}}(V, q, P)$ ▷ geometry only, budget B_o
if $\rho =$ audit **then**
  $(s_a, x_a) \leftarrow \textsc{Audit}_{\pi,\mathcal{T}}(V, P, q, c)$ ▷ overlay inspection, budget B_a
  $y \leftarrow \textsc{Adjudicate}_{\pi}(q, c, \{(s_a, x_a)\})$ ▷ degenerate fan-in, $K = 1$
else
  $\{(i_k, b_k)\}_{k=1}^{K} \leftarrow \textsc{Plan}_{\pi}(q, c)$ ▷ model-proposed budgets, capped
  for $k = 1 \dots K$ **do**
    $\mathcal{M}_k \leftarrow (q, c, i_k)$ ▷ path-isolated context
    $(s_k, x_k) \leftarrow \textsc{Verify}_{\pi,\mathcal{T}}(V, \mathcal{M}_k)$ ▷ budget b_k
  end for
  $y \leftarrow \textsc{Adjudicate}_{\pi}(q, c, \{(s_k, x_k)\})$ ▷ multimodal fan-in
end if
return y

ORIENT, AUDIT, and VERIFY passes the controller, which rejects calls beyond the phase budget or outside the allowed mode set.

B IMPLEMENTATION DETAILS

B.1 COMPUTATIONAL ENVIRONMENT AND MODEL SERVING

In our experiment setting, the open-sourced backbone Qwen3.6-27B (Thinking mode disabled) is served locally with vLLM (v0.20.0) on two NVIDIA H100 80 GB GPUs, using tensor-parallel size 2, a maximum model length of 32,768 tokens, and GPU memory utilization 0.8, with all other serving options left at their defaults. The proprietary backbone, Gemini-3.5-Flash, is accessed through its official API. Both backbones are sampled at temperature 0 for deterministic decoding. Our code is publicly available at `https://anonymous.4open.science/r/Med3DNavi-5530`.

B.2 TOOL PRIOR CONSTRUCTION

As stated in **Med3DNavi settings**, we reuse the structured-report design of 3DMedAgent (Wang et al., 2026): the tool prior P is a textual report derived from the VISTA3D (He et al., 2025) segmentation mask. Concretely, for each case we assemble three kinds of blocks. A scan-level block reports the HU intensity range and robust percentiles (p1, p5, p50, p95, p99) of the whole volume. For every organ present in the mask, an organ block reports its volume in cm^3 (the foreground voxel count after resampling the mask to 1 mm isotropic spacing, divided by 1000), a trimmed mean$\pm$std HU computed within the organ mask after excluding lesion voxels and discarding the outer 5% tails, and a 3D bounding box given as per-plane slice-index ranges measured at the original resolution. The blocks are concatenated into a single report that is supplied to the model as P; the same VISTA3D mask is loaded by the viewer for mask overlay, so that the audit route can re-ground the report against the image. A representative example is shown below:

Representative Tool Prior Report

```
[CT_HU_META]
HU_range:  min=-1000.0, max=1000.0
HU_percentiles:  p1=-1000.0, p5=-1000.0, p50=-979.0,
    p95=133.0, p99=250.0

Spleen:
```

```
Normal size (volume:  153.7 cm^3).
Mean HU value:  108.8 +/- 25.0.
3D Bounding Box (axial:  77-128, sagittal:  62-118,
    coronal:  96-141).
Liver:
Normal size (volume:  1679.5 cm^3).
Mean HU value:  116.1 +/- 23.9.
3D Bounding Box (axial:  5-141, sagittal:  81-228,
    coronal:  100-196).
Pancreas:
Normal size (volume:  80.4 cm^3).
Mean HU value:  99.1 +/- 25.9.
3D Bounding Box (axial:  30-92, sagittal:  80-174,
    coronal:  124-171).
Kidney:
Normal size (right kidney volume:  204.1 cm^3;
    left kidney volume:  151.0 cm^3;
    total kidney volume:  355.1 cm^3).
Mean HU value:  158.6 +/- 30.8.
Right kidney 3D Bounding Box
    (axial:  48-126, sagittal:  156-205, coronal:  95-133).
Left kidney 3D Bounding Box
    (axial:  32-95, sagittal:  73-125, coronal:  102-137).
Colon:
Volume:  325.2 cm^3.
3D Bounding Box (axial:  0-88, sagittal:  63-218,
    coronal:  121-196).
```

B.3 HU Window Presets

The preset set $\mathcal{W}$ of the viewer (Sec. 3.2) contains six clinical windows, listed in Tab. 4 with their center u and width w in HU. The full table is injected as a window-reference card into the three tool-using phases (Orient, Verify, and Audit), so the model can either name a preset or specify a free (u, w) pair; when a call omits windowing, the viewer falls back to the soft-tissue default ($u = 40$, $w = 400$). Narrow windows (brain, liver) preserve subtle soft-tissue contrast, whereas wide windows (lung, bone) prevent air- and bone-density structures from saturating.

Table 4: Clinical HU window presets $\mathcal{W}$ available to the viewer.

Preset	Center u (HU)	Width w (HU)
Soft tissue / abdomen	40	400
Lung	-600	1500
Bone	400	1800
Brain	40	80
Mediastinum	50	350
Liver	60	160

B.4 Prompt Templates

In this section, we list the system prompts used in the SVRP phases. To keep the presentation concise, we omit routine prompt components, such as output-format control, and retain only the key instructions most relevant to the protocol.

Phase Orient (shared)

```
You are a radiology navigation agent in the ORIENT phase.
Your ONLY goal is to establish the spatial location of the
anatomical region relevant to the question.  You are
```

```
building a coordinate frame, NOT making observations.

## CRITICAL CONSTRAINT
This phase determines WHERE, not WHAT. You must NOT
describe or evaluate density, morphology, contour, or any
other visual feature.  Your summary must contain ONLY
spatial information (organ name, slice range, position).

## Route decision
- "audit":  a preliminary_answer can be formed from the
   report alone.
- "autonomous":  the report is absent or the question needs
   open-ended visual reasoning.
When in doubt, choose "autonomous".
```

Audit Route

```
You are a radiology auditor in AUDIT mode.  The ORIENT
phase has determined that the existing evidence (clinical
report) is sufficient to form a preliminary answer.  Your
task is to audit that evidence, then confirm or revise the
answer.

## Three-step audit protocol
Step 1 - Restate the preliminary conclusion and identify
   WHICH pieces of evidence it depends on.
Step 2 - Design and execute the audit:  if the answer
   depends on a segmentation measurement, use
   overlay_mask=true to inspect the boundary for false
   positives / negatives and boundary drift.
Step 3 - Revise or confirm:  reliable -> confirm; minor
   issues -> adjust; clearly wrong -> replace with your
   own visual assessment.

## Revision discipline
The bar for OVERRIDING the report is HIGH. For HU /
density questions the report's voxel-level values are MORE
reliable than your visual estimate; only override if the
mask is clearly misplaced.
```

Autonomous Route (Plan + Verify)

```
[PLAN] You are a radiology examination planner.  Based on
the question and the spatial priors from ORIENT, design up
to {max_paths} independent CT check paths.  Each path is a
self-contained inspection strategy:  HOW to look, not WHAT
the answer is.  Paths must differ substantively in plane,
window, search strategy, criteria, stance, or mask-overlay
audit.
## CRITICAL CONSTRAINT
A path is a CHECK STRATEGY, not an answer guess.  The
instruction must never presuppose a conclusion.

[VERIFY] You are a radiology inspector executing ONE
independent check path in an isolated context.  Follow the
instruction's plane, window, search strategy and criteria.
Your judgment must be based ENTIRELY on what you observe in
the images during THIS inspection.  Do not use any prior
diagnostic labels from the clinical report as evidence.
```

Phase Adjudicate (shared)

```
You are a senior radiologist in the ADJUDICATE phase.  You
receive the conclusions and key images from {num_paths}
independent inspection paths.  Synthesize them into a final
answer.
```

```
## Evidence source calibration
- QUANTITATIVE answers: prefer the report's numbers UNLESS
  a path with overlay_mask=true reveals a clear
  segmentation failure.
- QUALITATIVE answers: the VLM visual evidence from paths
  takes priority.
Weigh the evidence across paths: do not simply count how
many paths agree; assess the QUALITY and consistency of the
observations. If paths disagree, examine the key images
yourself to resolve the conflict.
```

C Supplementary Experiments

C.1 Viewer-Grounding Feasibility

Med3DNavi requires that a backbone VLM can operate the programmable CT viewer, namely, localize anatomy well enough to issue meaningful slice-range and ROI commands. We test this assumption with two grounding tasks (Tab. 5): a *coronal* probe, which simulates prior-free orientation by predicting an organ's z-interval from a concatenated coronal overview, and an *axial* probe, which evaluates ROI control by predicting an organ bounding box from a single axial slice. We report IoU against mask-derived ground truth, together with a random baseline to calibrate chance level.

Table 5: Viewer-grounding feasibility (IoU). The coronal probe emulates prior-free Orient (z-interval localization); the axial probe emulates the ROI function (per-slice bbox).

	Coronal (Orient proxy: z-interval)					Axial (ROI proxy: bbox)				
Model	Lung	Liver	Kidney	Pancreas	Mean	Lung	Liver	Kidney	Pancreas	Mean
Qwen3.6-27B	0.73	0.72	0.46	0.38	0.57	0.66	0.49	0.33	0.22	0.43
Gemini-3.5-Flash	0.83	0.85	0.75	0.61	0.76	0.76	0.75	0.53	0.19	0.56
Random	0.35	0.26	0.27	0.23	0.28	0.11	0.10	0.09	0.03	0.08

Across both probes and both backbones, every organ cell exceeds the random baseline by a wide margin: mean coronal IoU is 0.57 (Qwen3.6-27B) and 0.76 (Gemini-3.5-Flash) against 0.28 for random, and mean axial IoU is 0.43 and 0.56 against 0.08 for random (≈5–7×). This level of grounding is already sufficient for our setting, where the VLM can make repeated viewer calls and progressively refine its localization. Current VLMs therefore carry usable anatomical spatial grounding ability and can issue meaningful viewer commands rather than searching at random, validating the premise that a generalist VLM can operate our CT viewer.

C.2 Path Count Analysis

We analyze how the number of verification paths K in the autonomous route trades accuracy against test-time cost. Figure 5 fixes $K \in \{1, 2, 3\}$ and compares against the adaptive setting on the autonomous-branch average; Fig. 6 shows, per subtype, the mean number of paths the adaptive planner actually issues against the accuracy it obtains. Together they support three observations.

Multi-path verification is the key driver of improvement, but its returns quickly saturate. Accuracy increases only when multiple independent paths are verified in isolated contexts and fused by adjudication ($K = 2$: 58.9; $K = 3$: 60.2), showing that the autonomous route benefits from the multi-path structure. At the same time, the marginal increase drops rapidly as K grows ($+3.3$ from $K = 1$ to 2, $+1.3$ from 2 to 3, and only $+0.6$ beyond), indicating that most of the gain is already realized by $K = 3$. This supports capping the verification budget at three paths, which keeps test-time cost bounded while sacrificing little accuracy.

Adaptive K is Pareto-dominant. The adaptive setting attains the highest accuracy (60.8) while issuing fewer paths on average ($\overline{K} = 2.44$) than the best fixed budget ($K = 3$, 60.2). In Fig. 5 the adaptive point lies above and to the left of the fixed-K curve, i.e. it is simultaneously more accurate and cheaper, validating the complexity-aware path selection performed in the Plan phase.

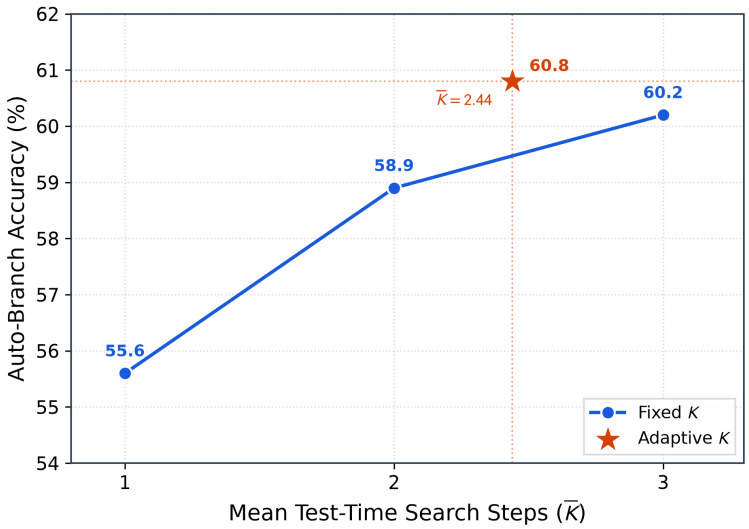

Figure 5: Test-time search scaling on the autonomous route. Fixing the verify-path count to $K \in \{1, 2, 3\}$ traces a saturating accuracy curve; the adaptive setting (star) reaches the highest accuracy at a lower mean path count ($\overline{K} = 2.44$), sitting above and to the left of the fixed-K curve.

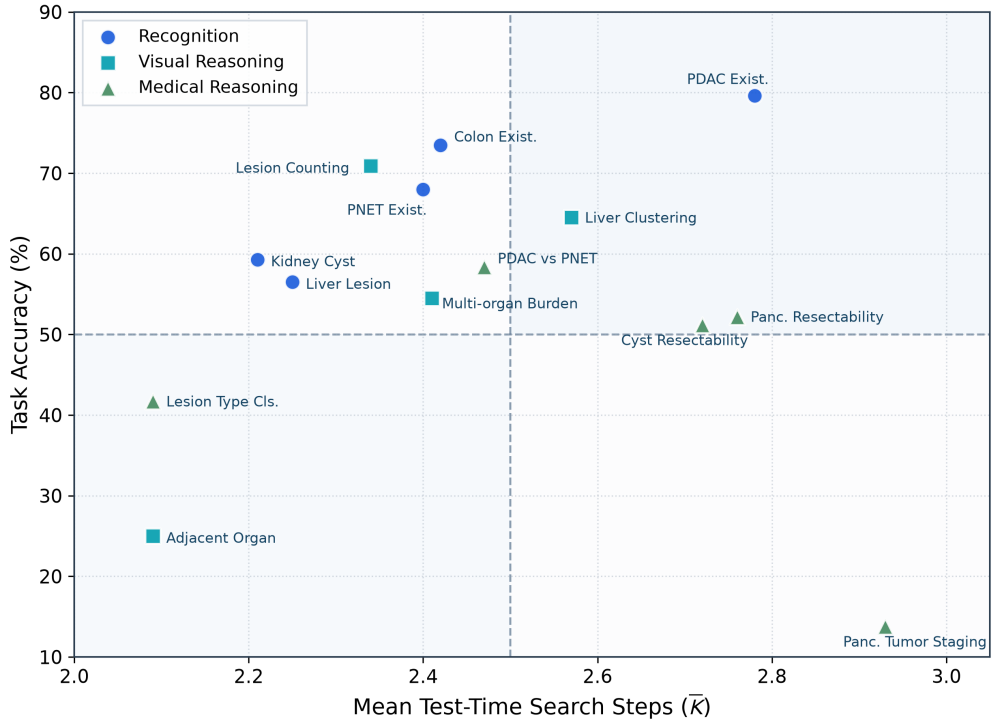

Figure 6: Per-subtype allocation of verify paths under the adaptive setting on the autonomous route. Each marker is a question subtype; the horizontal axis is the mean number of verify paths the planner issues and the vertical axis is accuracy.

The planner allocates search on demand. Figure 6 shows that the adaptive mean path count is differentiated across subtasks (2.09–2.93) rather than degenerating to the cap. Subtasks that reward deeper exploration receive more paths and convert them into accuracy (e.g., PDAC existence $\overline{K} = 2.78$, 79.6%; liver lesion clustering $\overline{K} = 2.57$, 64.5%), whereas early-converging subtasks are

resolved with fewer (e.g., lesion counting $\overline{K} = 2.34$, 70.9%; colon lesion existence $\overline{K} = 2.42$, 73.5%). A few subtasks remain hard even with more paths (e.g., pancreatic tumor staging, adjacent organ), marking clear directions for future work.

C.3 Cross-Backbone Routing Analysis

To test whether the Orient router reflects task structure rather than stochasticity, we compare the per-question route assignments produced by Qwen3.6-27B and Gemini-3.5-Flash. The two backbones agree on the route for 80.7% of questions (Cohen's $\kappa = 0.62$) and on the dominant route for 36 of 42 subtypes, and their per-subtype audit fractions are strongly correlated (Pearson $r = 0.75$). Fig. 7 overlays the two per-subtype audit-fraction profiles sorted by their mean fraction: the curves coincide across both anchor regimes and diverge only inside the transition band. We also summarize the following observations:

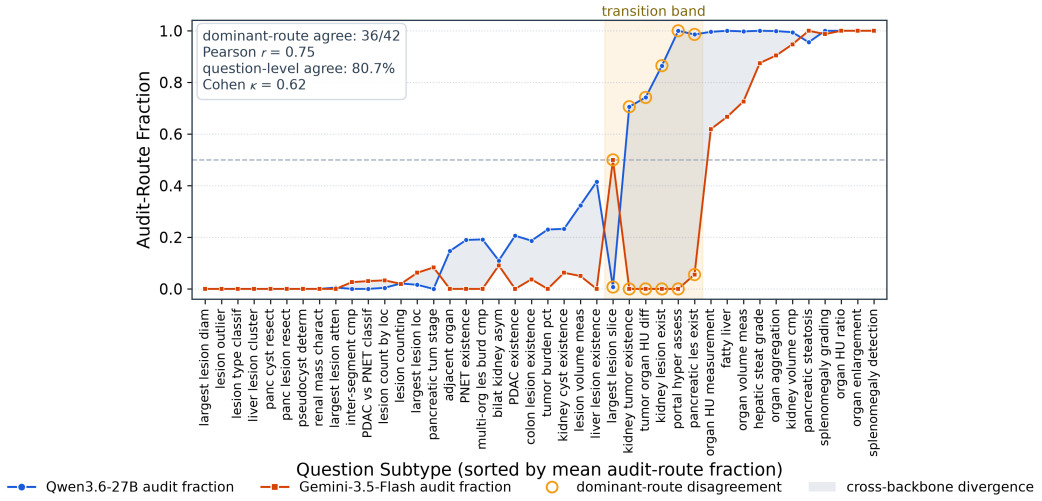

Figure 7: Cross-backbone routing consistency. The per-subtype audit-route fractions of Qwen3.6-27B and Gemini-3.5-Flash are overlaid as two line profiles over the 42 subtypes sorted by their mean fraction; the shaded area marks cross-backbone divergence, and rings mark the six subtypes whose dominant route differs. The profiles coincide at the autonomous (left) and audit (right) anchors and separate only within the transition band.

Routing is task-driven, not backbone-driven. For most cases, agreement is near-perfect in the two anchor regimes (the flat plateaus at both ends of Fig. 7): purely quantitative subtypes (e.g., organ HU ratio, organ enlargement, splenomegaly detection) are routed to audit by both backbones (fractions $\approx$1.0), and purely open-ended subtypes (e.g., lesion counting, PDAC vs PNET, lesion type classification) are routed to autonomous by both (fractions $\approx$0.0). The router therefore tracks properties of the task and the tool rather than model preference, supporting the backbone-agnostic claim.

Disagreement is confined to the transition band and is one-directional. The six subtypes whose dominant route differs (pancreatic lesion existence, kidney lesion and tumor existence, tumor organ HU difference, portal hypertension, largest lesion slice) are exactly the partially measurable band of Fig. 3 (the shaded span in Fig. 7). The disagreement is systematic rather than noisy: of the 193 discordant questions, 178 are Gemini-autonomous / Qwen-audit versus only 15 in the opposite direction, i.e. Gemini is uniformly less willing to take the tool prior on trust and more often explores the volume from scratch. Since SVRP supports both routes, this bias compromises neither correctness nor the protocol's guarantees.

Routing drift does not erode the gain. Despite Gemini's lower overall audit fraction (0.384 vs 0.578), the Gemini instantiation still improves over its 2D baseline on every subtask and attains

the best DeepTumorVQA average (Tab. 1). The benefit of SVRP is therefore robust to backbone-induced routing variation rather than contingent on a particular routing distribution.

C.4 Intra-Subtype Routing Variation Analysis

Figure 3 shows that several subtypes split across both the audit and autonomous routes. The general reason is that routing is determined case by case by whether the decisive evidence for answering the question is already captured by the available tool prior. Whenever this decisive evidence varies across instances within the same subtype, that subtype will naturally mix the two routes. If the prior contains reliable, answer-critical signal, VLM can resolve the case through the audit route by re-grounding and verifying that signal directly against the image. If the prior is missing, incomplete, or not trustworthy for the specific case, VLM instead routes to autonomous exploration, where the model searches for the needed evidence through multiple independent inspection paths. This explains the subtype-level mixtures in Fig. 3. In our setting, lesion masks are one concrete form of tool prior. Current segmentation models remain less accurate and less stable at the lesion level than at organ level. Therefore, subtypes whose decisive evidence is often well captured by the mask tend to favor audit, whereas subtypes for which the mask often lacks reliable lesion-level signal tend to favor autonomous exploration. For example, focal kidney findings are often segmented reliably enough for the overlay to be decisive, whereas colon lesions, PDAC, PNET more often lack usable lesion-level signal in the mask and therefore route predominantly to autonomous exploration.

D Limitations

Med3DNavi is a multi-stage agent, so its active exploration requires more model and tool invocations than single-shot VLM inference. Although we cap the number of paths and viewer calls to keep the interaction cost bounded, latency and API cost remain higher, which may limit deployment in time-sensitive settings. In addition, the audit route and mask-overlay functionality depend on upstream anatomical segmentation (VISTA3D in our experiments). When segmentation is unavailable, out of label scope, or inaccurate, the model cannot obtain reliable quantitative priors or overlays and must fall back to the prior-free autonomous route; lesion-level priors are even less consistently available and reliable. Finally, our experiments are limited to two tumor-centric CT benchmarks. While the framework is in principle modality-agnostic, its generalization to other volumetric modalities, broader anatomies, real clinical workflows, and different backbone capabilities remains to be validated.