MedAura: Native 3D–2D Evidence-Seeking Agents for CT Reasoning

Anonymous ACL submission

Abstract

Computed Tomography (CT) reporting is an evidence-seeking task: a reader forms a volumetric impression, inspects slice-level evidence, and decides whether each finding is sufficiently supported. Static-input 3D medical Vision-Language Models (VLMs) improve M3D-Cap report-style captioning and M3D-VQA, while recent medical agents show the value of tool use. We study a different supervision target: a CT model that requests evidence from the same volume, tracks whether the returned slices support each claim, and decides whether to answer, query again, or withhold.

We present MEDAURA, a supervised CT evidence-seeking framework that combines compact 3D memory with on-demand 2D slice inspection. At inference, the model requests a plane and range, receives a fixed same-volume slice packet, records evidence sufficiency and cited slices, then produces a final answer or asks for more evidence. Training separates claims: Stage 2A builds a single-pass task foundation, improving MedGemma by +28.9 points in M3D-VQA close accuracy and +7.4 points in M3D-Cap METEOR; Stage 2B tests packet sensitivity; and Stage 3 trains the evidence-seeking loop, yielding 21,951 GPT-audited trajectory turns for release upon acceptance. Stage 3 makes the evidence path auditable but remains below Stage 2A on long-form M3D-Cap, indicating that future CT agents need better finding-level slice coverage and stronger preservation of single-pass task strength inside multi-turn evidence seeking. MEDAURA contributes an auditable CT-agent interface, trajectory data, and a diagnostic account of why valid tool use alone does not yet solve grounded CT report-style generation.

1 Introduction

Computed Tomography (CT) reporting is not only recognition; it is evidence accounting. A report may contain findings across organs, planes, and locations; a useful CT model must therefore decide not only *what* to report, but also *which slices* support each claim and *when* the cited evidence is sufficient. This motivates a CT agent that keeps compact volumetric context, requests local same-volume slice evidence, and stops only when the cited evidence covers the findings.

Existing work covers parts of this workflow. Fixed-input 3D medical VLMs compress volumes and improve static-input M3D-Cap report-style captioning and M3D-VQA performance, and 3D–2D architectures show that preserving planar slice evidence helps volumetric language tasks (Bai et al., 2024; Xin et al., 2025; Shi et al., 2024; Lian et al., 2026). Throughout, CAP denotes the M3D-Cap benchmark split and VQA denotes M3D-VQA unless otherwise noted. MedGemma 1.5 4B provides a capable native multi-slice 2D image route (Sellergren et al., 2026). Separately, medical imaging agents such as VoxelPrompt, CT-Agent, RadAgent, MedVistaGym, and MedAgent-Pro show that tools and stepwise reasoning can be useful in medical imaging and diagnosis (Hoopes et al., 2024; Mao et al., 2025; Roschewitz et al., 2026; Lu et al., 2026; Wang et al., 2025). Our comparison with these systems focuses on the interface and supervision paradigm rather than a head-to-head task leaderboard: can a single CT decoder be trained under a structured protocol that links compact 3D memory, model-driven same-volume slice requests, returned 2D packets, evidence sufficiency, and explicit routing decisions (final/query-more/withhold)?

MEDAURA studies this question as supervised model behavior rather than a prompted tool wrapper. The model receives Active3D volume slots, chooses a plane/range through `QUERY_SLICES`, observes a fixed top-3 same-volume 2D packet, and writes an `EvidenceState` with support, sufficiency, cited slices, and next action. The runtime owns reslicing and packet construction; the model owns

uncertainty, evidence use, and routing.

We evaluate in stages because task accuracy and agent capability support different claims: Stage 2A is the single-pass task foundation, Stage 2B is a controlled 3D+2D packet-sensitivity gate, and Stage 3 trains the native tool loop. Table 1 reports the scale of each training signal. Stage 3 yields 21,951 GPT-audited trajectory turns over 8,121 trajectories plus 8,255 sidecar audit rows, which we will release publicly upon acceptance. The planned release is not ordinary M3D-Cap/M3D-VQA SFT or a free-form chain-of-thought corpus: it records tool calls, runtime observations, `EvidenceState` updates, route labels, supported final answers, and sidecar audit rows. Appendix A includes a representative M3D-VQA-derived release row in Table 8.

Our empirical findings are deliberately bounded. Stage 2A is the strongest MedAura task configuration, while Stage 3 learns the tool interface and shows controlled packet sensitivity but does not close the long-form M3D-Cap gap. The Stage 3-to-Stage 2A M3D-Cap gap is a localized task-metric shortfall rather than the task-line claim for Stage 3: Stage 3 supervises the tool-loop interface, not single-pass M3D-Cap. Counterfactual packet-sensitivity tests and judge diagnostics are consistent with finding-level visual coverage, rather than tool syntax alone, as the main remaining bottleneck.

Our contributions are:

- We formulate CT evidence seeking as auditable evidence accounting: compact 3D memory, same-volume slice requests, cited evidence, sufficiency decisions, and final/query-more/withhold routing, supervised as a single-decoder target.
- We will publicly release a Stage 3 `EvidenceState` trajectory dataset for CT evidence acquisition upon acceptance, with 21,951 accepted turns and 8,255 sidecar audit rows over 8,121 trajectories, plus provenance keys for auditing.
- We provide a staged evaluation separating task performance from agent behavior: Stage 2A is the strongest MEDAURA task line, Stage 2B supports controlled packet sensitivity, and Stage 3 learns the interface but does not solve long-form M3D-Cap.
- We diagnose the Stage 3-to-Stage 2A M3D-Cap gap, motivating future work on preserving single-pass task strength inside the multi-turn loop and stopping only when cited slices cover the report findings.

2 Related Work

3D medical VLMs and volumetric representation. M3D-LaMed, Med3DVLM, Med-2E3, and data-efficient 2D-encoder 3D VLMs establish fixed-input M3D-Cap/M3D-VQA performance as an important baseline family; newer 3D–2D architectures further show the value of preserving planar slice evidence before compression (Bai et al., 2024; Xin et al., 2025; Shi et al., 2024; Lian et al., 2026). MEDAURA studies a different interface: same-volume slice inspection is a model action, not only a preselected input.

Radiology report generation and grounded reporting. Report-generation work improves long-form generation through memory, cross-modal alignment, token-imbalance handling, observation-level structure, grounded reporting, and factual evaluation (Chen et al., 2020, 2021; Wu et al., 2023; Zhou and Wang, 2024; Bannur et al., 2024; Ostmeier et al., 2024; Nguyen et al., 2026). These systems improve generation from available evidence; MEDAURA studies evidence acquisition before finalization.

Tool-using and medical imaging agents. ReAct, Toolformer, MM-ReAct, and ViperGPT motivate reasoning with actions or executable visual programs (Yao et al., 2023; Schick et al., 2023; Yang et al., 2023; Suris et al., 2023). Medical imaging agents demonstrate that tool use and stepwise reasoning can improve medical image analysis, CT QA/reporting, and multimodal diagnosis (Hoopes et al., 2024; Mao et al., 2025; Roschewitz et al., 2026; Lu et al., 2026; Wang et al., 2025). MEDAURA is not claimed to outperform these systems head-to-head; instead, it studies a different interface and supervision target: same-volume CT plane/range requests, fixed returned slice packets, EvidenceState sufficiency fields, and final/query-more/withhold routing as supervised trajectory data. Appendix A.1 summarizes the contract differences.

Medical visual reasoning and grounding-aware evaluation. Recent work in medical reasoning and reinforcement learning (Pan et al., 2025; Lai et al., 2025; Jiang et al., 2026) alongside grounding-aware evaluation (Le-Duc et al., 2025; Zafar et al.,

2026; Nguyen et al., 2025) shows that accuracy can mask weak visual dependence. MEDAURA follows this lesson by reporting task metrics separately from packet-sensitivity, counterfactual, tool-validity, and EvidenceState diagnostics.

3 Method

3.1 Backbone and Token Interface

MEDAURA is built on MedGemma 1.5 4B, a Gemma 3 medical VLM with a native medical SigLIP image pathway (Sellergren et al., 2026). All reported data and evaluation are CT. We add a 24.6M-parameter Active3D projector, yielding a 4.33B-parameter model, and train Stage 3 with a 29.8M-parameter LM LoRA while freezing DCFormer, the projector, the native 2D tower, embeddings, and LM head (Appendix C).

3.2 Native 3D+2D Architecture

Given a CT volume V and query q, MEDAURA uses a compact 3D route plus native 2D slice observations. DCFormer features initialized from Med3DVLM-style weights (Xin et al., 2025) are projected into 256 Active3D vectors:

$$T_{3D} = P_\theta(H_{\text{low}}, H_{\text{high}}) \in \mathbb{R}^{256\times d}. \quad (1)$$

These vectors replace the embedding states at `<3d_volume_slot>` positions before decoder attention, making the 3D route an in-decoder token-replacement interface rather than an external classifier head. Projector details are in Appendix C.

When the agent calls `QUERY_SLICES`, the runtime returns selected same-volume 2D slices through MedGemma's native image-token pathway. For a later turn t with retrieved packet $\mathcal{P}_t$, the decoder input is

$$X_t = [T_{\text{hist}};\, T_{3D};\, E_{2D}(\mathcal{P}_t);\, T_{\text{obs/evid}}]. \quad (2)$$

This late fusion combines task text, tool history, compact 3D memory, native 2D image tokens, and evidence-state text. Each packet adds three slices (768 local 2D visual tokens) around the 256-token global 3D memory, preserving MedGemma's local reader while making the visual bottleneck explicit: Stage 3 can inspect only the evidence it asks the runtime to retrieve. The compact 3D route is for global planning; long-form M3D-Cap outputs may require finding-level evidence that is not covered by a few top-3 packets or preserved in compressed 3D semantics.

Algorithm 1 MedAura evidence-seeking inference contract

Require: CT volume V, user query q, maximum turns M
$T_{3D} \leftarrow \text{Active3D}(\text{DCFormer}(V))$
$h_0 \leftarrow [q; T_{3D}]$
Step-0 model prior: $y_0 \leftarrow \text{LM}(h_0)$
parse `QUERY_SLICES` (p_0, s_0, e_0) from y_0
validate $p_0 \in \{\text{axial}, \text{coronal}, \text{sagittal}\}$ and $0 \le s_0 < e_0 \le 1$
if validation fails **then**
  return `INVALID_QUERY` sidecar
end if
for $t = 1, \ldots, M$ **do**
  Runtime step: $\mathcal{C}_t \leftarrow \text{Candidates}(V, p_{t-1}, s_{t-1}, e_{t-1})$
  $\mathcal{P}_t \leftarrow \text{Top3}_{i\in\mathcal{C}_t}\, S(i)$
  $o_t \leftarrow \text{Observe}(p_{t-1}, \mathcal{P}_t)$
  update history $h_t \leftarrow [h_{t-1}; y_{t-1}; o_t; \mathcal{P}_t]$
  Model step: $y_t \leftarrow \text{LM}(h_t)$
  if y_t contains `final_answer` or `withhold` **then**
    return y_t
  end if
  parse `QUERY_SLICES` (p_t, s_t, e_t) from y_t
  validate $p_t \in \{\text{axial}, \text{coronal}, \text{sagittal}\}$ and $0 \le s_t < e_t \le 1$
  if validation fails **then**
    return `INVALID_QUERY` sidecar
  end if
end for
return `MAX_TURNS_REACHED` sidecar

3.3 Tool Loop

The action space contains one tool:

$$\begin{aligned} a_t &= \text{QS}(p_t, s_t, e_t), \\ p_t &\in \{\text{axial}, \text{coronal}, \text{sagittal}\}, \\ 0 &\le s_t < e_t \le 1. \end{aligned} \quad (3)$$

Here QS denotes `QUERY_SLICES`. The model chooses the plane/range and, in the EvidenceState line, a short retrieval intent. The runtime canonical-reslices the same volume, ranks up to 24 candidates with MedSigLIP relevance plus history-aware novelty penalties, and returns the three highest-scoring distinct slices:

$$\mathcal{P}_t = \text{Top3}_{i\in\mathcal{C}_t}\, S(i), \qquad |\mathcal{P}_t| = 3, \quad (4)$$

where $S(i)$ combines MedSigLIP relevance with range-overlap, reuse, and redundancy penalties. We fix $k = 3$; Appendix E gives the full score and Table 35 reports the $k = 6$ diagnostic.

The agent loop has two modes. Step 0 uses only the query and Active3D slots to produce a cautious prior and first tool call. Later turns receive history, observation metadata, retrieved 2D images, and Active3D slots; they then answer, query again, or withhold.

Step 0 must emit the first query; final/withhold routing begins only after a runtime observation. Ob-

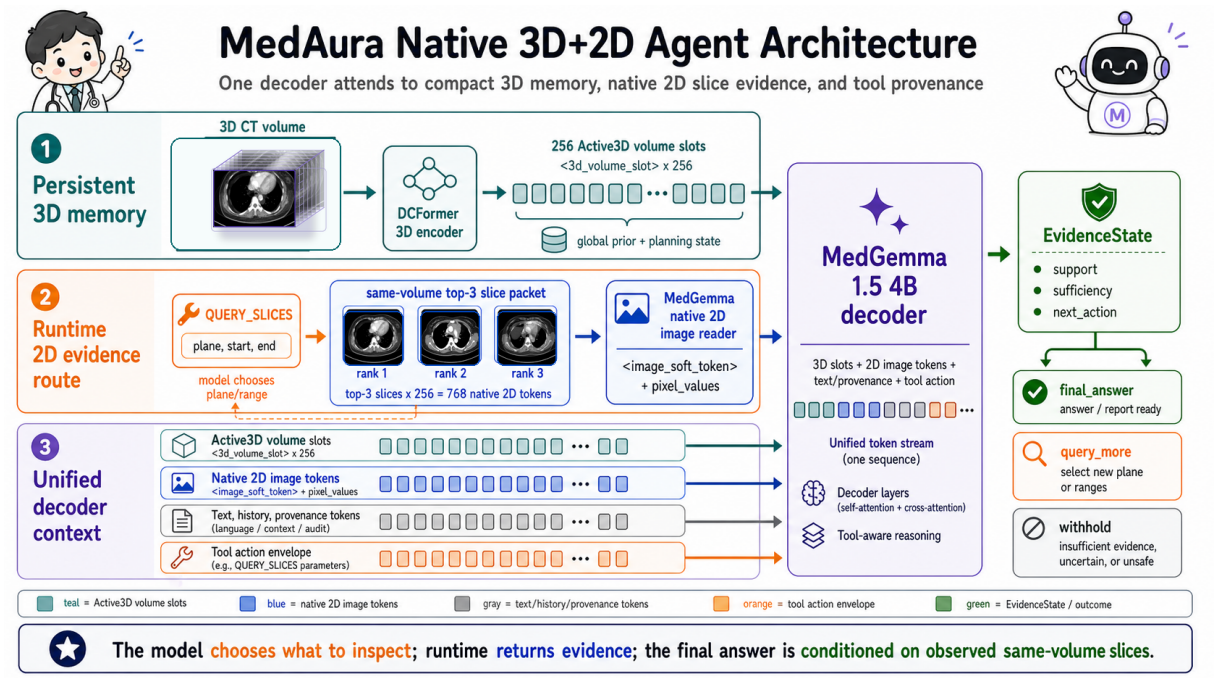

Figure 1: MEDAURA architecture. DCFormer volume features provide persistent Active3D slots, retrieved CT slices use MedGemma's native image pathway, and text/history/evidence state enter the same decoder sequence. The CT thumbnails are reconstructed from stored qualitative runtime provenance. Colored small blocks denote token-route classes; purple denotes the decoder module/background, not an additional tensor route.

servations and images are visible context, not supervised targets; supervised targets are `QUERY_SLICES` actions, EvidenceState fields, final answers, query-more actions, and withheld decisions. Evaluation-only forced-final recovery is never mixed into positive Stage 3 SFT rows.

3.4 Epistemic Step Decomposition

We supervise the 3D prior separately from evidence-conditioned updates. Step 0 contains uncertainty, retrieval intent, and the first tool call; later turns add packet review, slice-cited evidence, route decision, and EvidenceState. The invariant is that global 3D context suggests where to look, but final claims depend on observed evidence. This matters for M3D-Cap because one caption may contain many findings and each finding may require different slice evidence.

3.5 Stage-wise Training

Training is staged to separate claims. Stage 1 aligns the Active3D socket on M3D-Cap; Stage 2A trains the single-pass M3D-Cap/VQA foundation; Stage 2B trains controlled 3D+native 2D packet-sensitivity rows; Stage 3 initializes from Stage 2B and trains turn-level tool-loop trajectories with tool calls, `EvidenceState`, final answers, query-more actions, or withholding.

All supervised stages use the standard autoregressive language-modeling objective, but Stage 3 uses route-aware token weights:

$$\mathcal{L}_{\mathrm{SFT}} = -\frac{\sum_{j\in\mathcal{J}} \alpha_j \log p_\theta(y_j \mid x, y_{<j})}{\max(1, \sum_{j\in\mathcal{J}} \alpha_j)}. \quad (5)$$

Here $\mathcal{J}$ is the set of assistant target positions after the one-token autoregressive shift; prompt tokens, padding, runtime observations, and other `-100` labels are excluded from the loss. The weight α_j is determined by the row tag and span type: Step 0 boosts tool syntax, `QUERY_SLICES`, plane, and range tokens; query-more turns boost evidence-state and next-action spans; final turns boost evidence-state, `<final_answer>`, and VQA option prefixes. This makes the training objective match the deployed agent contract rather than treating all generated tokens as equally important prose. The separation is deliberate: Stage 2A is evaluated as the task-level foundation in Section 5.2, Stage 2B as a packet-sensitivity gate in Section 5.3, and Stage 3 as the native tool-loop interface rather than as a direct leaderboard replacement.

4 Data Construction

4.1 Source Tasks

We use M3D-Cap for CT report-style captioning and M3D-VQA for visual question answer-

Line	Source	Scale	Target / role
Stage 1	M3D-Cap	116,092 source rows	report-style caption text; Active3D socket alignment
Stage 2A	M3D tasks	105,824 CAP + 935,692 VQA; 317,472 sampled/epoch	CAP captions/VQA answers; task foundation
Stage 2B	same-volume 3D+2D	33,012 total: 16,506 real / 16,506 corrupted	evidence object, sufficiency/citation; packet-sensitivity diagnostic
Stage 3	audited traces	21,951 turns; 8,121 trajectories; 8,255 sidecars; 11,677 CAP / 10,274 VQA	EvidenceState tool/route; agent data

Table 1: Training signals and scale used by the reported MedAura lines. The rows are intentionally not interchangeable: Stage 2A is the benchmark foundation, Stage 2B is a controlled 3D+2D packet-sensitivity diagnostic, and Stage 3 is the native EvidenceState agent training data. For Stage 2A, 317,472 sampled/epoch denotes the balanced training mixture rather than the total available rows. Detailed lineage and data-construction variants are recorded in the appendix.

Span	Producer	Visible	Loss	Purpose
Active3D slots	encoder	yes	no	global 3D context
Step 0 thought	teacher	yes	yes	cautious prior and uncertainty
`QUERY_SLICES` call	teacher	yes	yes	evidence-seeking action
obs./images	runtime	yes	no	environment feedback
evidence state	teacher	yes	yes	provenance and sufficiency
final answer	audited teacher	yes	yes	M3D-Cap text or VQA answer

Table 2: Stage 3 data-contract boundary. Runtime observations and retrieved images are input context, not targets; the model is trained to decide how to use them through tool actions, EvidenceState fields, and final/query-more outputs.

ing about plane, phase, organ, abnormality, and location (Bai et al., 2024). Stage 1–2A use the original benchmark supervision: a volume paired with a report-style caption or answer. Stage 2B and Stage 3 add evidence-seeking supervision: the requested evidence, returned packet, sufficiency, and final/query-more/withhold route.

The MEDAURA Stage 3 EvidenceState trajectory dataset will be made public upon acceptance. The planned release contains accepted train-ready trajectory rows, sidecar audits, metadata, and provenance keys rather than repackaged CT volumes. Stage 2B and Stage 3 rows are built from non-test M3D source records; benchmark evaluation uses held-out M3D-Cap `test1k` and M3D-VQA `test5k` records. No supervision rows are derived from the reported test splits. Released rows carry provenance keys; our enforceable leakage boundary is source-split plus volume/sample-key separation (details in Appendix A).

These counts come from the final training audits. Stage 2B real rows pair live Active3D volume slots with native 2D slice packets from the same source volume. The training set then adds one deterministic corrupted counterpart per real row, cycling across shuffled 2D packets, removed cited slices, and zeroed 2D evidence. The Stage 2B numbers are therefore packet-sensitivity training counts, not benchmark evaluation counts.

4.2 Unified Stage 3 Teacher Protocol

The Stage 3 teacher is a data-construction pipeline, not a deployment-time dependency. From a benchmark CT case, it writes a cautious 3D prior and `QUERY_SLICES` action; the runtime reslices the same volume and returns a fixed top-3 packet; the teacher then emits an EvidenceState plus final, query-more, or withhold.

The reference target (report or answer) is visible during offline construction and audit; the student never sees it at deployment. Thus Stage 3 is teacher-forced interface data, not evidence of autonomous evidence discovery or a clinical policy: target tool calls, support labels, and final answers can be influenced by the reference, while the student must learn the routing behavior without that reference. A fixed OpenAI-compatible multimodal backend (temperature 0, high-detail images) and deterministic validators check format, citation, and route validity; valid turns become SFT rows and rejected turns become sidecars. The pipeline used approximately 196.6M API tokens, and the released rows and sidecars avoid rerunning the teacher (Appendix I).

The released EvidenceState line yields 21,951 supervised turns over 8,121 trajectories plus 8,255 sidecars (mean/max 1.70/4 tool queries; lineage in Appendix F).

5 Experiments

5.1 Experimental Setup and Compared Lines

We evaluate three claims separately. Benchmark metrics test final M3D-Cap/M3D-VQA quality; packet-sensitivity diagnostics test sufficiency and citation behavior under controlled corruption; agent diagnostics test whether a model can emit `QUERY_SLICES` calls, consume observations, update EvidenceState, query more, answer, or withhold. M3D-Cap uses BLEU-1, ROUGE-1, METEOR, and BERTScore-F1; VQA uses open-ended lexical/semantic metrics and close-ended option accuracy, following M3D/Med3DVLM (Bai et al., 2024; Xin et al., 2025). Compared lines in-

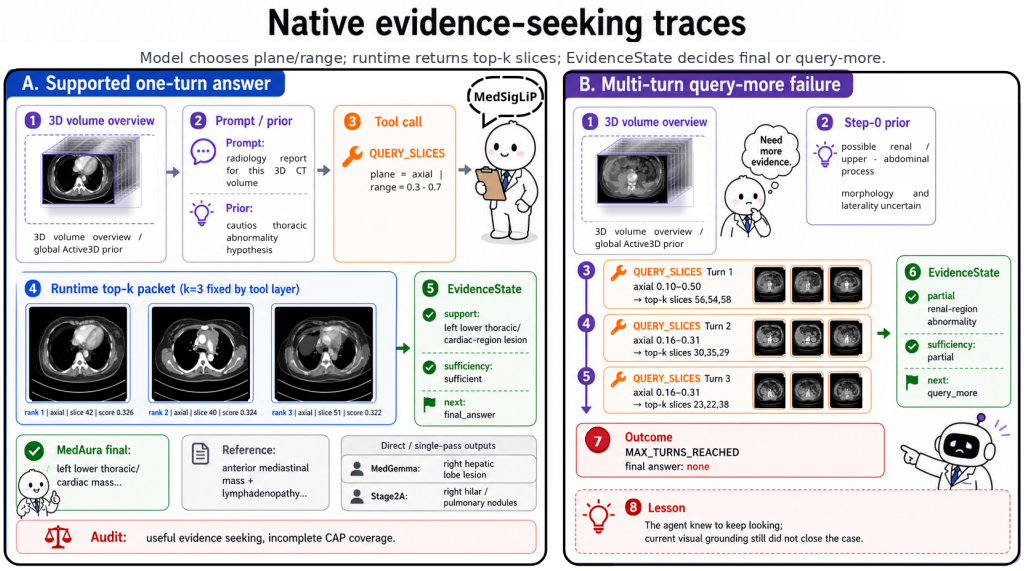

Figure 2: Native evidence-seeking traces reconstructed from frozen runtime provenance. Panel A shows a one-turn supported answer: the model requests an axial range, the runtime returns fixed top-3 same-volume CT slices, and EvidenceState routes to a final answer. Panel B shows a query-more failure: the EvidenceState marks the evidence as incomplete and keeps querying, but the available packets do not close the case before the turn limit. The figure illustrates traceable agent behavior and the finding-coverage bottleneck; it is not a claim of perfect M3D-Cap output.

clude MedGemma 1.5 4B IT, Active3D Stage 2A, Stage 3 RegionHistory, Stage 3 EvidenceState, M3D-LaMed, Med3DVLM, and the data-efficient 2D-encoder 3D VLM (Bai et al., 2024; Xin et al., 2025; Lian et al., 2026). Appendix A–I gives per-type metrics, ablations, counterfactuals, and judge protocols.

Single-run caveat. All training lines are single-run due to the 196.6M-token teacher construction cost and multi-stage training pipeline. We therefore interpret Stage 3 deltas as diagnostic directions rather than statistically validated improvements, and restrict our claims accordingly. Small deltas and CAP200/VQA200 subsets are diagnostic, not statistically powered rankings.

5.2 Task-Level M3D-Cap/M3D-VQA Results

Table 3 reports only final-answer benchmark performance; agent capability is tested separately in Section 5.3.

Stage 2A is the highest-scoring MedAura task line, improving MedGemma by 28.94 VQA-close points and 7.39 M3D-Cap METEOR. Stage 3 should be read as an interface/data line: EvidenceState improves over RegionHistory on M3D-Cap overlap and VQA-close, but remains below Stage 2A despite additional local 2D packets. This localized gap is informative because it cannot be explained by lack of 2D visual input alone.

5.3 Packet-Sensitivity and Native Agent Diagnostics

Table 4 reports controlled packet-sensitivity and agent-readiness diagnostics. Stage 2B uses a held-out 512-record split, separate from Table 1, contrasting 256 real same-volume 3D+2D packets with 256 corrupted packets that shuffle, zero, or remove visual support. The resulting sufficiency and citation changes indicate packet sensitivity, not complete clinical grounding. Stage 3 then extends this gate into a multi-turn EvidenceState loop; runtime-transfer stress tests show that prompting non-agent baselines with the same format rarely yields a functional CT agent in our setup.

5.4 Result Synthesis and Bottleneck Analysis

Stage 2A provides the highest final M3D-Cap/M3D-VQA scores among MedAura lines, whereas Stage 3 operates the native interface and yields the released trajectories. We therefore use the localized Stage 3-to-Stage 2A M3D-Cap gap diagnostically: once tool syntax works, what still prevents accurate long-form reporting?

Appendix G tests three plausible explanations. Prompt transfer rarely completes the native loop, while Stage 3 parses tool syntax, observations, and EvidenceState routing (Table 33). Packet variants such as $k = 6$, soft-range fallbacks, and frozen-2D reranking do not close the Stage 2A gap in our diagnostic evaluations (Table 35). Counterfactual

Line	Size / class	CAP B1	CAP R1	CAP M	CAP BERT	VQA-open M	Open BERT	VQA-close
MedGemma 1.5 4B IT (Sellergren et al., 2026)	4.30B	16.32	17.15	11.86	81.60	9.09	81.91	50.92
Active3D Stage 2A (ours)	4.33B	25.24 (+8.92)	24.62 (+7.47)	19.25 (+7.39)	84.71 (+3.11)	36.69 (+27.60)	91.94 (+10.03)	79.86 (+28.94)
Stage 3 RegionHistory (ours)	4.33B	6.30 (-10.02)	13.50 (-3.65)	7.39 (-4.47)	83.67 (+2.07)	23.62 (+14.53)	87.03 (+5.12)	61.30 (+10.38)
Stage 3 EvidenceState (ours)	4.33B	11.73 (-4.59)	17.49 (+0.34)	11.72 (-0.14)	83.04 (+1.44)	21.64 (+12.55)	86.73 (+4.82)	62.08 (+11.16)
M3D-LaMed (Bai et al., 2024)	6.9B	15.15	19.55	14.38	88.46	33.58	91.53	75.78
Med3DVLM (Xin et al., 2025)	7.6B	36.88	40.25	36.42	88.11	36.76	92.18	79.75
2D-Encoder 3D VLM (Lian et al., 2026)	7B-class n/r	**50.26**	**53.37**	**50.13**	**89.98**	**44.46**	**94.73**	**82.90**

Table 3: Task-level M3D-Cap/M3D-VQA performance. Stage 2A is the highest-scoring MedAura task-level line; Stage 3 rows train native tool-loop behavior and release trajectory data, but are not claimed as task-metric improvements over Stage 2A. CAP and open-ended VQA metrics are percentage-style values; VQA-close is accuracy in percent. Signed deltas are absolute differences from MedGemma 1.5 4B IT (green positive, red negative). The size column uses exact reported totals when available and backbone class otherwise; n/r denotes exact total not reported. Full per-type benchmark and architecture tables are in Appendix B.

Diagnostic	Value	Interpretation
Stage 2B evidence-object rate	0.998	emits controlled 3D+2D evidence object
Stage 2B final-answer rate	0.992	preserves terminal answer format
Stage 2B target sufficiency acc.	0.938	accepts matching same-volume evidence
Stage 2B corrupted suff. drop	1.000	marks corrupted packets insufficient
Stage 2B corrupted citation change	0.996	changes cited evidence under corruption
Stage 3 EvidenceState parse rate	1.000	parses the runtime grammar
Stage 3 native-runtime CAP200 success	0.745	operates the loop without prompt transfer
MedGemma-runtime CAP200 success	0.040	prompt transfer is weak in this setup

Table 4: Compact packet-sensitivity and agent-readiness diagnostics. These are not leaderboard M3D-Cap/M3D-VQA scores; they test evidence sensitivity and native runtime execution. Stage 2B diagnostics are computed from 512 controlled gate records rather than benchmark splits. The Stage 3 native-runtime CAP200 success row is from the runtime-transfer stress test with forced-final recovery disabled, so it should not be compared directly with the reported intent-guided packet-selector diagnostic in Appendix G.

packets alter sufficiency judgments, citations, and outputs, making a purely image-invariant account unlikely (Table 42).

The remaining gap is therefore most consistent with a combination of incomplete finding-level visual coverage, compact-token semantics, text priors, difficulty preserving Stage 2A's single-pass score inside the multi-turn loop, runtime-selection limits, and teacher-forced policy mismatch. Stage 3 has additional local 2D packets yet still falls below Stage 2A, including on VQA categories such as plane recognition, where local evidence should be less limiting (Plane close accuracy drops from 99.20 to 89.80; Tables 10– 11). Long-form M3D-Cap outputs contain multiple findings, while the current agent keeps 256 global 3D tokens and inspects only a few local packets. Because targets are visible during offline construction, teacher trajectories can also encode routes that are harder to reproduce without the reference at inference time. The current 4.33B model is smaller than several 7B-class 3D medical VLM references (Table 13), so we do not interpret the gap as a pure failure of agency. Appendix Tables 15 and 17 summarize this bottleneck.

6 Discussion

The result is a CT evidence-seeking agent interface, not a claim that the loop alone solves long-form M3D-Cap. MedAura keeps MedGemma's native 2D reader because local understanding is strong, but Stage 3 exposes it through a limited sequence of selected packets. The appendix ablations support this reading: tool syntax and EvidenceState parsing are learnable, but packet widening, soft-range fallbacks, and frozen-2D reranking do not recover Stage 2A-level CAP (Tables 31 and 17). The released trajectories provide an audited scaffold for testing stronger visual foundations under the same contract. Future work should prioritize 2D-derived volumetric enrichment, and rewards that stop the query loop only when cited slices fully cover the report findings.

7 Conclusion

We presented MEDAURA, a MedGemma-based native 3D+2D CT agent with Active3D slots, native 2D slice reading, and a supervised evidence loop. Stage 3 contributes auditable tool calls, EvidenceState updates, sidecars, and traceable same-volume packets. The loop is learnable, but grounded M3D-Cap performance remains constrained by incomplete cited-slice coverage of the findings, and difficulty preserving single-pass task strength inside the multi-turn loop.

8 Limitations

The action space contains only `QUERY_SLICES`, and top-3 selection is delegated to a frozen MedSigLIP runtime. Clinically useful systems would need localization, segmentation, patient-history tools, prospective validation, and human oversight. Released provenance exposes source split and volume/sample keys, but not patient identifiers; our leakage boundary is source-split plus key separation. Stage 3 trajectories are GPT-audited rather than radiologist-validated, and the low absolute judge scores (Appendix I) underscore that no system here should be treated as clinically usable. Reference-audited teacher forcing may encode routes easier for the offline teacher than for the deployed student, and the same model alias is used for construction and judging, so scores are diagnostic rather than independent validation. We provide interface/supervision contrasts to recent medical-agent systems rather than direct task-score comparisons, because their tool APIs, runtimes, datasets, and clinical scopes differ from the MedAura runtime. Results are single-run diagnostics; small ablation deltas, wall-clock cost, and GPU-memory comparisons are not claimed.

Ethical Considerations

MEDAURA is a research system, not an autonomous clinical tool. Errors in CT interpretation, slice retrieval, EvidenceState reasoning, or final generation can produce unsafe conclusions; deployment would require prospective validation, clinical oversight, uncertainty reporting, and human-in-the-loop safeguards. The released trajectories contain model-generated evidence states and slice selections, not radiologist-validated clinical ground truth. M3D-Cap/M3D-VQA and medical pretraining corpora may also reflect demographic, geographic, institutional, and acquisition-protocol biases; performance should not be assumed to generalize without fairness and external-site evaluation.

Data Availability

Upon acceptance, we will publicly release the Stage 3 EvidenceState trajectory dataset, including train-ready trajectory JSONL, sidecar audit JSONL, manifests, schema/EDA metadata, provenance keys, and abbreviated CAP/VQA sample rows. The release will not include raw CT volumes, images, or patient identifiers. Appendix A includes a representative M3D-VQA-derived row in Table 8. The release is intended for agent-interface learning and auditing, not as radiologist-validated clinical ground truth.

Code Availability

Upon acceptance, we will publicly release the research code for the MEDAURA training and evaluation pipeline, including the Active3D interface, `QUERY_SLICES` runtime, trajectory validators, packet-sensitivity diagnostics, and scripts used to reproduce the reported tables from the released artifacts. Repository links are omitted from the submission version to preserve double-blind anonymity.

References

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*.

Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Perez-Garcia, Valentina Salvatelli, Harshita Sharma, Felix Meissen, Mercy Ranjit, Shaury Srivastav, Julia Gong, Noel C. F. Codella, Fabian Falck, Ozan Oktay, Matthew P. Lungren, Maria Teodora Wetscherek, and 2 others. 2024. MAIRA-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 2021. Cross-modal memory networks for radiology report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 5904–5914. Association for Computational Linguistics.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 1439–1449. Association for Computational Linguistics.

Andrew Hoopes, Victor Ion Butoi, John Guttag, and Adrian V. Dalca. 2024. VoxelPrompt: A vision-language agent for grounded medical image analysis. *arXiv preprint arXiv:2410.08397*.

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

Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, and Xiaofeng Yang. 2025. Med-R1: Reinforcement learning for generalizable medical reasoning in vision-language models. *arXiv preprint arXiv:2503.13939*.

Khai Le-Duc, Duy M. H. Nguyen, Phuong T. H. Trinh, Tien-Phat Nguyen, Nghiem T. Diep, An Ngo, Tung Vu, Trinh Vuong, Anh-Tien Nguyen, Mau Nguyen, Van Trung Hoang, Khai-Nguyen Nguyen, Hy Nguyen, Chris Ngo, Anji Liu, Nhat Ho, Anne-Christin Hauschild, Khanh Xuan Nguyen, Thanh Nguyen-Tang, and 5 others. 2025. S-chain: Structured visual chain-of-thought for medicine. *arXiv preprint arXiv:2510.22728*.

Yi Lian, Yanchun Xie, Yang Jiang, Lei Wang, and Hailong Yu. 2026. A data-efficient 3d medical vision-language model using only a 2d encoder. *Scientific Reports*, 16:8809.

Meng Lu, Yuxing Lu, Yuchen Zhuang, Megan Mullins, Yang Xie, Guanghua Xiao, Charles Fleming, Wenqi Shi, and Xuan Wang. 2026. MEDVISTAGYM: A scalable training environment for thinking with medical images via tool-integrated reinforcement learning. *arXiv preprint arXiv:2601.07107*.

Yuren Mao, Wenyi Xu, Yuyang Qin, and Yunjun Gao. 2025. CT-Agent: A multimodal-LLM agent for 3d CT radiology question answering. *arXiv preprint arXiv:2505.16229*.

Cong Huy Nguyen, Son Dinh Nguyen, Guanlin Li, Tuan Dung Nguyen, Aditya Narayan Sankaran, Mai Huy Thong, Thanh Trung Nguyen, Mai Hong Son, Reza Farahbakhsh, Phi Le Nguyen, and Noel Crespi. 2026. Region-grounded report generation for 3d medical imaging: A fine-grained dataset and graph-enhanced framework. *arXiv preprint arXiv:2604.18145*.

Tuan Dung Nguyen, Cong Huy Nguyen, Nguyen Ho, Aditya Narayan Sankaran, Satwik Ramchandre, Thanh Trung Nguyen, Reza Farahbakhsh, Phi Le Nguyen, and Noel Crespi. 2025. Localizing before answering: A benchmark for grounded medical visual question answering. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 7670–7678. IJCAI.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson, Michael Moseley, Curtis Langlotz, Akshay S. Chaudhari, and Jean-Benoit Delbrouck. 2024. GREEN: Generative radiology report evaluation and error notation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 374–390. Association for Computational Linguistics.

Jiazhen Pan, Che Liu Ye, Junde Zhang, Junzhi Chen, Zhuoran Zhao, Alison Harrison, Shishir Nair, Ben Glocker, Daniel Rueckert, and Georgios Kaissis. 2025. MedVLM-R1: Incentivizing medical reasoning capability of vision-language models via reinforcement learning. *arXiv preprint arXiv:2502.19634*.

Mélanie Roschewitz, Kenneth Styppa, Yitian Tao, Jiwoong Sohn, Jean-Benoit Delbrouck, Benjamin Gundersen, Nicolas Deperrois, Christian Bluethgen, Julia Vogt, Bjoern Menze, Farhad Nooralahzadeh, Michael Krauthammer, and Michael Moor. 2026. RadAgent: A tool-using AI agent for stepwise interpretation of chest CT. *arXiv preprint arXiv:2604.15231*.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*.

Andrew Sellergren, Chufan Gao, Fereshteh Mahvar, Timo Kohlberger, Fayaz Jamil, Madeleine Traverse, Alberto Tono, Bashir Sadjad, Lin Yang, Charles Lau, Liron Yatziv, Tiffany Chen, Bram Sterling, Kenneth Philbrick, Richa Tiwari, Yun Liu, Madhuram Jajoo, Chandrashekar Sankarapu, Swapnil Vispute, and 23 others. 2026. MedGemma 1.5 technical report. *arXiv preprint arXiv:2604.05081*.

Yiming Shi, Xun Zhu, Kaiwen Wang, Ying Hu, Chenyi Guo, Miao Li, and Ji Wu. 2024. Med-2E3: A 2d-enhanced 3d medical multimodal large language model. *arXiv preprint arXiv:2411.12783*.

Didac Suris, Sachit Menon, and Carl Vondrick. 2023. ViperGPT: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Ziyue Wang, Junde Wu, Chang Han Low, and Yueming Jin. 2025. MedAgent-Pro: Towards multi-modal evidence-based medical diagnosis via reasoning agentic workflow. *arXiv preprint arXiv:2503.18968*.

Yuexin Wu, I-Chan Huang, and Xiaolei Huang. 2023. Timer: Token imbalance adaptation for radiology report generation. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 209 of *Proceedings of Machine Learning Research*, pages 72–85. PMLR.

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*.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu,

Ce Liu, Michael Zeng, and Lijuan Wang. 2023. MM-REACT: Prompting ChatGPT for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*.

Anas Zafar, Leema Krishna Murali, and Ashish Vashist. 2026. Beyond accuracy: Evaluating visual grounding in multimodal medical reasoning. *arXiv preprint arXiv:2603.03437*.

Yuanpin Zhou and Huogen Wang. 2024. Divide and conquer radiology report generation via observation level fine-grained pretraining and prompt tuning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7597–7610. Association for Computational Linguistics.

Appendix Roadmap

The following roadmap provides a structural guide to the supplementary material. To facilitate review, we explicitly categorize each section by its epistemic role: claim-bearing evidence, reproducibility details, design context, or illustrative examples.

Section	Contents	Role in the paper
A	Source data, Stage 3 release composition, representative release row, and interface contrast	Claim-bearing data context
B	Full M3D-Cap/M3D-VQA reference tables supporting the main text	Claim-bearing result support
C	Visual-token budgets and the M3D-Cap grounding bottleneck	Claim-bearing analysis
D	Stage lineage, data-generation variants, and diagnostic-only lines	Design context
E	Selector formulas, token/tag contract, and implementation-level routing details	Reproducibility detail
F	Supervised row schema, route targets, and runtime boundaries	Reproducibility detail
G	Runtime-transfer, packet-selector, and capacity diagnostics	Claim-bearing diagnostics
H	Qualitative traces showing supported answers and failure modes	Illustrative examples
I	Perturbation sensitivity and shared-teacher/judge caveats	Claim-bearing diagnostics
J	Future reward-interface sketch (not used as evidence)	Proposed extension

A Source Data and EDA

This appendix records the data-level facts needed to interpret the main paper without overloading the main text. Table 5 and Figure 3 summarize the dataset composition and the MedAura capability matrix. The central distinction is that the benchmark datasets provide final reports or answers, while Stage 3 requires additional trajectories containing tool calls, observations, evidence-state updates, and withheld decisions. Figure 4 and Figure 5 detail the runtime loop and the Stage 3 supervision contract, with the final EvidenceState data composition shown in Table 6.

Data source	Task	Input	Target	Metrics
M3D-Cap	report-style captioning	3D CT volume	free-text radiology-style caption	BLEU-1, ROUGE-1, METEOR, BERTScore-F1
M3D-VQA close	multiple-choice VQA	3D CT volume + question/options	option A/B/C/D	exact option accuracy
M3D-VQA open	open-ended VQA	3D CT volume + question	free-text answer	BLEU-1, ROUGE-1, METEOR, BERTScore-F1
Unified Stage 3	tool-loop SFT	CT volume + query + tool observations	tool call, evidence state, final/query-more/withhold	task metrics + system metrics

Table 5: Data and task inputs used in the paper. Stage 3 is derived from the benchmark tasks but adds agent-state supervision not present in the original datasets.

A.1 Recent Medical-Agent Interface Contrast

Table 7 summarizes how we position MEDAURA relative to recent medical-agent systems. The comparison is intentionally about interface and supervision design, not task-score superiority, because the systems use different datasets, tool runtimes, and clinical scopes. These systems are empirical prior work; the contrast below is the training object, not the existence of experiments.

Model	Native 3D memory	Native 2D reader	Tool action	Query more	Evidence state	Sidecar / RL-ready
MedGemma 1.5 4B	—	✓	—	—	—	—
M3D-LaMed / Med3DVLM	✓	—	—	—	—	—
2D-Encoder 3D VLM	◐	✓	—	—	—	—
MedAura Stage 2B	✓	✓	—	—	◐	—
MedAura Stage 3	✓	✓	✓	✓	✓	✓

Native agency requires actions, observations, evidence state, and failure routing.

✓ Yes ◐ Partial — No MedAura (Stage 2) MedAura (Stage 3)

Figure 3: Capability matrix for native medical agency. Standard VLM benchmarks measure final answers, whereas the MedAura agent interface also requires actions, observations, query-more behavior, evidence state, and failure routing.

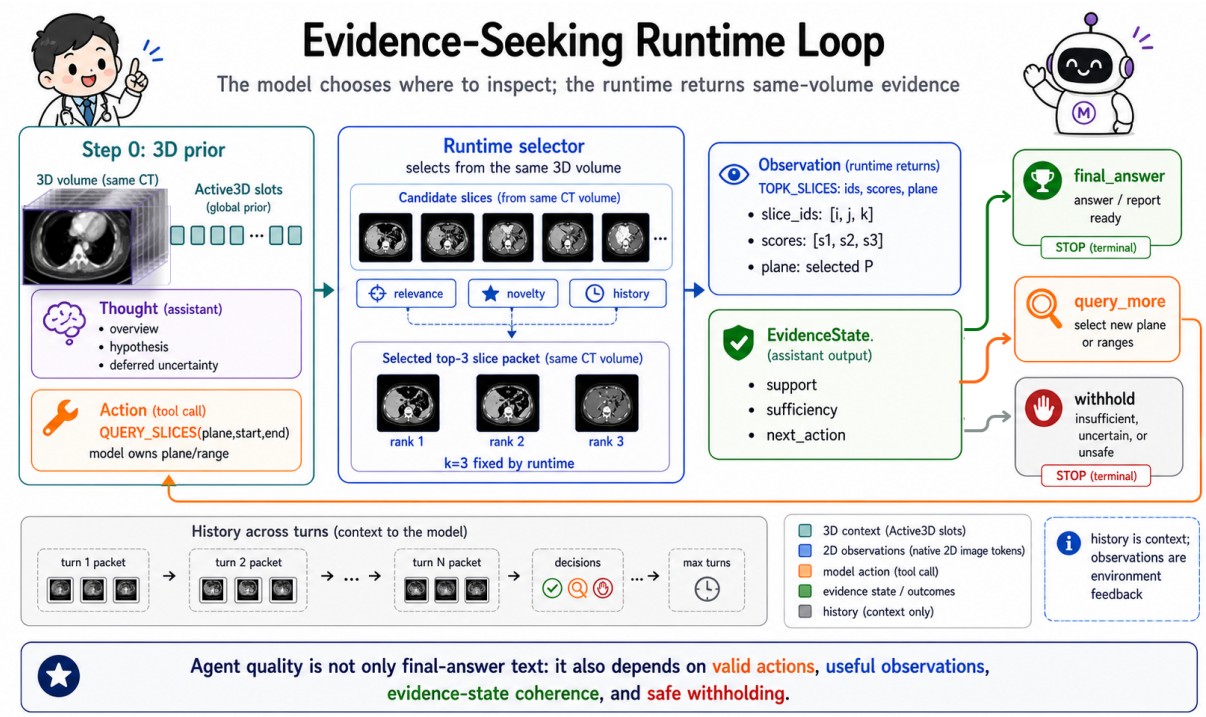

Figure 4: Detailed evidence-seeking runtime loop. This appendix figure expands the compact main algorithm by showing the model-owned action/evidence-state path and the runtime-owned same-volume slice-packet observations.

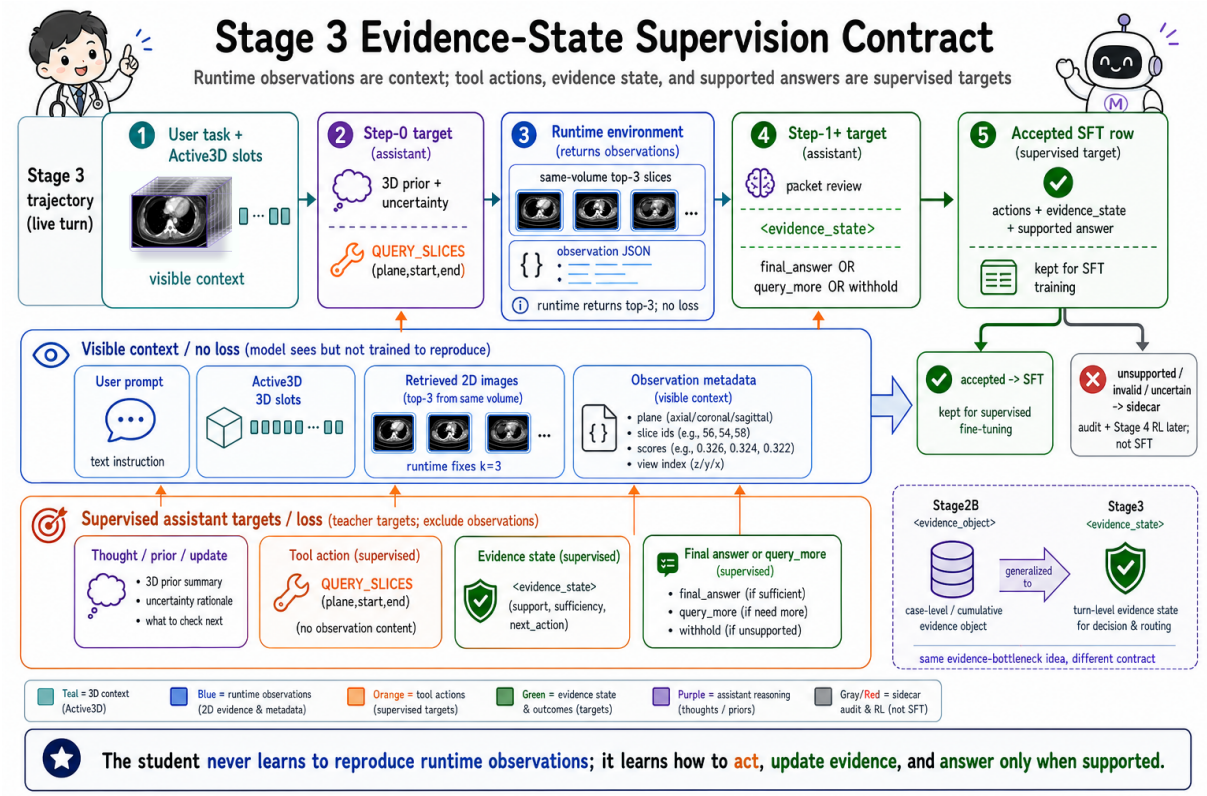

Figure 5: Stage 3 supervision contract. Runtime observations and retrieved images are visible context, while tool actions, EvidenceState fields, query-more decisions, final answers, and withheld decisions are supervised assistant targets after validation.

Stage 3 EvidenceState build field	Audited count	Interpretation
Accepted supervision	21,951 train-ready rows over 8,121 accepted trajectories	positive SFT rows after validation and judge filtering
Audit sidecars	8,255 sidecar rows	rejected, unsupported, or non-final cases retained for audit and preference/RL follow-up, not mixed into positive SFT
Task mixture	11,677 CAP rows; 10,274 VQA rows	M3D-Cap and VQA share the same agent contract, but M3D-Cap remains the harder multi-finding target
Route labels	8,121 step0-query; 5,709 query-more; 8,121 final-answer	the dataset contains both evidence-seeking decisions and terminal answers
Tool-call depth	13,830 queries; mean/max 1.70/4 per accepted trajectory	quantifies the native-agent runtime depth exposed by the released data
Turn structure	step0: 8,121; step1-final: 4,044; step1-query-more: 4,077; later query-more: 1,632; later final: 4,077	roughly half of trajectories require additional evidence after the first query
Support status	uncertain: 8,458; supported: 7,260; partial: 6,199; unsupported: 34	EvidenceState supervision includes sufficiency calibration rather than only supported final reports
Teacher quality	excellent: 10,928; usable: 10,949; borderline: 74	accepted rows are mostly high-quality teacher outputs under the construction audit
Contract audit	duplicate sample IDs: 0; prompt/target contract errors: 0; obsolete Step0 k fields: 0	verifies that the train-ready rows obey the reported Stage 3 contract

Table 6: Stage 3 EvidenceState data composition from the final build audit. These counts explain why the dataset is agent-trajectory supervision rather than ordinary M3D-Cap/M3D-VQA fine-tuning data: it contains explicit query decisions, evidence sufficiency labels, terminal answers, and sidecar failures for follow-up reward construction.

System	Empirical setting	Agent/tool object	Difference from MedAura contract
VoxelPrompt (Hoopes et al., 2024)	volumetric medical-image analysis, especially neuroimaging tasks such as segmentation, measurement, and lesion analysis	language agent emits executable code/instructions that invoke a jointly trained adaptable vision network	broad executable volumetric-image agent; not reported as a same-volume CT plane/range packet trajectory with EvidenceState and final/query-more/withhold targets
CT-Agent (Mao et al., 2025)	3D chest CT question answering and report generation	planning/action/memory modules, anatomy-aware plugins, and global-local token compression for CTQA	close CT-agent neighbor; its reported interface centers on anatomy-aware CTQA/reporting tools rather than the MedAura-style supervised plane/range request, fixed top-3 packet, and EvidenceState routing dataset
RadAgent (Roschewitz et al., 2026)	stepwise chest CT report interpretation	tool-augmented report generation with inspectable intermediate decisions and tool interactions	close reporting-agent contrast; MedAura instead studies an M3D-Cap/M3D-VQA supervision target over M3D-derived trajectories with explicit same-volume slice packets, sufficiency labels, and release-row sidecars
MedVistaGym (Lu et al., 2026)	medical VQA benchmarks	scalable tool-integrated environment for visual reasoning, tool selection, sub-image evidence integration, trajectory sampling, and RL	shares the motivation that training, not tool access alone, matters; MedAura is narrower and CT-specific, releasing supervised same-volume packet trajectories rather than a broad RL environment
MedAgent-Pro (Wang et al., 2025)	multimodal medical diagnosis across diseases/-modalities	hierarchical diagnostic workflow with guideline retrieval, patient-level reasoning, professional tools, and clinical validation	broader diagnostic-agent framework; MedAura focuses on CT evidence acquisition inside one decoder with auditable slice-packet and EvidenceState supervision
MEDAURA	CT M3D-Cap + M3D-VQA	Active3D memory, `QUERY_SLICES` plane/range requests, runtime top-3 same-volume packets, EvidenceState, and final/query-more/withhold routing	reported contribution: a supervised CT evidence-seeking contract, staged diagnostics, and a 21,951-turn trajectory release

Table 7: Recent medical-agent interface contrast. The table contrasts interface and supervision design, not head-to-head task performance. Prior systems demonstrate empirical tool use, stepwise interpretation, or tool-integrated training; MedAura focuses on a CT-specific supervised trajectory contract with model-owned same-volume slice requests, fixed returned slice packets, EvidenceState fields, and final/query-more/withhold routing.

A.2 Representative Stage 3 Release Row

Table 8 shows one abbreviated Stage 3 row derived from an M3D-VQA source record. The example is included to make the released JSONL schema concrete without redistributing raw CT volumes or patient identifiers.

Field	Abbreviated release-row example
Task and source	M3D-VQA-derived release row `sample_id=vqa_001dcc7b8594_step1`, `trajectory_id=vqa_001dcc7b8594::single`, over source volume key `ct_case/022273`; raw volume paths and patient identifiers are not redistributed in the release.
User query	"Where is the fungating mass located within the urinary bladder wall?"
Step-0 model action	`QUERY_SLICES(plane=sagittal, start=0.25, end=0.75)` after a cautious Active3D-only plan to inspect the bladder region.
Runtime observation	`TOPK_SLICES`; plane=`sagittal`; aligned bounds $z_{\text{start}} = 0.25, z_{\text{end}} = 0.75$; candidate count 24; top-3 slice indices $[163, 152, 168]$ with scores $0.3463, 0.3461, 0.3346$.
EvidenceState target	Sagittal `slice_index` 163 and 168 show the lesion at the superior/dome aspect of the urinary bladder; `slice_index` 152 provides adjacent context. `packet_status=observed_topk`; `supporting_slices=[163,152,168]`; `visual_sufficiency=sufficient`; `next_action=final_answer`.
Assistant target and audit	`<final_answer>C. Apex</final_answer>`; route=`final_answer`; support=`supported`; teacher quality=`excellent`; judge decision=`accept`.

Table 8: Representative Stage 3 EvidenceState release-row example derived from an M3D-VQA source record, shortened for space. The full dataset stores the user prompt, assistant target, runtime observation metadata, EvidenceState fields, route labels, audit decisions, and provenance keys; it does not redistribute raw CT volumes.

B External Benchmark Reference Tables

The following tables reproduce the benchmark frames needed to compare MEDAURA against the main M3D-family references. They are included in the appendix so the main text can stay concise while the paper still exposes the full metric set used by Med3DVLM and the data-efficient 2D-encoder 3D VLM.

C Visual Token Budget and M3D-Cap Grounding Requirements

This appendix section explains the visual bottleneck behind the Stage 3 negative result. The tables should be read in order: the token/tag contract defines which symbols are actual tensor routes, the token-budget table explains why the native 2D reader helps only for selected slices, and the requirement and diagnosis tables connect that bottleneck to finding-level M3D-Cap coverage. The goal is not to add another leaderboard claim, but to make clear why a valid tool loop still needs stronger visual memory and selection.

D Training Stages and Data Generation Lines

This section records lineage rather than introducing additional claims. The main paper relies on Stage 2A as the task foundation, Stage 2B as the controlled packet-sensitivity gate, and Stage 3 as the native tool-loop data line. The tables below document how those stages and data-generation variants map to training objectives, visual routes, and evidence roles so that internal construction names do not leak into the main narrative.

E Formal Implementation Details

This section collects implementation formulas that are important for reproducibility but too detailed for the main method. The selector is a runtime policy: the model chooses a plane/range and retrieval intent, while the runtime constructs a fixed same-volume top-3 packet. The formulas below therefore support the packet-ablation analysis rather than defining an additional learned retrieval head.

Novelty-aware slice packet selector. The runtime uses the same selector described in the main text. The implemented candidate score is

$$\begin{aligned} S(i) = R(i) &- 0.30 I_{\text{range}}(r) - 0.25 N_{\text{near}}(i) \\ &- 0.45 C_{\text{img}}(i) - O(i), \end{aligned} \tag{6}$$

Method	Visual/LLM route	B1	R1	M	BERT	Size/class
RadFM	3D ViT + Perceiver + MedLLaMA-13B	12.23	16.49	11.57	87.93	13B-class n/r
M3D-LaMed	3D ViT + Perceiver + LLaMA-2-7B	15.15	19.55	14.38	88.46	6.9B
Med3DVLM	DCFormer + MLP-Mixer + Qwen2.5-7B	36.88	40.25	36.42	88.11	7.6B
2D-Encoder 3D VLM	BioMedCLIP slice encoder + LLaMA-2-7B	**50.26**	**53.37**	**50.13**	**89.98**	7B-class n/r
MedGemma 1.5 4B IT	native medical SigLIP multi-slice reader	16.32	17.15	11.86	81.60	4.30B
Active3D Stage 2A	DCFormer Active3D slots + MedGemma	25.24	24.62	19.25	84.71	4.33B
Stage 3 RegionHistory	Active3D + retrieved 2D packet agent	6.30	13.50	7.39	83.67	4.33B
Stage 3 EvidenceState	Active3D + evidence-state packet agent	11.73	17.49	11.72	83.04	4.33B

Table 9: M3D-Cap benchmark frame. External numbers follow the M3D/Med3DVLM and data-efficient 2D-encoder reports. We omit inference FLOPs because they are not measured under a shared input-length/tool-runtime protocol for all systems. MedAura rows are separated by stage because Stage 2A is the task-level foundation, while Stage 3 rows test the agent interface. Best overall results are in **bold**.

Method	Metric	Plane	Phase	Organ	Abn.	Loc.	Mean
RadFM	B1	14.24	14.25	14.24	15.64	23.58	16.39
RadFM	R1	25.40	25.41	25.38	25.38	29.09	26.13
RadFM	M	20.62	20.63	20.61	20.60	24.19	21.33
RadFM	BERT	92.68	92.04	86.79	85.84	86.26	88.72
M3D-LaMed	B1	98.37	74.41	34.20	15.91	24.00	49.38
M3D-LaMed	R1	98.42	78.63	37.87	19.27	27.74	52.39
M3D-LaMed	M	49.20	63.58	23.78	12.83	18.50	33.58
M3D-LaMed	BERT	99.47	95.55	88.97	86.08	87.60	91.53
Med3DVLM	B1	98.85	78.17	40.22	18.99	25.66	52.38
Med3DVLM	R1	98.89	84.20	45.22	23.27	29.99	56.31
Med3DVLM	M	49.43	68.50	29.32	16.21	20.32	36.76
Med3DVLM	BERT	99.83	96.47	90.47	86.27	87.88	92.18
2D-Encoder 3D VLM	B1	99.07	**82.71**	**49.82**	**26.79**	**32.61**	**58.20**
2D-Encoder 3D VLM	R1	99.15	**89.32**	**52.39**	**29.37**	**38.79**	**61.80**
2D-Encoder 3D VLM	M	**52.32**	**75.91**	**40.06**	**25.81**	**28.19**	**44.46**
2D-Encoder 3D VLM	BERT	99.87	**97.13**	**93.75**	**91.27**	**91.64**	**94.73**
MedGemma 1.5 4B IT	B1	2.92	3.92	2.02	1.94	3.59	2.88
MedGemma 1.5 4B IT	R1	5.61	7.09	3.48	3.14	5.90	5.04
MedGemma 1.5 4B IT	M	10.64	13.68	6.11	4.81	10.20	9.09
MedGemma 1.5 4B IT	BERT	81.87	83.16	81.46	80.93	82.13	81.91
Active3D Stage 2A	B1	**99.20**	78.27	40.64	16.36	32.04	53.30
Active3D Stage 2A	R1	**99.20**	79.72	42.29	18.96	34.43	54.92
Active3D Stage 2A	M	49.60	72.12	26.31	11.96	23.49	36.69
Active3D Stage 2A	BERT	**99.88**	95.85	90.20	84.80	88.96	91.94
Stage 3 RegionHistory	B1	45.00	34.71	16.77	5.72	16.13	23.67
Stage 3 RegionHistory	R1	60.00	40.65	21.15	7.49	19.91	29.84
Stage 3 RegionHistory	M	40.91	39.59	16.14	5.53	15.93	23.62
Stage 3 RegionHistory	BERT	89.99	88.43	86.33	84.28	86.12	87.03
Stage 3 EvidenceState	B1	45.00	29.57	16.16	5.18	15.05	22.19
Stage 3 EvidenceState	R1	60.00	34.17	20.30	6.85	18.62	27.99
Stage 3 EvidenceState	M	40.91	31.35	15.79	5.09	15.04	21.64
Stage 3 EvidenceState	BERT	90.00	88.09	86.08	83.85	85.63	86.73

Table 10: Open-ended M3D-VQA benchmark frame by question type. Abn. denotes Abnormality and Loc. denotes Location. MedAura Stage 3 values are computed from saved E2E inference records, with BERTScore-F1 post-computed offline. Best overall results within each metric row are in **bold**.

Method	Plane	Phase	Organ	Abn.	Loc.	Mean
RadFM	19.65	28.70	16.80	18.92	14.88	19.79
M3D-LaMed	98.80	79.75	74.75	66.65	58.94	75.78
Med3DVLM	99.15	87.50	77.45	70.17	64.49	79.75
2D-Encoder 3D VLM	**99.82**	**90.04**	**79.93**	74.41	**70.32**	**82.90**
MedGemma 1.5 4B IT	73.70	22.20	56.50	55.30	46.90	50.92
Active3D Stage 2A	99.20	81.90	76.60	**76.90**	64.70	79.86
Stage 3 RegionHistory	88.50	33.70	64.60	64.70	55.00	61.30
Stage 3 EvidenceState	89.80	46.30	61.00	60.90	52.40	62.08

Table 11: Closed-ended M3D-VQA accuracy by question type. Stage 2A and both Stage 3 lines are filled from saved official-style records. Best overall results are in **bold**.

Model	Task	N	R@1	R@5	R@10
PMC-CLIP	IR	2000	1.15	4.35	7.60
PMC-CLIP	TR	2000	3.15	8.55	13.55
M3D-LaMed	IR	2000	19.10	47.45	62.25
M3D-LaMed	TR	2000	18.45	47.30	62.15
DCFormer-S SigLIP / Med3DVLM	IR	2000	61.00	87.20	91.25
DCFormer-S SigLIP / Med3DVLM	TR	2000	63.90	88.05	91.90
MedAura Active3D	IR/TR	n/e	n/e	n/e	n/e

Table 12: Image-text retrieval reference table from the Med3DVLM evaluation setup. IR denotes image retrieval and TR denotes text retrieval. n/e denotes not evaluated; MedAura currently targets M3D-Cap/M3D-VQA and agentic evidence use, so retrieval is kept as a reference rather than a claimed result.

Method	Visual strategy	LLM backbone	Total params	3D encoder data	Agentic?
RadFM	3D ViT encoder + Perceiver	MedLLaMA-13B	13B-class, exact total n/r	∼16M	No
M3D-LaMed	3D ViT + 3D spatial pooling Perceiver	LLaMA-2-7B	6.9B	120k	No
Med3DVLM	DCFormer 3D encoder + dual-stream MLP-Mixer	Qwen2.5-7B-Instruct	7.6B	120k	No
2D-Encoder 3D VLM	frozen 2D BioMedCLIP slice encoder + spatial-frequency fusion + feature injection	LLaMA-2-7B	7B-class, exact total n/r	0	No
MedGemma 1.5 4B IT	native medical SigLIP multi-image reader, no native 3D encoder	Gemma 3 / MedGemma 1.5 4B	4.30B	n/a	No
MedAura Active3D	DCFormer global memory + MedGemma native 2D slice reader	Gemma 3 / MedGemma 1.5 4B	4.33B	reused 3D encoder init	Yes

Table 13: Architecture and model-size comparison. n/r denotes not reported. Total parameters are reported when explicitly available; otherwise we report the backbone class to avoid overclaiming. MedAura is smaller than 7B-class 3D VLMs because it is built on MedGemma 1.5 4B/Gemma 3, while adding a lightweight Active3D projector and a native evidence-seeking interface.

Token / tag	Source	Status	Role in MedAura
`<image_soft_token>`	native MedGemma token	active tensor route	Placeholder counted against `pixel_values`; activates MedGemma's native 2D image pathway. It is never used for 3D.
`<3d_volume_slot>`	paper alias for base Gemma unused marker	active tensor route	Active3D volume-slot placeholder. In implementation it is instantiated as Gemma base `<unused10>` to preserve tokenizer/checkpoint compatibility; the input embedding is replaced by projected 3D volume tokens before decoder attention.
`<unused1>`	base Gemma unused token	active action envelope	Stage 3 `QUERY_SLICES` wrapper. The token row is not visual; behavior is learned through SFT/LoRA.
`<mode_3d>` / `<mode_2d>`	custom added tokens	reserved soft boundary	LM-readable phase labels for 3D overview and 2D evidence. The reported Stage 3 line uses natural-language phase instructions rather than consistently wrapping every phase with these tokens.
`<slice_pack>`	custom added token	optional provenance	Packet metadata boundary. The reported Stage 3 train path instead uses image-adjacent `[Slice Packet Item ...]` metadata after each attached image.
`<start_of_volume>`	custom added token	reserved	Disabled in the reported line after earlier confounds; retained only for additional ablations.
`<evidence_state>`	plain multi-token tag	active Stage 3 target	Current EvidenceState bottleneck before final/query-more decisions; supervised as text, not a custom token row.
`<evidence_object>`	plain multi-token tag	Stage 2B controlled target	Used for controlled Stage 2B packet-sensitivity gates and older Stage 3 calibration; not mixed into the main EvidenceState SFT line.
`<final_answer>`	plain multi-token tag	active target	Final M3D-Cap text or VQA option answer.

Table 14: Main token and tag contract. We list only protocol-relevant tokens, not the full Gemma vocabulary. Tensor routing is activated by placeholders plus tensors, not by textual mode tags alone. Explicit modality-boundary wrappers are reserved for an ablation and release-cleanup pass.

Visual source	Token budget	Role	Implication
Active3D volume memory	256 tokens for the full CT volume	global 3D prior and planning state	efficient but compact; may miss subtle or distributed M3D-Cap findings
One MedGemma 2D image	256 native visual tokens per slice	fine local slice reading	strong local detail, but only for selected slices
Current Stage 3 packet	top-3 slices = 768 native 2D tokens	targeted evidence after `QUERY_SLICES`	improves locality but can under-cover multi-finding reports
MedGemma direct multi-slice baseline	many 2D images, each 256 tokens	broad 2D stack reading	strong baseline, but lacks trained 3D memory and agentic query/update policy
Future enriched 3D route	larger/stronger 3D memory or 2D-derived volumetric features	richer global evidence state	intended to reduce M3D-Cap grounding bottleneck while preserving agency

Table 15: Visual token budget. MedAura intentionally keeps the native MedGemma 2D image route while adding compact Active3D volume memory. The same 256-token budget that makes the 3D route efficient also explains why long-form M3D-Cap may need stronger visual-token enrichment.

CAP requirement	Why it matters	MedAura supervision/eval handle
finding coverage	reports may contain multiple positive and negative findings, not one answer label	cumulative evidence state, judge checks, GPT-judge rubric
organ/region binding	a fluent report can still attach a finding to the wrong anatomy	reference-audited region repair, region/plane consistency checks
laterality/location	clinical utility depends on side and anatomical location	slice/range provenance and evidence-state support fields
slice/range support	each visual claim should be traceable to observed evidence	supporting-slice fields, unseen-slice citation filtering
sufficiency/uncertainty	the model should query or withhold when evidence is partial	final/query-more/withhold grammar and sidecar policy
unsupported-claim avoidance	M3D-Cap metrics may not fully penalize hallucinated findings	counterfactual and GPT-judge protocols

Table 16: Finding-level M3D-Cap requirements. M3D-Cap is harder than close-ended VQA because every generated finding should be covered, localized, and supported by evidence rather than merely matching a final option.

Observed symptom	Interpretation	Next measurement or fix
Stage 2A VQA is competitive, M3D-Cap remains behind stronger 3D report models	compact 3D prior is useful for categorical reasoning but weak for multi-finding synthesis	report finding coverage, GPT-judge, stronger 3D/packet evidence
Stage 2B corruption-sensitivity criteria are met	outputs are sensitive to same-volume 2D evidence under controlled conditions	extend to full counterfactual M3D-Cap/M3D-VQA tests
Stage 3 parses tool loop but does not beat Stage 2A on completed task metrics	agent grammar is learnable, but grammar alone does not create grounded report coverage	packet selector, answer-time 3D review, RL route calibration
Stage 3 has additional native 2D packets but remains below Stage 2A on M3D-Cap	local 2D access alone does not preserve Stage 2A's single-pass score inside the multi-turn loop	cited-slice coverage, teacher-forced route mismatch, and stronger visual-memory objectives
EvidenceState improves coherence over RegionHistory	data repair and judge filtering help trajectory quality	keep EvidenceState as released data line and ablation anchor
Counterfactual packets alter evidence-state and final behavior	returned visual evidence affects behavior under perturbation; the failure is incomplete grounding rather than purely image-invariant behavior	vision-assisted judging, lesion-level/localization probes
M3D-Cap metrics still need improvement after cleaner data	visual representation/coverage and single-pass score preservation are bottlenecks, not only teacher formatting	DCFormer base/large, 2D-derived feature enrichment, larger-packet diagnostics

Table 17: Visual-grounding bottleneck diagnosis. The localized Stage 3-to-Stage 2A M3D-Cap gap is used diagnostically: valid tool traces and visual sensitivity are not enough unless the visual representation and packet evidence cover the final findings.

Candidate direction	Purpose	Primary test	Priority
DCFormer-base upsize-init	test whether the current decomp-small 3D encoder is under-powered; compare scale-aware upsize initialization with a direct-upsize control while keeping 256 volume slots	Stage 2A M3D-Cap/M3D-VQA, counterfactual 3D sensitivity; Stage 3 only if Stage 2A improves	diagnostic: base/large upsize and direct-upsize controls did not establish a stronger Stage 2A foundation
Frozen-2D embedding packet reranker (P8)	retrieval-only use of frozen MedSigLIP slice embeddings for relevance, adjacent-slice novelty, MMR diversity, and coverage-aware top-3 selection while preserving MedGemma's native 2D reader	M3D-Cap/M3D-VQA diagnostic subset, pairwise packet cosine, near-duplicate rate	appendix diagnostic; the small hard-range diagnostic reached CAP20 M 10.34 and VQA-close20 60.0 but remained below P4 on M3D-Cap/stability
2D-derived volumetric enrichment	train a bridge from strong 2D medical slice features into compact 3D/evidence tokens; this is the main visual-enrichment direction suggested by recent 2D-encoder 3D VLMs and is distinct from P8 packet reranking	Stage 2A M3D-Cap/M3D-VQA, finding coverage, visual reliance, counterfactual sensitivity	proposed extension
larger-packet slice diagnostic	test whether more local evidence helps M3D-Cap while monitoring token budget	M3D-Cap metrics, support labels, latency	diagnostic; $k = 6$ was negative, so the reported policy stays fixed top-3
answer-time 3D review	keep compact 3D evidence near final synthesis	M3D-Cap, unsupported claims, shuffled-3D sensitivity	medium
RL/GRPO route calibration	reward final/query decisions only when cited slices cover report findings	route reward, final correctness, counterfactual reward	medium
shared-expert route MoE	preserve the frozen medical FFN path while adding small trainable experts for route/task specialization	M3D-Cap/M3D-VQA score preservation, route validity, expert-usage audit	diagnostic; route-capacity diagnostics produced mixed M3D-Cap/M3D-VQA tradeoffs and are not used as the reported model
512-token direct route	test larger visual memory after smaller changes are measured	memory, CAP, grounding sensitivity	proposed extension

Table 18: Model-quality extension path. The present paper focuses on the native 3D+2D agent interface; these variants target the visual-token bottleneck exposed by M3D-Cap.

Stage 3 line	Split	BLEU-1 / Acc.	ROUGE-1	METEOR	BERTScore
Stage 3 RegionHistory	CAP test1k	6.30	13.50	7.39	83.67
Stage 3 RegionHistory	VQA-close test5k	61.30	–	–	–
Stage 3 RegionHistory	VQA-open test5k	23.67	29.84	23.62	87.03
Stage 3 EvidenceState	CAP test1k	11.73	17.49	11.72	83.04
Stage 3 EvidenceState	VQA-close test5k	62.08	–	–	–
Stage 3 EvidenceState	VQA-open test5k	22.19	27.99	21.64	86.73

Table 19: Stage 3 evaluation completion grid. Values are computed from saved E2E inference records; BERTScore-F1 is post-computed offline from predictions and references after generation.

Metadata field	Meaning	Paper policy
Source filename plane	Orientation/reconstruction implied by the source series name, e.g., axial portal venous CT.	Treat as source metadata, not necessarily the only view available at runtime.
Runtime observation plane	Canonical axial, coronal, or sagittal view requested by `QUERY_SLICES` and returned by the runtime.	Source of truth for Step 1 image loading and evidence claims.
Coarse region	Conservative anatomy hint such as brain, chest, abdomen, pelvis, spine, or unknown.	Use as a retrieval/windowing hint; do not treat as a hidden answer label.
Reference-audited region	Teacher-refined anatomical focus used during data construction.	Preserve for audit and trajectory repair; keep student prompts deployment-style.

Table 20: Plane and region policy. A source volume named as axial can still be canonically resliced into coronal or sagittal runtime observations; those views are reformatted slices from the same 3D voxel block, not independent acquisitions.

Stage	Objective	Visual route	Trainable modules	Eval role
Stage 1	Align Active3D socket	3D slots only	projector	mechanical route check
Stage 2A	3D-only M3D-Cap/M3D-VQA foundation	Active3D slots	DCFormer, projector, LM LoRA	task-level line
Stage 2B	3D+2D packet sensitivity	Active3D + native 2D slices	packet-sensitivity adapters/LoRA	controlled gate
Stage 3	tool-loop SFT	3D prior/review + retrieved 2D packets	LM LoRA first pass	agent/data contribution
Stage 4	preference/RL calibration	same as Stage 3	policy/adapters to be determined	proposed extension

Table 21: Stage separation. The main paper keeps benchmark, packet-sensitivity, and agent-readiness claims separate to avoid overclaiming.

Data line	Change introduced	Empirical lesson	Role in this paper
Early epistemic trajectories	early Step 0/Step 1 decomposition and teacher-data experiments	useful for design history, not the reported main line	appendix context
Unified agent-loop trajectories	final/query/withhold grammar with region- and history-aware retrieval	tool-loop execution is feasible, but task semantics can remain weak	diagnostic ablation
Evidence-State trajectories	retrieval intent, support labels, evidence-state supervision, and judge filtering	improves trajectory coherence and provides the reported Stage 3 data line	data/system contribution
Preference/RL trajectories	decision-span preferences and counterfactual evidence conditions	intended to reward stopping only when cited slices cover report findings	proposed extension

Table 22: Data generation lines used to organize the appendix. Implementation suffixes are omitted from the manuscript narrative and retained only in operational tracking.

where $R(i)$ is weighted MedSigLIP text-image relevance, I_{range} is maximum overlap with previous same-plane query ranges, N_{near} is a near-slice penalty, C_{img} is image-cosine redundancy against retrieved history, and $O(i)$ penalizes soft-range fallback candidates outside the requested interval. Let $\mathcal{H}_p$ be the previous retrieved slices from the same canonical plane p. Range overlap is the maximum interval IoU:

$$\begin{aligned} I_{\text{range}}(r) &= \max_{h \in \mathcal{H}_p} \frac{u(r,h)}{v(r,h)}, \\ u(r,h) &= \max(0, \min(e_r, e_h) - \max(s_r, s_h)), \\ v(r,h) &= \max(e_r, e_h) - \min(s_r, s_h). \end{aligned} \tag{7}$$

where $r = (s_r, e_r)$ is the current normalized query interval and (s_h, e_h) is a previous same-plane interval; the value is zero when no same-plane history exists. Candidate-level image redundancy is normalized as

$$\begin{aligned} c_i &= \max_{h \in \mathcal{H}_p} \cos(\phi_I(i), \phi_I(h)), \\ C_{\text{img}}(i) &= \text{clip}_{[0,1]} \frac{c_i - 0.94}{0.985 - 0.94}. \end{aligned} \tag{8}$$

The near-slice penalty is discrete in the implementation:

$$N_{\text{near}}(i) = \begin{cases} 1.0, & d_i \le 2, \\ 0.5, & 2 < d_i \le 5, \\ 0, & d_i > 5, \end{cases} \tag{9}$$

where d_i is the minimum absolute slice-index distance to previously returned same-plane slices. For the reported hard-range selector, $O(i) = 0$. In the soft-range diagnostic, candidates outside the model-requested interval receive $O(i) = 0.12(1 + \delta_i)$, where δ_i is the distance from the candidate normalized slice position to the nearest query boundary. The runtime then sorts by $S(i)$, deduplicates slice ids, and returns exactly three slices. A packet is rejected as redundant only when all hard gates hold:

$$\begin{aligned} \text{IoU} \ge 0.85, \quad n_{\text{new}} = 0, \\ n_{\text{near}} \ge 2, \quad \max \cos \ge 0.985. \end{aligned} \tag{10}$$

This means repeated tool calls can still pass if they introduce new slices or substantially different visual evidence.

Algorithm 2 Runtime slice-packet selector. The model supplies only $q_t = (p_t, s_t, e_t)$ and intent text; the runtime owns candidate extraction, MedSigLIP ranking, deduplication, redundancy rejection, and fixed top-3 selection.

Require: CT volume V, query $q_t = (p_t, s_t, e_t)$, history H_t, row context x_t
Ensure: observation packet o_t or rejection sidecar

```
k ← 3
Q_t ← BuildQueries(x_t, q_t)                          ▷ |Q_t| ≤ 7 weighted MedSigLIP prompts
C_t ← Candidates(V, p_t, [s_t, e_t]; 24)
if SoftRange = 1 for a diagnostic run then
    C_t ← C_t ∪ ExpandedRange(V, p_t, s_t, e_t) ∪ SparsePlane(V, p_t)
end if
for all i ∈ C_t do
    R(i) ← max_{q∈Q_t} w_q cos(φ_T(q), φ_I(i))
    S(i) ← R(i) − 0.30 I_range(r_t) − 0.25 N_near(i) − 0.45 C_img(i) − O(i)
end for
P_t ← FirstKUnique_{i∈SortDesc(C_t,S)}(i, k)
ν_t ← NoveltySummary(P_t, H_t, p_t, r_t)
if Redundant(ν_t) then
    return QUERY_REJECTED_REDUNDANT_PACKET
end if
return o_t = TOPK_SLICES(p_t, P_t, ν_t)
```

Turn type	Main boosted spans	Purpose
Step 0 planner	<unused1> 3.0, QUERY_SLICES 4.5, plane 3.5, start/end 3.0, plan span 2.0	learn valid evidence-seeking actions while downweighting non-terminal clinical prose
Query-more turn	evidence-state 2.5, support fields 3.0, next action 3.5, QUERY_SLICES 4.0, plane 3.0, start/end 2.5	learn when evidence is partial and how to request additional evidence
Final turn	evidence-state 2.5, support fields 3.0, final-answer field 3.5, final-answer span 2.0, tags 3.0	bind terminal synthesis to the evidence state rather than free-form fluency
VQA final prefix	answer option prefix weight multiplied by the per-row choice weight	keep close-ended VQA parseable and reduce malformed terminal answers

Table 23: Stage 3 token-weight policy implemented by the collator. Values are multiplicative token weights for matched spans. They are not new supervision labels; they shape which parts of the teacher trajectory are emphasized. The weights were selected from early development stability diagnostics rather than exhaustively tuned, and are reported for reproducibility.

Turn tag	Inputs	Supervised target	Purpose
step0	user query + Active3D slots	3D prior, deferred uncertainty, QUERY_SLICES call	learn where to inspect
step1_final	first observation + slice packet	evidence state + final answer	answer from first evidence packet
step1_query_more	first observation + slice packet	evidence state + second QUERY_SLICES call	continue when evidence is partial
step_query_more	later observation + history	updated evidence state + QUERY_SLICES call	multi-turn refinement
step_final	cumulative observations + history	final evidence state + final answer	supported terminal synthesis

Table 24: Unified Stage 3 turn tags. Unsupported, invalid, failed, or non-final trajectories are withheld from first-pass SFT and retained for audit or later preference/RL training.

Span	Supervised?	Role
User query and task prompt	no	deployment-style request and task context
Active3D slots	no	tensor-level 3D visual input
Retrieved 2D images	no	native MedGemma visual input
Observation metadata	no	environment feedback: plane, slice indices, scores, normalized positions
Thought / prior / update	yes	model's epistemic state and belief update
QUERY_SLICES action	yes	model-generated tool decision
Evidence state	yes	compact support/sufficiency/citation state near the decision
Final answer	yes	M3D-Cap text or VQA answer target

Table 25: Input/target boundary for Stage 3 SFT. This distinction prevents the model from being trained to reproduce environment observations while still learning how to react to them.

Stage	Input or action	Quality control	Output
Registry task	CT volume, M3D-Cap/M3D-VQA prompt, benchmark target	deduplicate by volume and align reference/registry keys	task row
Step 0 teacher	3D prior and task prompt	require cautious hypothesis, deferred uncertainty, valid plane/range	first QUERY_SLICES call
Runtime retrieval	model-selected plane/range	fixed top-3, MedSigLIP relevance, novelty/history checks	slice-packet observation
Step 1+ teacher	packet images, observation metadata, history, reference target for offline audit	citation subset, support status, route consistency, no unseen-slice evidence	evidence state plus final/query-more/withhold
Judge/filter	complete trajectory	deterministic validators plus fixed GPT judge	accepted SFT rows or sidecar

Table 26: Stage 3 data lineage. The final SFT release contains only accepted train-ready turns. Sidecars preserve failures, unsupported evidence, and ambiguous cases for audit and Stage 4 preference/RL design.

```
<thought>
[Overview] macroscopic 3D findings / exclusions.
[Hypothesis] suspected pathology or cautious negative prior.
[Retrieval Intent] evidence to check next.
[Deferred] unresolved from 3D overview alone.
[Plan] call QUERY_SLICES with plane/start/end.
</thought>
<unused1>{"name":"QUERY_SLICES",
 "arguments":{"plane":"axial","start":0.40,"end":0.85}}
</unused1>

<observation>
{"obs":"TOPK_SLICES","topk":[
 {"slice_index":56,"score":0.0545,"z_norm":0.44}, ...]}
</observation>

<thought>
[Tool Quality Check] packet usability.
[Evidence] cited returned slices only.
[Resolved] what the packet confirms or rejects.
[Grounded Synthesis] 3D prior + 2D evidence.
[Sufficiency] sufficient / partial / marginal / inconclusive.
</thought>
<evidence_state>
packet_status, supporting_slices, unresolved_evidence,
visual_sufficiency, next_action
</evidence_state>
<final_answer>
M3D-Cap: report-style text. VQA: A./B./C./D. answer format.
</final_answer>
```

Figure 6: Canonical Unified Stage 3 trajectory skeleton. The observation is runtime context rather than a supervised target; the model is supervised on the epistemic update, tool action, evidence state, and supported final answer.

Concept	Definition	Reported interpretation
Source filename plane	Dataset metadata such as an axial reconstructed NIfTI stack.	acquisition/reconstruction orientation, not the only possible view
3D volume	Loaded voxel block used by Active3D and runtime tools.	same CT volume can be resliced into canonical views
Runtime observation plane	Plane requested by `QUERY_SLICES` and returned in the observation.	source of truth for Step 1 image loading
Cross-plane query-more	Later turns may request axial, coronal, or sagittal evidence.	valid resliced evidence, not independent acquired series

Table 27: Plane policy for Stage 3 data. An axial source volume can still produce coronal or sagittal runtime observations by reslicing the same 3D voxel block. We therefore report runtime planes as canonical evidence views, not separate acquired studies.

F Unified Stage 3 Trajectory Contract

Stage 3 data is a trajectory flattened into supervised assistant turns. Environment observations are context. Tool calls, thoughts, evidence-state updates, query-more decisions, and final answers are assistant targets.

Canonical trajectory skeleton. A supported Stage 3 trajectory follows the form below. The runtime fixes the slice budget; older construction rows may contain a generated `k`, but deployment ignores it. The model owns the canonical plane/range decision and the evidence-state update.

M3D-Cap and M3D-VQA sample interpretation. M3D-Cap trajectories are multi-finding: a final report should be supported by the cumulative evidence state, not by a single generic packet. VQA trajectories use the same evidence loop, but the terminal answer is normalized to the benchmark format; close-ended answers must begin with the selected option label and text. Qualitative model-output examples

EvidenceState audit	Rows	Traj.	Axial obs.	Coronal obs.	Sagittal obs.	Multi-plane traj.
Accepted train-ready data	21,951	8,121	8,247	3,241	2,342	2,410

Table 28: EvidenceState data audit. Observation counts are runtime evidence planes. Multi-plane trajectories are cases where later turns requested a different canonical resliced view to resolve remaining uncertainty.

```
Current reported Stage3 input packing

Step0 user side:
  [Active3D volume slots inserted by collator: <3d_volume_slot> x 256]
  Task: Step0 3D planning for a native medical imaging agent.
  Use only the active 3D volume tokens and text metadata available at Step0.
  No 2D slice images have been retrieved yet.
  Question/Input: ...
  Body region hint: ...
  View Plane or axis metadata: ...

Step0 assistant target:
  <thought>
  [Overview] cautious 3D overview.
  [Hypothesis] possible target/finding requiring 2D confirmation.
  [Retrieval Intent] short visual-evidence query for packet ranking.
  [Deferred] unresolved from 3D overview alone.
  [Plan] call QUERY_SLICES(plane,start,end).
  </thought>
  <unused1>{"name":"QUERY_SLICES",
   "arguments":{"plane":"axial","start":0.30,"end":0.70}}</unused1>

Step1 user side:
  [Slice Packet]
  The following 2D slices were retrieved from the same 3D CT volume.
  <image>
  [Slice Packet Item 1] rank=0; plane=axial; slice_index=56;
    z_norm=0.44; query_range=0.30-0.70; score=...; relevance=...
  <image>
  [Slice Packet Item 2] rank=1; plane=axial; slice_index=64; ...
  <image>
  [Slice Packet Item 3] rank=2; plane=axial; slice_index=43; ...
  [Active3D volume slots inserted near answer boundary]
  prior Step0, observation JSON, and task instruction

Step1 assistant target:
  <thought>packet audit, evidence, unresolved evidence, synthesis</thought>
  <evidence_state>
  packet_status, supporting_slices, visual_sufficiency, next_action
  </evidence_state>
  <final_answer>supported M3D-Cap text or VQA answer</final_answer>
```

Figure 7: Current Stage 3 prompt and metadata packing. The reported line uses natural-language phase instructions and image-adjacent slice provenance rather than consistently wrapping phases with `<mode_3d>` or `<mode_2d>`. The visual routes are still tensor-driven: Active3D volume-slot replacement for 3D and MedGemma native image placeholders plus `pixel_values` for 2D.

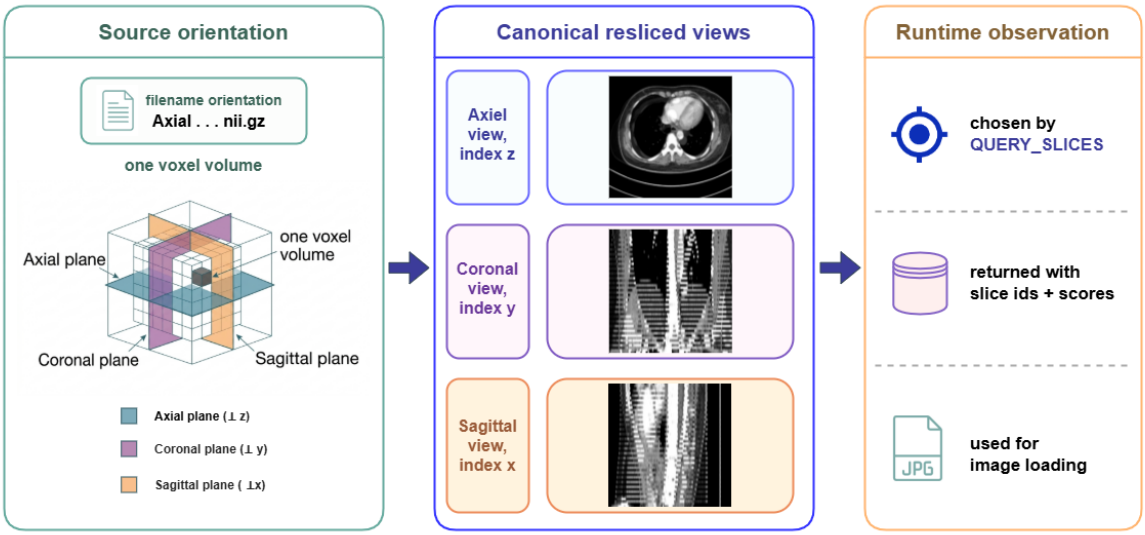

Figure 8: Source orientation versus runtime observation plane. A source filename may describe the acquired/reconstructed stack orientation, while the runtime can request canonical axial, coronal, or sagittal resliced views from the same voxel block. The shown axial thumbnail is an example frozen slice; the coronal and sagittal panels illustrate canonical reslice options rather than independent acquisitions.

Public line	Role	Main limitation addressed	Paper role
Early epistemic data	early Step0/Step1 pipeline and ablation source	useful but inconsistent trajectory coherence	design context
Stage 3 RegionHistory	first evaluated unified Stage 3 agent-loop SFT line	valid runtime loop but weak M3D-Cap grounding and noisy inherited priors	diagnostic baseline
Stage 3 EvidenceState	repaired expert-trajectory line used for current Stage 3	retrieval intent, reference-audited coherence repair, support labels, judge-all filtering	reported Stage 3 line
Stage 4 preference/RL data	preference/RL follow-up source	uses sidecars and model failures after SFT stabilizes	preference/RL scaffold

Table 29: Data-line naming. The paper reports reader-facing stage names and keeps construction variants as appendix traceability rather than method names.

Data ablation	Question	Measurement	Status
RegionHistory vs EvidenceState	Does explicit support state repair noisy trajectories?	task metrics plus parse, support, final/query-more rates	main comparison
Judge/validator off vs on	Do filters prevent unsupported supervision?	sidecar reasons, invalid finals, unseen-slice citations	audit replay
Final-only vs query-more	Does agency help when the first packet is partial?	final-turn distribution, M3D-Cap support, withhold rate	future diagnostic
Reference-hidden vs reference-audited teacher	Does reference auditing improve coherence without changing deployment prompt?	contradiction rate, region/-plane consistency, support labels	construction audit
Sidecar excluded vs mixed	Are unsupported cases kept out of SFT?	final-answer corruption and route contradictions	policy check

Table 30: Data-quality audit matrix. These ablations test whether the trajectory data teaches evidence-dependent behavior rather than only fluent medical text generation.

are reported separately from the data contract so that benchmark targets and generated diagnostics are not conflated. The data/contract takeaway is that Stage 3 contributes supervised environment trajectories, not only additional prose: accepted rows train actions and EvidenceState fields, while sidecars preserve unsupported or unsafe behavior for audit and future reward construction.

G Ablation Study Matrix

This appendix reports MedAura ablations that separate the agent-specific design choices. External component ablations from prior 3D medical VLM papers are used only as context for the visual-grounding hypothesis, not as MedAura result tables.

How to read the evidence. The agent contribution is not established by a higher M3D-Cap score alone. It is supported by Stage 2B packet-sensitivity diagnostics, Stage 3 parse/runtime success, runtime-transfer failure of prompted baselines, and fixed qualitative traces showing model-owned `QUERY_SLICES` decisions. The retrieval contribution is also limited on purpose: P4 is the reported fixed top-3 policy because it gives the most stable M3D-Cap/open-VQA tradeoff among the evaluated diagnostics, while P5, P6, and P8 test tempting alternatives.

P8 should not be read as the 2D-to-3D visual-encoder bridge. It is only a retrieval-time reranker that reuses frozen 2D MedSigLIP embeddings to reduce repeated or near-duplicate slices inside the same three-slice packet. The P8 diagnostic supports the same cautionary reading: cleaner packet diversity alone did not solve M3D-Cap captioning in our small diagnostic subset, so the remaining weakness is consistent with finding-level evidence coverage and compact 3D visual representation rather than duplicate-slice selection alone.

We organize the appendix evidence in three layers. First, capability ablations ask whether a line has native 3D memory, native 2D evidence reading, tool actions, query-more behavior, and evidence-state supervision. Second, transfer ablations ask whether the same runtime can be induced by prompting external or non-agent baselines. Third, bottleneck ablations ask why a valid agent loop still lags the highest-scoring final-answer Stage 2A line on M3D-Cap/M3D-VQA. This separation is important: many

rows are negative or diagnostic by design, and their role is to locate the remaining grounding bottleneck rather than to advance a new model variant.

Concretely, the appendix has four evidence clusters. The data/contract cluster establishes what supervision Stage 3 adds beyond benchmark task supervision. The capability/transfer cluster tests whether tool use is learned or merely prompted. The bottleneck cluster localizes failures to visual evidence richness, packet coverage, cited-slice coverage, and finding-level binding. The counterfactual/judge cluster tests whether evidence-state and final-answer behavior changes when the visual packet or 3D route is perturbed. These clusters are meant to support one argument: MEDAURA has a trained native-agent interface, but stronger visual grounding is still needed for finding-complete report-style generation. Rows marked as proposed extensions are included only to delimit unclaimed next steps; they are not used as evidence for the reported claims. Claim-bearing evidence is restricted to reported/evaluated rows, fixed qualitative traces, and the diagnostic tables explicitly reported here. Small-subset diagnostics use fixed subsets for rapid iteration and should be read as diagnostic indicators, not statistically powered model rankings.

The ablations also act as a diagnostic map. Runtime transfer shows that prompt formatting alone does not sustain the loop in this setup; grammar diagnostics show that invalid tool syntax is not the main failure; the $k = 6$ packet diagnostic and P8 reranker diagnostic show that raw packet widening or redundancy-only reranking do not close the gap in the evaluated subsets; the answer-time review/MoE diagnostics show that simple placement or decoder-capacity changes did not close the gap in our evaluations; and the localized Stage 3-to-Stage 2A gap shows that wrapping the model in the multi-turn loop can fail to preserve Stage 2A's single-pass score even when extra 2D evidence is available. The remaining failure is unlikely to be purely image-invariant: counterfactuals show visual sensitivity, but the system still has under-resolved visual evidence and incomplete cited-slice coverage for finding-complete M3D-Cap.

External visual-route evidence. Prior component ablations motivate the visual-token bottleneck diagnosis but are not counted as MedAura ablations. Med3DVLM reports that removing DCFormer strongly degrades both M3D-Cap and M3D-VQA performance, supporting the need for a native global 3D route (Xin et al., 2025). The data-efficient 2D-encoder 3D VLM reports that strong 2D medical slice features, nonredundant slice selection, and feature-injection modules improve volumetric M3D-Cap/M3D-VQA performance, supporting a 2D-derived visual-enrichment direction (Lian et al., 2026). MedAura's completed P8 experiment is narrower: it only reranks which slices are sent through the existing MedGemma image route and does not train new visual tokens. The corresponding MedAura architecture direction is therefore the 2D-derived volumetric bridge in Table 18, not P8 packet reranking.

The next capability/transfer tables ask whether this interface is learned or only prompted. The intended conclusion is narrow: a model can have a high single-pass benchmark score and still fail the observation–EvidenceState–final runtime contract.

Table 32 defines the capability gap that motivates the paper. It is a taxonomy, not a score table: single-pass systems may have high benchmark scores while still lacking explicit tool action, query-more behavior, and evidence-state supervision.

For the agent-control data audit, we define four audited subsets from the same Stage 3 training contract: direct-final rows, one-query-final rows, multi-turn rows, and packet-conditioned final rows. On the first 2,000 CAP rows, the multi-turn slice contains 627 Step0 query rows, 747 query-more rows, and 626 final rows, indicating that the data can support controlled agent grammar ablations. The enhanced audit also finds valid JSON tool payloads, canonical planes, and normalized ranges for all audited one-query and multi-turn query rows (A1/A2 parse, plane, and range validity are all 1.000). The A2 multi-turn slice uses all three canonical reslicing planes (870 axial, 258 coronal, 246 sagittal query rows). This contract audit is used only to define comparable ablation subsets; task metrics still require model-output evaluation.

Table 33 then tests whether these capabilities can be induced by prompting alone. In this setup, they cannot: MedGemma can often emit a first syntactically valid tool call, but the observation–evidence–final loop does not persist without Stage 3 training.

Table 35 asks whether the remaining M3D-Cap weakness can be fixed by changing slice retrieval alone. The negative P6 result is important: adding more slices increases context but lowers stability, so the reported runtime keeps a fixed top-3 packet and treats coverage as a finding-level selection problem rather

Ablation axis	Completed evidence	Main lesson	Claim status / implication
Baseline/stage anchors	MedGemma, Stage 2A, Stage 2B gate, Stage 3 RegionHistory, Stage 3 EvidenceState	Stage 2A is the highest-scoring task-level foundation among MedAura lines; Stage 3 EvidenceState is the reported native-agent line but not a leaderboard win.	Reported anchors; not a leaderboard claim for Stage 3.
Agent runtime transfer	CAP200 transfer with MedGemma-runtime, Stage 2A-runtime, Med3DVLM-runtime, and Stage 3 EvidenceState native runtime	In our setup, prompting non-agent baselines with the tool interface does not induce the full native runtime contract.	Claim-bearing capability diagnostic.
Slice-packet selector	P4/P5/P6 CAP200 and VQA-close200; P4/P5 VQA-open200	Intent-guided fixed top-3 is the most stable evaluated packet policy; soft range helps close VQA but hurts M3D-Cap/open-VQA; the larger $k = 6$ diagnostic hurts stability.	Points to finding/evidence-gated selection rather than larger raw packets.
Counterfactual packet sensitivity	CAP200, CAP100, and VQA-close100 with blank, off-target, donor-volume, and drop-support packets	EvidenceState changes under every perturbation; same-volume off-target packets produce the largest disruption.	CAP200 five-condition GPT-judge is reported as a diagnostic, not a benchmark replacement.
Answer-time 3D review	Stage 3-initialized 64-slot final-review diagnostic; CAP32 protocol pass; CAP200 quality below P4	Placement alone did not fix M3D-Cap grounding in the completed diagnostic once the grammar was preserved.	Diagnostic evidence unless paired with a stronger evidence-gated objective.
3D visual route	Prior 52-token/256-cache lessons, current 256 direct slots, Stage 2B 3D+2D binding, Stage 2A zero/shuffled 3D-slot diagnostic	Direct slots are viable and behaviorally active under zero/shuffle perturbations, but M3D-Cap still exposes a 3D token semantics and coverage bottleneck.	Supports stronger visual-token enrichment as a proposed extension.
DCFormer scale	DCFormer base/large upsize and direct-upsize controls with a scale-aware 256-token projector	Base/large upsize and direct-upsize controls did not establish a stronger Stage 2A foundation.	Ablation evidence; next visual-route work should use 2D-derived volumetric enrichment or stronger slice-to-volume supervision.
Route grammar/data contract	Direct-final, one-query, multi-turn, and packet-final contract audits; RegionHistory/EvidenceState evaluations	The reported grammar audit is valid; EvidenceState is the repaired SFT contract. A Stage 3-from-Stage 2A initialization diagnostic improved VQA-close200 to 64.5%, but CAP200 remained weak (M 8.01) with 29.0% forced-final rate.	Retain packet-sensitivity initialization as the reported Stage 3 line; Stage 2A-init remains diagnostic until paired with a CAP-specific objective.
Shared-expert MoE	Isolated scaffold over Stage 3 EvidenceState: frozen shared FFN path plus small routed residual experts in late and global LM layers	Route-capacity diagnostics produced mixed M3D-Cap/M3D-VQA tradeoffs and are not used as the reported model.	Retain as diagnostic until paired with a task/route-conditioned M2 variant.
Boundary tokens and Stage 4 RL	Design only	<mode_3d> and <mode_2d> are text phase markers, not tensor routers; RL should optimize evidence decisions after SFT.	Proposed extension.

Table 31: Ablation status. Evaluated small-subset diagnostics are separated from official task benchmarks and proposed-extension rows, which are not used as evidence for the reported claims. The DCFormer-scale lane is retained as a diagnostic; the next short-horizon visual priority is 2D-derived volumetric enrichment rather than another Stage 3 retrain from base-upsize.

Line	Native 3D	Native 2D	Tool action	Query-more	Evidence state	Empirical role
MedGemma 1.5 4B direct multi-slice	–	✓	–	–	–	strong 2D/multi-slice baseline
M3D-LaMed / Med3DVLM	✓	–	–	–	–	single-pass 3D VLM baselines
Active3D Stage 2A	✓	–	–	–	–	3D foundation without agency
Active3D Stage 2B	✓	✓	–	–	partial	controlled packet-sensitivity gate
Stage 3 RegionHistory	✓	✓	✓	✓	partial	diagnostic agent loop
Stage 3 EvidenceState	✓	✓	✓	✓	✓	repaired agent candidate

Table 32: Agent-capability ablation. Standard task metrics compare final answers, but they do not establish whether a system is an agent. This table separates native visual state, native 2D evidence reading, explicit tool actions, multi-turn query-more behavior, and evidence-state supervision.

than a raw packet-width problem.

P8 addresses a plausible alternative explanation. Since MedGemma already has a strong native medical SigLIP image pathway, one might expect a frozen-2D reranker to solve the evidence problem by selecting less redundant slices. Our P8 test does exactly that in a retrieval-only way: it reuses frozen MedSigLIP slice embeddings to balance relevance, adjacent-slice novelty, MMR-style diversity, and coverage before selecting the same fixed three-slice packet. It does not train a new 3D encoder and does not create richer volume tokens. The result is diagnostic: redundancy drops in the evaluated subset, but M3D-Cap metrics do not improve enough for duplicate-slice selection to explain the gap by itself; the remaining gap is more consistent with finding-level visual coverage and representation quality.

Table 36 separates route grammar from semantic grounding. The repaired EvidenceState contract makes actions parseable and auditable, but M3D-Cap quality still depends on visual evidence quality and the model's ability to bind findings to the right slices.

The answer-time review and MoE diagnostics below are capacity/placement tests. They are included because they tested plausible non-visual fixes and did not remove the M3D-Cap gap, strengthening the conclusion that better evidence representation and finding-level selection are the next targets.

H Qualitative Case Studies

Qualitative examples are drawn from evaluated inference outputs, not from hand-written idealized trajectories. Each case is recorded with the same row index, volume reference, runtime plane, queried slice indices, and provenance metadata used by the quantitative tables. This makes the qualitative figure auditable and prevents cherry-picked visual evidence from drifting away from the reported metrics.

Selection and audit policy. Qualitative cases are selected by deterministic filters over inference outputs. CAP cases include the target prompt, reference report, baseline outputs, Active3D Stage 2A output, Stage 3 EvidenceState query, returned slice packet, evidence state, and final answer when available. The evaluated provenance set contains success/partial-success, CAP-failure, query-more, and counterfactual candidates; slice images are reconstructed from stored runtime provenance rather than manually chosen after reading the answer. All shown cases have Stage 3 EvidenceState slice packets and deterministic joins to MedGemma single-pass, Stage 2A direct, and Stage 3 official outputs. Runtime-transfer rows are retained only as protocol-stress evidence: Stage 2A-runtime has 50/50 non-empty Step 0 responses but 0/50 valid tool calls; MedGemma-runtime has 50/50 non-empty Step 0 responses, 50/50 parsed tool calls, and 2/50 final answers; Med3DVLM-runtime has 0/50 valid tool calls or final answers. These rows are not official single-pass benchmark predictions.

Visual layout. Each qualitative card shows the target prompt, representative runtime slices, model output, evidence state, and a one-line audit summary. Long raw trajectories are omitted from the main figure and retained in the supplementary trace record.

Provenance traceability. Every case used in a figure has a provenance row containing `record_index`, `volume_ref`, model alias, runtime plane, `top_indices`, window preset, prediction text, reference text,

Evaluated policy	Runtime given	3D prior	2D read	Tool parse	Query-more	Evidence	Withhold	Status / interpretation
MedGemma 1.5 4B IT	MEDAURA prompt + runtime	–	✓	1.00	0.02	0.005	0.96	CAP200: first tool syntax transfers, but the observation–evidence–final loop does not (runtime success 0.04)
M3D-LaMed	MEDAURA prompt + runtime	✓	–	n/a	n/a	n/a	n/a	reference scaffold only unless an executable checkpoint is available
Med3DVLM	MEDAURA prompt + runtime	✓	–	0.00	0.00	0.00	0.00	CAP200: remains single-pass under the external tool prompt (200 invalid queries)
2D-Encoder 3D VLM	MEDAURA prompt + runtime	partial	✓	n/a	n/a	n/a	n/a	literature reference; executable comparison depends on compatible check-point/runtime availability
Active3D Stage 2A	MEDAURA prompt + runtime	✓	–	0.04	0.00	0.00	0.03	CAP200: 3D foundation alone does not learn the agent protocol (runtime success 0.01)
Stage 3 EvidenceState	native trained runtime	✓	✓	1.00	0.425	1.00	0.255	CAP200: trained native-agent line operates the full runtime (runtime success **0.745**)

Table 33: Runtime-transfer agent stress test on CAP200 with forced-final recovery disabled. Every executable baseline receives the same external MedAura runtime protocol: tool instructions, `QUERY_SLICES` syntax, retrieved slice-packet observations, evidence-state prompt, and final/withhold policy. Values in the Tool/Query/Evidence/Withhold columns are rates. This table tests native medical agency rather than raw task accuracy: a model can score well in single-pass inference yet fail to operate the evidence-seeking loop.

Axis	Variant	Purpose	Primary measurement	Status
3D route	52-token cache/prefix	test aggressive volume compression	task metrics, grounding sensitivity	prior diagnostic
3D route	256-token cache/prefix	denser memory prefix	task metrics, evidence use	prior diagnostic
3D route	256 direct Active3D slots	canonical direct tensor route	task metrics, route liveness	reported main route
3D route	3D+2D binding	test coexistence with native 2D slices	sufficiency/citation gates	evaluated
3D route	64-slot answer-time final review	test whether compact 3D evidence near final synthesis improves M3D-Cap grounding	M3D-Cap/VQA diagnostic subset, runtime success, counterfactual sensitivity	diagnostic result: Stage 3 init resolves the protocol mismatch, CAP200 M 7.82 remains below P4
3D route	zero/shuffled 3D-slot sensitivity	test whether direct Active3D slots affect output under perturbation	output-change rate under zeroed and shuffled 3D slots	diagnostic result: n=200, zero-change 48.0, shuffled-change 50.0
3D route	512 direct slots or stronger DCFormer	test spatial-token/encoder bottleneck	CAP, grounding diagnostics	proposed extension
boundary tokens	explicit <mode_3d>/<mode_2d> wrappers	test whether LM-readable phase labels improve 3D/2D binding without changing tensor routes	parse stability, evidence-state consistency, task metrics	proposed diagnostic

Table 34: 3D route/token and boundary-token ablation scaffold. This paper reports the 256 direct-slot route and the Stage 2B packet-sensitivity gate. Explicit modality wrappers are listed as an additional data/prompt cleanup diagnostic, not as a current Stage 3 reported claim.

Packet line	Name	Purpose	Primary measurement	Status
P0	uniform range top-3	naive slice-packet baseline	task score, packet relevance	prior diagnostic
P1	MedSigLIP single-query top-3	generic visual relevance	task score, packet precision	prior diagnostic
P2	MedSigLIP multi-prompt top-3	task/question-aware relevance	task score, relevance diagnostics	implemented component
P3	novelty-aware MedSigLIP top-3	reduce repeated/redundant packets	redundancy, runtime success	implemented runtime
P4	intent-guided novelty packet	use retrieval intent in packet scoring	CAP200 M **11.71**; VQA-close200 59.0; VQA-open200 M **22.32**	selected runtime policy
P5	soft-range intent-guided packet	recover from narrow/wrong ranges	CAP200 M 11.23; VQA-close200 **62.0**; VQA-open200 M 22.08	VQA-close above P4; M3D-Cap/open-VQA below P4
P6	larger-packet $k = 6$ diagnostic	test M3D-Cap multi-finding coverage beyond fixed top-3	CAP200 k=6 M 9.61; VQA-close200 57.5	below P4/P5; not a reported line
P7	balanced multi-plane packet	test coverage across canonical views	plane consistency, task metrics	proposed diagnostic
P8	frozen-2D embedding reranker	reduce within-packet visual redundancy while preserving relevance	soft CAP20 M 8.32 / VQA-close20 45.0; hard CAP20 M 10.34 / VQA-close20 60.0	appendix diagnostic; below P4 CAP/stability; not a new encoder

Table 35: Slice-packet selector ablation scaffold. Current retrieval is not random; it already uses MedSigLIP relevance plus novelty. The CAP200 diagnostics show that neither soft range expansion, plain $k = 6$ packet growth, nor diversity-only frozen-2D packet reranking addresses the current M3D-Cap bottleneck, motivating finding-level evidence coverage and stronger visual evidence rather than larger or less-redundant packets alone.

Grammar line	Behavior	Main lesson	Status
no-agent single-pass	direct answer from available visual context	strong baseline; no evidence-seeking policy	evaluated as Stage 2A
terminal Step 1	one query, one packet, final answer	too rigid for M3D-Cap coverage	replaced by EvidenceState line
single-loop query/final	tool call exists but always terminates	tool can become ceremonial	prior diagnostic
unified final/-query/withhold	model may answer, query more, or withhold	runtime works; semantics need grounding	diagnostic result
EvidenceState grammar	adds retrieval intent, support labels, evidence state	repaired agent-data line	diagnostic result
Stage 2A initialization	trains the same EvidenceState contract from the Stage 2A checkpoint	better VQA-close, worse M3D-Cap final-route stability	diagnostic result
RL route policy	reward stopping only when cited slices cover findings	optimize after SFT exposes actions	proposed extension

Table 36: Stage 3 route-grammar ablation scaffold. The main paper treats grammar validity and benchmark quality as separate claims.

Variant	Purpose	Primary measurement	Status
no answer-time 3D review	baseline with standard Active3D placement	task metrics and counterfactual sensitivity	implemented baseline
textual 3D review summary	place compact global summary before evidence/final spans	M3D-Cap support, unsupported claims	proposed
pooled 16/32/64 3D review tokens	reinsert compact visual 3D state near answer generation	CAP, grounding sensitivity, memory	diagnostic result: CAP32 protocol pass; CAP200 M 7.82 below P4
evidence-state 3D reference	use evidence state as compact support bridge	support consistency, final quality	partially implemented
active 3D review tool	let model request refreshed 3D context	route quality, visual reliance	proposed extension

Table 37: Answer-time 3D review ablation scaffold. This tests whether long CAP outputs suffer because global 3D evidence is too far from final synthesis.

MoE line	Train scope	Design	Data / init	Status
Dense Stage 3 EvidenceState	LM LoRA only	current repaired EvidenceState agent line	GPT-repaired Stage 3 EvidenceState data; Stage 2B init	completed anchor
MoE-M1 last2	routed residual FFN experts only	frozen shared FFN path plus 2 small experts, top-1 router, load-balance audit	same GPT-repaired Stage 3 EvidenceState data; Stage 3 EvidenceState init	CAP200: M **12.22**, R1 **18.03**, B1 **12.58**; VQA-close200 Acc. 58.5
MoE-M1 global3	routed residual FFN experts only	same M1 design placed at upper full-attention layers 17/23/29	same GPT-repaired Stage 3 EvidenceState data; Stage 3 EvidenceState init	CAP200: M 11.54, R1 17.20, B1 10.99; VQA-close200 Acc. **63.0**
MoE-M2 task/route gated	routed experts plus explicit task/route conditioning	task/query/final specialist routing	same data plus task/route markers	proposed extension

Table 38: Shared-expert MoE ablation scaffold. This diagnostic does not change the Active3D tensor route or MedGemma native 2D image route; it tests whether LM-side route-specific capacity can preserve medical answering while learning Stage 3 agent behavior. The first M1 diagnostic trained on the GPT-repaired Stage 3-EvidenceState data with balanced expert usage. The last-layer diagnostic slightly increased M3D-Cap overlap but weakened VQA-close, while the global-layer diagnostic improved VQA-close but lowered CAP. We therefore treat M1 as a diagnostic route-capacity result, not as the main reported model.

Case type	Required evidence	Models to compare	Audit label
Native-agent trace	Step 0 prior, `QUERY_SLICES` payload, top-3 slice packet, evidence state, final/withhold decision	Stage 3 EvidenceState; optional Stage 3 RegionHistory	valid query, useful plane/range, supported final answer
M3D-Cap baseline contrast	reference report, real packet slices, final outputs from each system	MedGemma, Med3DVLM, Stage 2A, Stage 3 EvidenceState	finding coverage, correct region, unsupported claim risk
M3D-Cap failure	missing finding or unsupported claim with retrieved packet shown	same model set	weak packet, over-general report, wrong organ/region, missing abnormality
VQA close	question/options, predicted option, queried packet if applicable	MedGemma, Med3DVLM, Stage 2A, Stage 3 EvidenceState	correct option, option-format validity, evidence-dependent correction
Query-more trace	first query, first packet, follow-up plane/range, second packet, final decision	Stage 3 variants	useful query-more, redundant query, wrong plane/range
Counterfactual case	real packet versus blank/off-target/donor-volume packet and resulting evidence state/final answer	Stage 2B/Stage 3	answer changed, citation changed, sufficiency dropped

Table 39: Qualitative example audit schema. Figures use frozen inference outputs and runtime-derived slice images rather than hand-written idealized outputs.

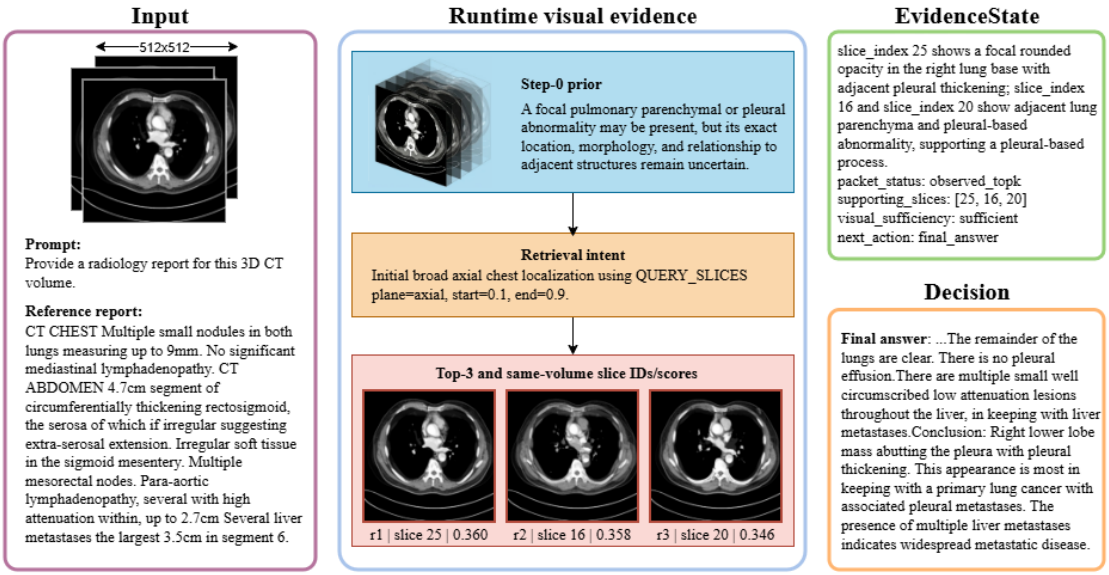

Figure 9: Detailed qualitative case card for a one-turn evidence-supported native-agent trace. The case is clinically partial rather than a perfect CAP match: the model localizes supported thoracic and hepatic findings but still misses part of the reference report. The card is reconstructed from M3D-Cap sample `caption_000057`; all CT thumbnails are runtime-derived same-volume slices.

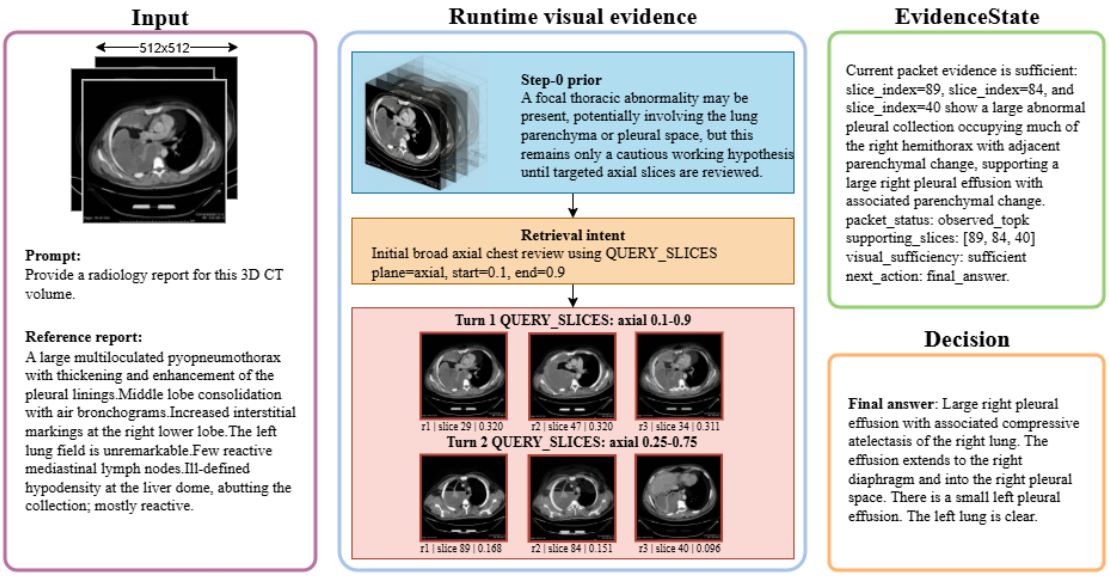

Figure 10: Detailed qualitative case card for a two-turn native-agent trace that reaches a supported final answer after querying additional slices. The case shows useful evidence seeking but remains clinically partial: the model captures the dominant pleural process while simplifying the diagnosis and missing secondary findings. The card is reconstructed from M3D-Cap sample `caption_000105`; all CT thumbnails are runtime-derived same-volume slices.

and the runtime-derived slice-image identifiers. The judge-packet reconstruction pipeline reconstructs queried slices from saved runtime provenance, matching the evidence seen by the model.

I Counterfactual and GPT-Judge Protocol

Official benchmark metrics remain necessary, but they do not fully measure clinical usefulness or visual reliance. We therefore define two supplementary protocols: counterfactual grounding diagnostics and GPT-assisted clinical judging. These are not replacements for benchmark scores. Their role is explanatory: they test whether the agent state changes under evidence perturbations and whether clinical quality remains limited even when the runtime protocol is valid.

The Stage 3 teacher and judge use the same fixed OpenAI-compatible chat-completions backend configured as model alias `gpt-5.5`. All reported teacher/judge calls use temperature 0, high image detail, 5000 teacher tokens, 1200 judge tokens, judge-all filtering, and two repair attempts when needed. The implementation does not set an explicit `reasoning_effort` or `thinking` parameter. Including development, repair, accepted-row, sidecar, and judge calls, the data construction program used approximately 196.6M API tokens: 144.4M input tokens and 52.3M output tokens across 63,875 requests. This budget is reported to make the resource cost of the released trajectory data explicit. Reproducing the teacher program depends on continued availability of the configured API endpoint and a substantial token budget; releasing accepted rows and sidecars decouples downstream training from rerunning the teacher pipeline. The released accepted trajectories contain 13,830 tool calls, with mean/max 1.70/4 calls per trajectory; wall-clock inference cost depends on environment-dependent CT reslicing and MedSigLIP packet-scoring implementation, so we report tool-call depth rather than a non-portable runtime number.

Shared-model confound. The same model alias (`gpt-5.5`) is used for both trajectory construction (teacher) and GPT-assisted judging. We acknowledge this confound and mitigate it in three ways: (1) the judge receives masked model identity and a rubric-anchored scoring protocol (Table 44) rather than open-ended quality assessment; (2) deterministic validators handle format, citation, and route checks independently of the GPT judge; and (3) the counterfactual perturbation design (Tables 42 and 43) tests *relative* changes within the same judge, so absolute score calibration does not affect the perturbation-sensitivity conclusion. A fully independent judge model or human annotator pool would strengthen the

Stage 3 data step	Model/API role	Output kept for audit or SFT
Step 0 repair/generation	`gpt-5.5` teacher with reference-audited offline target construction	cautious 3D prior, retrieval intent, route proposal
Runtime observation	deterministic CT reslicing and MedSigLIP packet scoring	same-volume top-3 slice packet, plane/range/slice provenance
EvidenceState synthesis	`gpt-5.5` teacher conditioned on returned packet	support status, cited slices, sufficiency, next action, final/query-more/withhold target
Judge-all filtering	deterministic validators plus `gpt-5.5` judge	accepted SFT rows or sidecar rows with reject/support flags
Release construction	deterministic merge and audit pipeline	train-ready rows, sidecar audits, provenance keys; no raw CT redistribution

Table 40: Stage 3 teacher/judge data workflow. The reference target is visible only during offline target construction and auditing; student training and deployment-style inference receive the CT volume, query, runtime observation packets, and prior conversation state.

Perturbation	Question answered	Expected grounded behavior
real evidence	normal reference condition	answer uses valid evidence and citations
blank/zero packet	does image presence matter?	lower sufficiency or less specific answer
same-volume off-target packet	does requested location matter?	evidence state or answer changes
donor-volume same-region packet	does volume identity matter beyond region priors?	confidence drops or answer changes
drop cited slice	do cited slices affect the decision?	sufficiency/citation state changes
shuffled/zero 3D slots	does global 3D prior matter?	Step 0 plan or final answer degrades

Table 41: Counterfactual grounding diagnostics. These tests distinguish perturbation-sensitive behavior from fluent but image-invariant reasoning.

absolute-score interpretation; we consider this a limitation of the current diagnostic protocol.

The judge protocol has two modes. The *text-only* mode receives the prompt, reference, model answer, evidence state, and packet metadata; it measures clinical correctness, finding coverage, unsupported claims, and evidence-state coherence, but it is not evidence of visual grounding. The *vision-assisted* mode additionally receives the exact slice packet queried by the student during inference, reconstructed from saved volume, plane, slice-index, and windowing metadata. We use only the vision-assisted mode for visual-evidence claims.

For the reported Stage 3 interface, the C1–C4 diagnostics are implemented as inference-time perturbations over the returned slice packet: blank 2D packet, same-volume off-target packet, donor-volume same-region packet, and drop-cited-slice packet. Donor-volume means the packet provenance has a different `volume_ref` from the target case; we do not require explicit patient identifiers for this diagnostic. Shuffled/zero 3D slots are kept as a separate route ablation because they perturb the Active3D volume slots rather than the native 2D evidence packet. The Stage 2A n=200 3D-slot diagnostic changes 48.0% of outputs under zeroed slots and 50.0% under shuffled slots, supporting the interpretation that the direct 3D route is behaviorally active even though M3D-Cap remains limited by visual-token semantics and evidence coverage.

The 100% EvidenceState-change rate should be interpreted as sensitivity, not complete grounding. Two observations mitigate the concern that this reflects general instability rather than meaningful evidence dependence. First, the perturbation-specific pattern is non-uniform: same-volume off-target packets produce the largest disruption (0.630 sufficiency change, 0.700 runtime drop in CAP200) while drop-support perturbations produce the smallest (0.445, 0.045), which is inconsistent with blanket input fragility. Second, the off-target perturbation changes the *content* of the evidence state (e.g., cited slices and region descriptions change), not merely its length or formatting. Nevertheless, a benign-perturbation control, such as minor intensity jitter or noise-augmented slices, would help more rigorously separate meaningful evidence dependence from residual output instability; we leave that control to future work.

CAP200 perturbation	Final changed	State changed	Suff. changed	Runtime drop	Diagnostic interpretation
blank 2D packet	0.985	1.00	0.515	0.065	Image removal changes the trajectory, but M3D-Cap overlap can remain language-prior dominated.
same-volume off-target packet	0.960	1.00	0.630	0.700	The largest disruption: packet location materially affects the agent trajectory under perturbation.
donor-volume same-region packet	0.985	1.00	0.485	0.260	Volume-specific evidence changes the output beyond a coarse region prior.
drop-support slice	0.995	1.00	0.445	0.045	Citation/support perturbations are visible to the EvidenceState, but finalization can still occur.

Table 42: Stage 3-EvidenceState CAP200 counterfactual sensitivity. Values are paired rates against the real-evidence condition. The real condition has final/runtime rate 0.94 and CAP METEOR 0.1171; the same-volume off-target perturbation reduces runtime success to 0.26 and METEOR to 0.0333.

VQA-close100 perturbation	Final changed	State changed	Choice changed	Runtime drop	Diagnostic interpretation
blank 2D packet	0.34	1.00	0.34	0.03	The final letter is partly robust to blanking, but the evidence state changes in every paired case.
same-volume off-target packet	0.48	1.00	0.48	0.27	The largest VQA perturbation; real accuracy 0.58 falls to 0.48 and runtime success falls from 0.94 to 0.72.
donor-volume same-region packet	0.43	1.00	0.42	0.08	Volume provenance affects many outputs even when the coarse region is held fixed.
drop-support slice	0.27	1.00	0.27	0.02	Single-slice removal is visible to the state but less often changes the multiple-choice letter.

Table 43: Stage 3-EvidenceState VQA-close100 counterfactual sensitivity. Values are paired rates against the real-evidence condition. Multiple-choice letters are less sensitive than M3D-Cap text, but EvidenceState changes in 100/100 paired cases for every perturbation.

Judge dimension	Definition	Score format	Modes
clinical correctness	final answer/report agrees with the reference finding set	1–5	both
finding coverage	report mentions important positive and negative findings	1–5	both
anatomy/region binding	findings are attached to the correct organ, side, and region	1–5	both
unsupported claims	output avoids hallucinated findings not supported by reference or evidence	count	both
evidence use	answer is consistent with cited packet/evidence state	1–5	vision preferred
evidence-bound finding rate	reference findings both mentioned and supported by visible/evidence-state support	0–1	vision preferred
clinical usefulness	output would help a clinician more than a generic template	1–5	both

Table 44: GPT-assisted M3D-Cap judging rubric. Model identity is masked from the judge. Text-only judging measures reference/evidence-state consistency; vision-assisted judging additionally receives the exact queried slice packet and is the only setting used for visual-evidence claims.

System	Judge mode	Clin. corr.	Coverage	Binding	Evid. use	EBF	Unsup.	Useful.
MedGemma baseline	text-only	1.015	1.010	1.050	1.000	0.0005	5.000	1.010
MedGemma-runtime	text-only	1.000	1.000	1.000	1.005	0.0000	0.810	1.000
Stage 2A-Active3D	text-only	1.385	1.385	1.460	1.015	0.0079	5.040	1.350
Med3DVLM-runtime	text-only	1.000	1.000	1.000	1.000	0.0000	0.000	1.000
Stage 3 EvidenceState	text-only	1.125	1.085	1.310	1.565	0.0359	5.515	1.105
Stage 3 EvidenceState	vision-assisted	1.165	1.090	1.335	1.275	0.0518	5.505	1.150

Table 45: GPT-assisted M3D-Cap clinical judging diagnostic. Values shown are CAP200 fixed-prefix results. The MedGemma baseline row uses the single-pass S16 axial-montage CAP output. Stage 2A and Stage 3 rows use official CAP200 outputs; rows marked *runtime* use the MedAura runtime-transfer stress-test outputs rather than single-pass benchmark predictions. Systems without valid queried packet provenance are judged in text-only mode. Stage 3 EvidenceState also supports vision-assisted judging because its runtime records the exact packet observed before final synthesis. These values diagnose clinical weakness and evidence-state behavior and should not be read as benchmark replacements.

Counterfactual condition	n	Clin. corr.	Coverage	Binding	Evid. use	EBF	Unsup.	Useful.
real evidence	200	1.110	1.075	1.290	1.895	0.038	5.305	1.100
blank 2D packet	200	1.110	1.045	1.215	1.200	0.015	6.060	1.055
same-volume off-target packet	200	1.280	1.035	1.685	3.130	0.013	1.565	1.415
donor-volume same-region packet	200	1.155	1.055	1.665	2.185	0.020	4.890	1.125
drop cited/supporting slice	200	1.110	1.070	1.250	1.615	0.026	5.355	1.090

Table 46: Counterfactual GPT-judge clinical-degradation table. The reported run uses CAP200 fixed-prefix samples and five conditions, producing 1000 text-only judge tasks. Scores remain low across conditions, so this is interpreted as a diagnostic of evidence dependence rather than a claim of clinical report quality. Blank and drop-support packets reduce judged evidence use; off-target packets often make outputs more conservative, reducing unsupported claims despite not improving finding coverage.

J Future Stage 4 RL Interface

Stage 4 should optimize route quality rather than free-form fluency: answer when evidence is sufficient, query when uncertainty remains, and withhold when evidence is unsupported. Candidate rewards combine final-answer correctness, evidence-state consistency, citation validity, visual sufficiency calibration, counterfactual sensitivity, and tool format.

The current reward scaffold uses $R = 0.25R_{\text{task}} + 0.20R_{\text{evid}} + 0.20R_{\text{query}} + 0.15R_{\text{cf}} + 0.10R_{\text{stop}} + 0.05R_{\text{fmt}} + 0.05R_{\text{hall}}$. For close-ended VQA, a wrong parsed choice caps reward at 0.35 even if the tool call executed correctly. GRPO can use group-relative advantages over same-prompt trajectories; no RL-trained model is claimed.