3DMEDEVO: SELF-EVOLVING VLM AGENTS WITH EVIDENCE-GROUNDED MULTIMODAL SKILLS FOR 3D MEDICAL IMAGE ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

3D CT analysis integrates volumetric perception with clinical reasoning, and improves as experience accumulates over successive readings. Existing 3D medical methods typically use task-specific models or task-agnostic end-to-end predictors and rely on external teacher supervision. Memory-augmented approaches instead reuse experience across cases but often retain raw, noisy, and poorly curated interaction traces as text that does not fully preserve spatial detail. More importantly, they lack adoption-aware estimates of memory utility for future reasoning. Here, we present **3DMedEvo**, a **self-evolving runtime framework** that distills informative trajectories from frozen vision–language model (VLM) agents reading CT scans into structured multimodal skills through outcome-guided reflection, without additional teacher supervision. The resulting skills combine reusable procedures with visual exemplars in a repository organized into multiple branches, with guidance grounded in clinical evidence. 3DMedEvo further calibrates skill reliability from observed utility using an adoption-aware **Skill Credit Score (SCS)**, which guides retrieval for each task and uncertainty-aware governance. In turn, retrieved skills support reasoning by frozen VLM agents within a closed loop of skill acquisition, evaluation, and revision. Experiments on two established clinical benchmarks covering thoracic and abdominal CT show that 3DMedEvo outperforms reasoning agents using tools and representative agents using memory across recognition, quantitative measurement, visual reasoning, and medical reasoning tasks. The framework also generalizes across VLM backbones, offering a practical approach to self-evolution in 3D medical image analysis. All code and data will be made publicly available.

1 INTRODUCTION

Clinical 3D CT analysis is a dynamic, multi-step process. As evidence accumulates across slices and measurements, diagnostic hypotheses and reading priorities may change (Rubin et al., 2015; Bertram et al., 2016; Elstein Arthur & Schwarz, 2002; Barrows & Feltovich, 1987). To guide these adjustments, clinicians draw on prior case experience, linking findings to plausible interpretations and reading strategies (Norman, 2005; Patel & Itri, 2022; Norman et al., 2007). In turn, comparison and reflection distill this experience into reusable guidance, which feedback from later cases can refine (Charlin et al., 2007; Mamede et al., 2014). Despite advances in AI-assisted 3D CT analysis (Wasserthal et al., 2023; Hamamci et al., 2026; Li et al., 2026a), fixed reading policies and sensitivity to domain shifts can constrain adaptation to unfamiliar settings (Wang et al., 2026a). Therefore, experience-driven learning is needed to continually integrate medical knowledge with case-derived experience (Kolodner, 1992; Schmidt & Rikers, 2007), forming and refining reusable procedures to guide decisions in situations not fully covered by training data or textbooks.

Earlier approaches largely addressed isolated 3D image analysis tasks in individual clinical cases. Task-specific models provide anatomical segmentation and disease detection (Wasserthal et al., 2023; Cao et al., 2023), while vision–language models (VLMs) learn image–text representations for broader clinical interpretation (Blankemeier et al., 2026; Hu et al., 2025). However, these methods rely on computationally costly parameter updates and often require teacher-generated supervision (Wu et al., 2025a; Bai et al., 2024; Li et al., 2026a). Moreover, their black-box, single-step pre-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

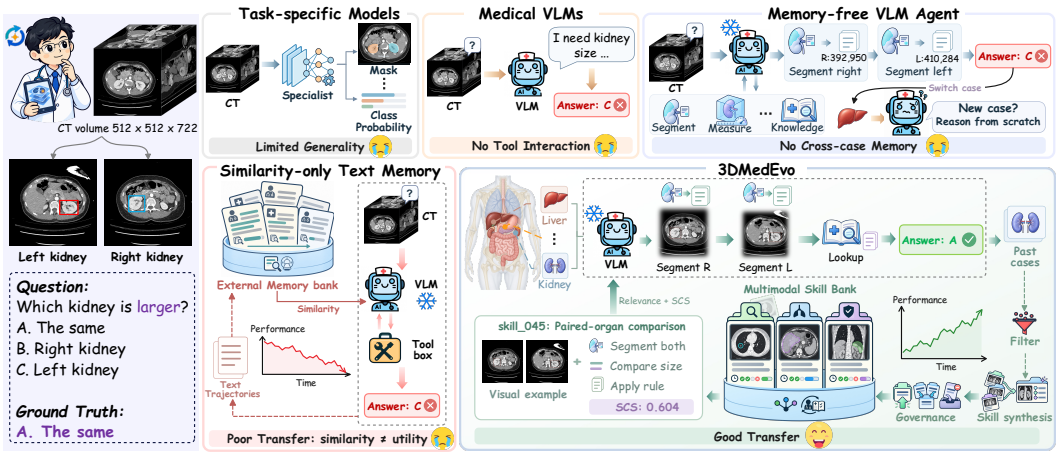


Figure 1: **Comparison of 3D medical image analysis paradigms.** Workflows of task-specific models, medical VLMs, memory-free VLM agents, similarity-only text memory, and 3DMedEvo.

dictions lack an explicit evidence-to-decision trace, hindering clinical interpretability (Wang et al., 2026a). To further ground clinical reasoning, tool-augmented agents enable VLMs to invoke specialized imaging tools and verify uncertain findings. For example, 3DMedAgent integrates anatomical measurements with regional and slice-level inspection, using findings to guide subsequent tool calls (Wang et al., 2026a). Nevertheless, its memory primarily supports case-specific reasoning rather than the accumulation and revision of transferable procedures across cases.

To address this gap, memory-based methods retain and retrieve experience across tasks, extending memory beyond the context of a single case (Wang et al., 2024; Suzgun et al., 2026; Zhang et al., 2026b) (Figure 1). However, storing past cases or trajectories (Guo et al., 2024; Shi et al., 2024; Zhou et al., 2025) often produces redundant, instance-specific memories, limiting their transferability across tasks. Recent approaches therefore distill execution experience into reusable insights and procedural skills that capture recurring patterns of reasoning and action (Zhao et al., 2024; Ni et al., 2026; Jiang et al., 2026). Such guidance is often refined through textual feedback on execution outcomes (Fang et al., 2026; Zhang et al., 2026a). In medicine, structured experience memories likewise organize clinical reasoning for reuse across cases (Han et al., 2026; Sun et al., 2026). However, three challenges remain in reusing CT-reading experience. First, case-based and skill-based memories often lack long-term utility assessment, allowing redundant or low-quality entries to accumulate. Second, text-centric abstraction suits language-only agents but loses spatial detail needed for 3D CT reasoning. Third, case-derived guidance requires clinical evidence.

In this work, we introduce **3DMedEvo**, a memory-based self-evolving framework that learns **evidence-grounded multimodal skills** for 3D medical image analysis without updating VLM parameters. It distills completed reading trajectories into **reusable skills** through outcome-guided reflection, without an additional teacher model. These skills are organized into three complementary branches for locating evidence, interpreting findings, and judging whether the evidence supports a conclusion. Each skill pairs a scoped procedure with visual exemplars, while its medical claims are checked against clinical evidence. 3DMedEvo then assigns each skill an adoption-aware **Skill Credit Score (SCS)**, an estimate of reuse value based on observed use and task outcomes. In read-only deployment, 3DMedEvo performs value-aware retrieval from the fixed bank, combining semantic relevance with this credit to guide the frozen VLM. To further refine the bank, posterior intervals temper promotion and retirement under sparse evidence, while merging reduces redundancy. This establishes a lifecycle for accumulating, refining, and reusing 3D CT-reading experience, integrating medical knowledge with feedback from subsequent cases. Our contributions are as follows:

- We introduce **3DMedEvo**, a teacher-free framework for 3D medical image analysis that integrates case experience with clinical knowledge and feedback to refine reading strategies without updating VLM parameters.
- We establish a **closed-loop lifecycle** for CT-reading skills from acquisition and evaluation to revision and reuse, integrating adoption-aware **SCS** with posterior-interval governance.

- Experiments on thoracic and abdominal CT benchmarks demonstrate **state-of-the-art** accuracy across the evaluated VLM backbones and support skill reuse across heterogeneous environments.

2 Related Work

Medical Vision–Language Models. Medical VLMs connect imaging findings with clinical language through image–text pretraining (Zhang et al., 2025), while medical MLLMs develop clinical reasoning through instruction tuning or reinforcement learning (Sellergren et al., 2025; Li et al., 2025). Extending these capabilities to 3D CT, RadFM unifies 2D and 3D radiological inputs, while M3D supports retrieval, question answering, localization, and segmentation (Wu et al., 2025a; Bai et al., 2024). Fine-grained CT–report grounding and volume-level alignment enable language-guided analysis at regional and whole-volume scales (Lin et al., 2024; Hamamci et al., 2026); Merlin adds electronic health records to CT–report supervision (Blankemeier et al., 2026). However, volumetric training requires substantial computation and paired or annotated CT data (Wang et al., 2026a). 3DMedEvo evolves CT-reading skills with a frozen VLM and no additional teacher model.

Medical Agents. Medical agents combine image inspection in radiology (Fallahpour et al., 2025) with expert discussion and adaptive collaboration (Tang et al., 2024; Kim et al., 2024) and tool-based clinical workflows (Li et al., 2024a; Wang et al., 2026b). Agentic reinforcement learning supports interleaved reasoning and tool use over medical images and videos (Li et al., 2026b; Jiang et al., 2025b). For CT, 3DMedAgent assembles case-level evidence through global-to-local inspection (Wang et al., 2026a). Other systems learn from simulated practice and workflow feedback (Li et al., 2024b; Zhu et al., 2025) or combine tool-sequence reuse and image-feature memory with policy training (Fan et al., 2026). 3DMedEvo proposes a memory-based self-evolution paradigm for 3D medical image analysis, turning reading experience into skills reusable across patients.

Self-Evolving Memory and Skills. Self-evolving and memory-augmented agents can reuse interaction experience and external knowledge to improve subsequent decisions without retraining for each task (Zhao et al., 2024; Zhang et al., 2026b). Reflection and tool-based correction refine responses (Shinn et al., 2023; Gou et al., 2024), while episodic and linked memories support experience reuse (Zhou et al., 2025; Xu et al., 2026b). Recent systems organize experience into workflows, procedural memories, multimodal skills, and clinical memory graphs (Wang et al., 2024; Fang et al., 2026; Ni et al., 2026; Zheng et al., 2025; Jiang et al., 2026; Han et al., 2026), evolving skills through hierarchical refinement, policy learning, or external editing (Li et al., 2026c; Xia et al., 2026; Yang et al., 2026). Memory-based methods use utility estimates to guide reuse (Zhang et al., 2026b; Sun et al., 2026). However, utility scores alone neither provide clinical grounding nor quantify uncertainty under sparse adoption. 3DMedEvo addresses these gaps through a complete self-evolution loop for 3D medical imaging, grounding multimodal skills in clinical evidence and coupling the Skill Credit Score with posterior intervals to guide reuse and evolution.

3 Method

3.1 Problem Formulation

We formulate 3DMedEvo as a memory-based Markov decision process (M-MDP) (Zhou et al., 2025), where memory guides 3D medical image analysis and evolves through interaction feedback.

Memory-Based Markov Decision Process. Given a 3D medical image volume V and a clinical question q, the agent maintains an interaction state $h_t \in \mathcal{H}$ comprising the task input and its action–observation history. A memory bank $\mathcal{K} \in \mathbb{K}$ retains experience from prior interactions, where $\mathbb{K}$ is the space of finite memory banks. The agent retrieves relevant memory $\mathcal{C} \subseteq \mathcal{K}$ to inform its next action. A vision–language model (VLM) p_θ generates an action $a_t \in \mathcal{A}$ according to $a_t \sim p_\theta(\cdot \mid h_t, \mathcal{C})$, encompassing reasoning decisions, tool calls, and final responses. The reward function R provides feedback $r_t = R(h_t, a_t)$, and the transition kernel T determines the next state, $h_{t+1} \sim T(\cdot \mid h_t, a_t)$. These components define the M-MDP:

$$\mathfrak{D}_{\text{mem}} = \langle \mathcal{H}, \mathcal{A}, T, R, \mathbb{K} \rangle. \tag{1}$$

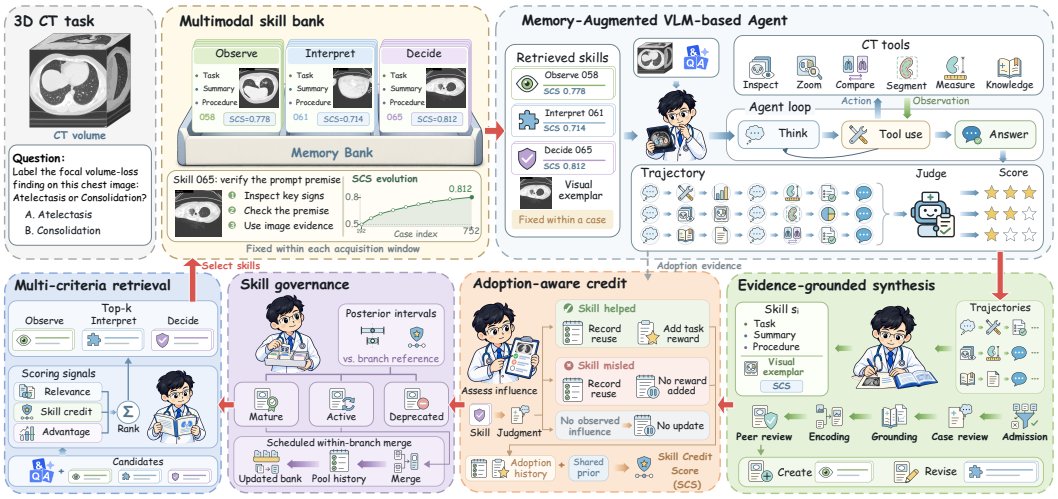


Figure 2: **Overview of 3DMedEvo.** CT-reading experience yields multimodal skills. Red arrows connect synthesis, credit updates, governance, and retrieval in a closed loop for cross-case reuse.

A trajectory τ records states, retrieved memory, actions, and feedback. A memory-update operator $\mathcal{U}$ incorporates this experience through $\mathcal{K}^+ = \mathcal{U}(\mathcal{K}, \tau)$. Under this formulation, adaptation proceeds through memory retrieval and evolution without updating the VLM parameters θ.

Memory-Augmented VLM-based Agent. Under the M-MDP formulation, 3DMedEvo maintains a memory bank $\mathcal{K}$ of skills distilled from prior cases. This reflects how clinicians consolidate image-reading experience into shared, reusable guidance. Each skill s_i contains an applicability description d_i, a brief experience summary b_i, and a transferable procedure ℓ_i, with an associated visual exemplar library $\mathcal{V}_i$. The effective adoption count n_i and accumulated credit c_i determine the Skill Credit Score SCS_i, which summarizes the skill's reliability. For N skills, the bank is represented as:

$$\mathcal{K} = \{s_i\}_{i=1}^{N}, \qquad s_i = (d_i, b_i, \ell_i, \mathcal{V}_i, n_i, c_i). \tag{2}$$

Across cases, $\mathcal{U}$ updates the bank with experience distilled from interaction trajectories and feedback. Given the initial task state h_0, 3DMedEvo selects a skill subset $\mathcal{C} \subseteq \mathcal{K}$ according to task relevance and historical adoption evidence. Their model-visible content serves as guidance. The resulting agent policy π is:

$$\pi(a_t \mid h_t, \mathcal{K}) = p_\theta(a_t \mid h_t, \mathcal{C}). \tag{3}$$

3.2 Multimodal Skill Memory

As illustrated in Figure 2, 3DMedEvo accumulates reusable CT-reading procedures and visual exemplars in an evolving skill bank while keeping the VLM frozen. Raw trajectories can retain redundant, instance-specific details (Sun et al., 2026; Zhao et al., 2024). Lengthy experience descriptions also consume context budget and can make relevant information harder to access (Liu et al., 2024). We therefore distill compact skills specifying applicability, operations, and verification rules. We detail retrieval (Section 3.3), skill synthesis (Section 3.4), adoption-aware credit (Section 3.5), and governance (Section 3.6), then integrate them into closed-loop evolution (Section 3.7).

Procedural branches. Radiologic error analyses distinguish visual search, recognition, and diagnostic decision errors (Patel et al., 2019). These distinctions motivate organizing skills around evidence acquisition, interpretation, and verification as complementary roles, not fixed stages. (1) **Observe branch** addresses visual search errors in locating relevant evidence. (2) **Interpret branch** addresses recognition errors in identifying and understanding observed findings. (3) **Decide branch** addresses diagnostic decision errors in judging whether the evidence supports a conclusion.

Layered multimodal skills. Volumetric CT interpretation relies on morphology and spatial relationships that compact textual rules may not fully convey. Each skill therefore pairs visual exemplars $\mathcal{V}_i$, which preserve views observed in prior cases as references for appearance and anatomical context, with a symbolic layer (d_i, b_i, ℓ_i) that retains transferable procedures. Together, the two layers provide context-relevant guidance while limiting information loss from text-only abstraction.

3.3 MULTI-CRITERIA SKILL RETRIEVAL

For a clinical question q, 3DMedEvo scores all available skills using three criteria: task relevance, adoption reliability, and relative advantage. Relevance identifies applicable experience as $R_i(q) = \cos(\psi(q), \psi(d_i))$, using a fixed text encoder ψ. Similarity alone does not establish reliability; SCS_i incorporates evidence from actual adoption and task outcomes, as detailed in Section 3.5. Since task rewards also reflect case difficulty, the relative advantage A_i of skill s_i summarizes its adoption-aware feedback against a historical reward baseline for comparable tasks. We normalize the three criteria to comparable scales with z_B and combine them using non-negative weights that sum to one:

$$\mathrm{score}(q, s_i; \mathcal{K}) = w_R z_B(R_i(q)) + w_{\mathrm{SCS}} z_B(\mathrm{SCS}_i) + w_A z_B(A_i). \tag{4}$$

We retain non-negative-scoring skills in descending order, with at most k per branch. The subset $\mathcal{C}$ remains fixed within the case. Details are provided in Appendix C.1.

3.4 EVIDENCE-GROUNDED SKILL SYNTHESIS

To base skill updates on informative experience, we maintain a gated trajectory buffer $\mathcal{B}^{(w)}$ for each acquisition window w of L cases. Let $\mathcal{D}^{(w)}$ denote the judged completed trajectories generated with skill bank $\mathcal{K}^{(w)}$, which remains fixed throughout the window.

Experience selection and review. Admission $\mathcal{I}$ filters infrastructure errors, uninformative repetitions, and uninformative successes:

$$\mathcal{B}^{(w)} = \{\tau \in \mathcal{D}^{(w)} : \mathcal{I}(\tau) = 1\}. \tag{5}$$

To extract a transferable lesson from each trajectory, Case Review $\mathcal{R}$ examines the reading process and outcome feedback for mechanisms behind success or failure. It produces up to three proposals $\rho_\tau \in \mathcal{R}\big(\tau; \mathcal{K}^{(w)}\big)$, each giving the lesson, a creation or local-revision intent, and the revision target.

Evidence-grounded encoding. Clinical guideline development couples evidence appraisal with the formulation of clear, actionable recommendations (Alonso-Coello et al., 2016; Brouwers et al., 2010). In 3DMedEvo, $\mathcal{G}$ checks the medical claims in each proposal against public clinical guidelines and research evidence, yielding $\tilde{\rho}_\tau = \mathcal{G}(\rho_\tau)$. No additional teacher supervision is introduced. The checked proposal $\tilde{\rho}_\tau$ retains uncertainty when support is insufficient and is rejected when contradicted. To make the retained lesson reusable, encoding produces a layered multimodal skill $\hat{s}_\tau = \mathcal{E}(\tilde{\rho}_\tau, \tau)$, pairing a scoped procedure and verification conditions with exemplars of morphology and anatomical context observed in τ.

Candidate acceptance. Before release, peer review $\mathcal{P}$ checks guidance for overstated evidence and unclear applicability. It assesses source fidelity, actionability, and same-branch novelty, accepting or rejecting without rewriting. For each trajectory, the skill-update operator $\mathcal{W}$ returns the accepted candidate update $\Delta_{\mathrm{skill}}\big(\mathcal{K}^{(w)}, \tau\big)$:

$$\Delta_{\mathrm{skill}}\big(\mathcal{K}^{(w)}, \tau\big) = \mathcal{W}\big(\mathcal{K}^{(w)}, \tau\big) = \bigcup_{\rho_\tau} \begin{cases} \{(\varnothing, \hat{s}_\tau)\}, & \text{creation accepted by } \mathcal{P}, \\ \{(s_{\rho_\tau}, \hat{s}_\tau)\}, & \text{revision accepted by } \mathcal{P}, \\ \varnothing, & \text{otherwise.} \end{cases} \tag{6}$$

Here, $s_{\rho_\tau} \in \mathcal{K}^{(w)}$ is the revision target specified by ρ_τ. At the window level, the accepted updates form $\Delta^{(w)}_{\mathrm{skill}} = \bigcup_{\tau \in \mathcal{B}^{(w)}} \mathcal{W}\big(\mathcal{K}^{(w)}, \tau\big)$. Applying these updates yields a candidate bank $\widetilde{\mathcal{K}}^{(w)}$ through the synthesis operator $\mathcal{S}$:

$$\widetilde{\mathcal{K}}^{(w)} = \mathcal{S}\big(\mathcal{K}^{(w)}, \Delta^{(w)}_{\mathrm{skill}}\big). \tag{7}$$

3.5 ADOPTION-AWARE SKILL CREDIT

3DMedEvo synthesizes skills from individual CT readings, but their value lies in reliable reuse across cases. For a retrieved skill s_i in trajectory τ, Admission (Section 3.4) provides the adoption label $g_{\tau,i} \in \{+1, -1, 0\}$: $+1$ denotes actual use supporting a correct key judgment, -1 denotes actual use leading to an incorrect key judgment, and 0 denotes no observable influence. Let r_τ denote the trajectory's task evaluation score, based on correctness or a task-specific rubric, and n_i, c_i the skill's effective adoption count and accumulated credit. To temper estimates from sparse

adoption histories, we combine the updated statistics with a shared prior of mean $v_0 \in (0, 1)$ and strength $\lambda > 0$ to compute the Skill Credit Score (SCS):

$$n_i \leftarrow n_i + \mathbf{1}[g_{\tau,i} \neq 0], \qquad c_i \leftarrow c_i + \mathbf{1}[g_{\tau,i} = +1] r_\tau, \tag{8}$$

$$\mathrm{SCS}_i = \frac{n_i}{n_i + \lambda} \frac{c_i}{n_i} + \frac{\lambda}{n_i + \lambda} v_0, \qquad n_i > 0. \tag{9}$$

All actual adoptions increase n_i, but only positive adoption contributes r_τ to c_i. As n_i increases, empirical performance receives more weight, while the score becomes less sensitive to individual adopted trajectories. Details are provided in Appendix C.3.

3.6 Posterior-Interval Governance

To maintain a compact bank of reusable skills, 3DMedEvo performs governance at scheduled window boundaries. For each retrieval-eligible skill s_i, 3DMedEvo derives an uncertainty interval $[L_i, U_i]$ around its credit SCS_i from updated adoption statistics. The lower and upper bounds give conservative and optimistic estimates; for the same score, fewer adoptions yield a wider interval (Appendix C.4). For each branch b, a reference rank K_b determines the governance reference β_b from the current retrieval-eligible pool. The governance operator $\mathcal{Q}$ compares each skill's credit interval with this reference:

$$\mathcal{Q}(s_i; \beta_b) = \begin{cases} \text{mature}, & L_i > \beta_b, \\ \text{deprecated}, & U_i < \beta_b, \\ \text{unchanged}, & \text{otherwise}. \end{cases} \tag{10}$$

Promotion uses the conservative estimate; deprecation requires the optimistic estimate to be uncompetitive. This tempers lifecycle changes when adoption evidence is limited. Within-branch merging pools credit statistics and deprecates lower-SCS near-duplicates. For a bank, $\mathcal{Q}$ applies the branch-wise lifecycle and consolidation rules, returning it unchanged between governance boundaries.

3.7 Closed-Loop Evolution

The above framework integrates retrieval, multimodal skill synthesis, adoption-aware credit, and governance into a unified closed loop for skill evolution. Let $\mathcal{F}$ denote the credit-update operator. We summarize the full process as the following operator composition:

$$\mathcal{K}^{(w+1)} = \mathcal{Q}\Big(\mathcal{F}\Big(\mathcal{S}\big(\mathcal{K}^{(w)}, \Delta^{(w)}_{\mathrm{skill}}\big), \mathcal{D}^{(w)}\Big)\Big). \tag{11}$$

During acquisition, each window uses a fixed bank $\mathcal{K}^{(w)}$. The updated bank $\mathcal{K}^{(w+1)}$ guides retrieval in the next window, whose feedback informs subsequent skill updates. After acquisition, 3DMedEvo enters read-only deployment, using the same retrieval rule to select skills. The final skill bank and its statistics remain fixed throughout deployment. Algorithm 1 details the complete procedure.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate 3DMedEvo on 3D CT question-answering benchmarks: DeepChestVQA (Wang et al., 2026a) for thoracic abnormality assessment and DeepTumorVQA (Chen et al., 2025; 2026) for abdominal tumor analysis. In Frozen Memory, all memory-based methods construct memories from DeepTumorVQA's training split, excluding evaluation-overlapping CT volumes. Memories and statistics remain fixed during in-domain testing on held-out DeepTumorVQA cases and out-of-domain testing on DeepChestVQA. In Evolving Memory, skills are accumulated and updated during evaluation. Benchmark and training-set details appear in Appendix H.

Implementation Details. 3DMedEvo uses GLM-5.2 (Z.ai, n.d.) and Qwen3.6-Plus (Alibaba Cloud, 2026b) as its primary backbones, with each frozen model shared across task execution and all LLM-based stages of skill evolution. We encode questions and skill applicability descriptions with text-embedding-v4 (Alibaba Cloud, 2026a). Our default retrieval configuration uses ranking weights

Table 1: **Frozen-memory accuracy (%) on DeepChestVQA and DeepTumorVQA.** Overall superscripts show changes from same-backbone ReAct. **Bold**/underlined scores denote the best/second-best per metric, separately for GLM-5.2 and Qwen3.6-Plus.

Backbone	Method	DeepChestVQA				DeepTumorVQA				
		Recognition	Visual Reasoning	Medical Reasoning	Overall	Recognition	Measurement	Visual Reasoning	Medical Reasoning	Overall
Closed-Source VLMs										
Claude Sonnet 4.6	ReAct	48.28	35.03	44.94	40.55	49.72	43.47	39.15	47.43	43.74
GPT-5.5	ReAct	51.72	36.90	42.70	41.22	58.24	44.03	44.68	47.13	47.76
Gemini-3-Flash	ReAct	62.93	40.91	44.57	45.57	56.53	48.30	48.51	57.70	51.84
Grok-4.5	ReAct	49.14	39.04	41.95	41.61	59.38	44.60	43.69	55.59	49.31
Open-Source Medical VLMs										
MedGemma-27B	ReAct	47.41	27.81	49.06	38.31	56.82	39.77	38.87	29.91	40.98
Lingshu-32B	ReAct	50.00	25.94	46.44	36.86	50.57	34.38	37.59	32.02	38.51
HuluMed-32B	ReAct	47.41	27.01	50.94	38.57	53.12	38.92	37.87	36.56	40.92
RadFM-14B	ReAct	48.28	24.33	36.70	32.36	52.84	23.01	33.48	37.76	36.09
3D Medical Agent Baselines										
GPT-5.5	3DMedAgent	57.76	38.77	42.70	43.06	55.40	36.93	34.18	40.48	40.23
Qwen3.6-Plus	3DMedAgent	56.03	31.82	49.06	41.61	53.13	31.53	38.01	31.72	38.56
Memory-Based Agent Baselines										
GLM-5.2	ReAct	38.79	31.28	44.57	37.12	53.41	44.60	40.43	51.66	46.03
	Reflexion	42.24	37.17	44.94	40.69$^{+3.57}$	57.95	46.88	40.57	49.55	47.07$^{+1.04}$
	CRITIC	38.79	34.22	**46.07**	39.10$^{+1.98}$	59.38	47.73	42.70	51.96	48.85$^{+2.82}$
	ExPeL	38.79	34.49	41.95	37.78$^{+0.66}$	59.94	46.02	43.69	47.73	48.22$^{+2.19}$
	AWM	42.24	35.29	44.57	39.63$^{+2.51}$	59.94	44.60	37.73	48.34	45.63$^{-0.40}$
	Evolver	43.10	37.43	44.19	40.69$^{+3.57}$	59.38	45.74	39.72	54.98	47.82$^{+1.79}$
	DC	39.66	37.17	43.82	39.89$^{+2.77}$	59.09	44.03	40.28	46.22	45.98$^{-0.05}$
	GSEM	43.10	38.50	44.57	41.35$^{+4.23}$	59.38	48.86	42.13	48.34	48.16$^{+2.13}$
	SkeMex	43.97	32.09	45.69	38.71$^{+1.59}$	**61.93**	48.86	40.71	54.98	49.37$^{+3.34}$
	3DMedEvo	**58.62**	**43.85**	45.32	**46.63**$^{+9.51}$	56.25	**53.41**	**50.07**	**58.91**	**53.68**$^{+7.65}$
Qwen3.6-Plus	ReAct	51.72	30.48	49.06	40.29	49.15	46.59	44.26	47.43	46.32
	Reflexion	50.00	33.69	41.95	39.10$^{-1.19}$	60.51	46.02	42.13	52.57	48.62$^{+2.30}$
	CRITIC	55.17	34.76	43.45	40.95$^{+0.66}$	56.53	44.60	40.85	50.76	46.67$^{+0.35}$
	ExPeL	50.86	29.41	43.07	37.52$^{-2.77}$	57.67	46.88	41.70	49.24	47.41$^{+1.09}$
	AWM	51.72	35.29	45.69	41.48$^{+1.19}$	59.66	45.17	45.96	55.89	50.46$^{+4.14}$
	Evolver	53.45	33.42	43.82	40.16$^{-0.13}$	60.23	44.89	43.83	49.85	48.51$^{+2.19}$
	DC	50.86	25.67	42.70	35.54$^{-4.75}$	57.67	44.03	39.15	44.71	44.94$^{-1.38}$
	GSEM	52.59	32.09	43.82	39.37$^{-0.92}$	59.09	47.44	43.40	49.85	48.62$^{+2.30}$
	SkeMex	48.28	29.95	42.32	37.12$^{-3.17}$	**61.93**	45.17	44.11	53.47	49.71$^{+3.39}$
	3DMedEvo	**62.07**	**42.25**	**50.94**	**48.35**$^{+8.06}$	60.51	**51.14**	**49.79**	**62.84**	**54.71**$^{+8.39}$

$(w_R, w_{\text{SCS}}, w_A) = (0.6, 0.3, 0.1)$, clipping bound $B = 2$, non-negative score retention, and a per-branch quota of $k = 2$. For credit, we set the SCS prior to $v_0 = 0.5$ and $\lambda = 2$, use $\lambda_a = 2$ for advantage shrinkage, and retain revision history with $\gamma = 0.5$. Governance uses posterior multiplier $z = 1$ and runs every three acquisition windows of $L = 32$ cases. Full settings and hyperparameter sensitivity are deferred to Appendices G and B.6.

Baselines & Metrics. We compare 3DMedEvo with (i) medical specialist models as domain-specific references (Sellergren et al., 2025; Xu et al., 2026a; Jiang et al., 2025a; Wu et al., 2025a); (ii) memory-free tool-using agents (Yao et al., 2022; Wang et al., 2026a); (iii) reflection and self-correction methods (Shinn et al., 2023; Gou et al., 2024); and (iv) experience-based memory agents (Zhao et al., 2024; Wang et al., 2024; Wu et al., 2025b; Suzgun et al., 2026; Han et al., 2026; Sun et al., 2026). All baselines have access to the same tool suite for fair comparison (Appendix D). Baseline details are provided in Appendix E. We use Gemini-2.5-Flash (Google DeepMind, 2025) as an answer-only judge to extract option letters and report exact-match accuracy (Appendix F).

4.2 Main Results

Frozen Memory. We first evaluate accuracy on in-domain DeepTumorVQA and out-of-domain DeepChestVQA with two frozen backbones and a fixed skill library. As shown in Table 1, 3DMedEvo achieves higher overall accuracy than the strongest closed-source VLM and specialized medical models on both benchmarks. This advantage also holds against memory-based agents

Table 2: **Evolving-memory accuracy (%) on DeepChestVQA and DeepTumorVQA.** R1–R3 denote evaluation rounds. Δ reports the change from the first to the third round.

Method	DeepChestVQA Qwen3.5-27B				DeepChestVQA GLM-5.2				DeepTumorVQA Qwen3.6-Plus			
	R1	R2	R3	Δ	R1	R2	R3	Δ	R1	R2	R3	Δ
Reflexion	35.40	36.46	37.12	+1.72	35.67	36.86	35.54	-0.13	48.96	49.08	49.13	+0.17
CRITIC	38.31	38.04	38.31	0.00	38.57	36.72	38.44	-0.13	47.41	48.44	47.99	+0.58
ExPeL	40.55	38.04	38.44	-2.11	36.59	39.10	37.38	+0.79	48.04	48.62	48.27	+0.23
AWM	37.25	38.04	38.04	+0.79	36.86	36.86	37.91	+1.06	48.10	47.47	47.70	-0.40
Evolver	38.04	39.37	39.89	**+1.85**	38.31	38.97	38.71	+0.40	48.97	49.83	49.94	+0.98
DC	36.72	37.65	37.25	+0.53	35.80	36.86	36.33	+0.53	45.86	47.59	48.74	**+2.87**
GSEM	38.84	37.38	36.99	-1.85	36.99	36.33	38.31	+1.32	48.56	48.51	47.70	-0.86
SkeMex	39.10	39.10	39.63	+0.53	35.93	37.78	36.20	+0.26	48.79	48.62	49.08	+0.29
3DMedEvo	**42.01**	**42.27**	**42.93**	+0.92	**51.92**	**51.65**	**53.63**	**+1.72**	**55.92**	**56.38**	**56.95**	+1.03

Table 3: **Ablation study on memory content.** VR, MR, Rec, and Ovr. denote visual reasoning, medical reasoning, recognition, and overall accuracy.

Variant	VR	MR	Rec	Ovr.	Δ
3DMedEvo	**49.20**	**53.56**	**68.10**	**53.63**	–
Memory-free ReAct	31.28	44.57	38.79	37.12	−16.51
Raw trajectory	33.16	47.19	43.97	39.76	−13.87
Partial skills	45.72	49.81	65.52	50.20	−3.43
Text-only skills	43.32	54.68	54.31	49.01	−4.62

Table 4: **Ablation study on branch partition.** Obs, Int, Dec: Observe, Interpret, Decide. Part.: branch-specific retrieval quotas, deduplication, and governance. Checkmarks indicate enabled components.

Variant	Obs	Int	Dec	Part.	VR	MR	Rec	Ovr.	Δ
3DMedEvo	✓	✓	✓	✓	**49.20**	**53.56**	68.10	**53.63**	–
Flat repository	✓	✓	✓		48.13	49.81	54.31	49.67	−3.96
Observe only	✓				48.66	47.19	56.03	49.27	−4.36
Interpret only		✓			45.72	52.43	67.24	51.39	−2.24
Decide only			✓		39.30	48.69	**69.83**	47.29	−6.34

using the same backbone. On in-domain DeepTumorVQA, 3DMedEvo exceeds the strongest memory baselines by 4.31 and 4.25 for GLM-5.2 and Qwen3.6-Plus, respectively. The task breakdown shows gains in measurement, visual reasoning, and medical reasoning with both backbones. We then evaluate the same frozen library on out-of-domain DeepChestVQA, where the gains over the strongest same-backbone memory baselines are 5.28 and 6.87, respectively. 3DMedEvo also leads the same-backbone comparisons in recognition and visual reasoning. By contrast, some memory baselines underperform ReAct. Taken together, these results suggest that CT-reading experience captured in skills can generalize across cases and potentially even across disease types, without test-time updates. Detailed subtype results are provided in Appendix B.1.

Evolving Memory. We further evaluate Evolving Memory, where skills are continually accumulated and updated during online testing. We use three backbones on two CT benchmarks (Table 2). Three rounds revisit the same questions, but each R1 question is answered before any feedback on it enters the bank. Even in R1, 3DMedEvo leads the strongest baseline by 1.46, 13.35, and 6.95, and it ranks first in every round with a net gain from R1 to R3 in all three settings. Final accuracies reach 42.93% and 53.63% on DeepChestVQA and 56.95% on DeepTumorVQA. These scores exceed Evolver, the strongest final-round baseline, by 3.04, 14.92, and 7.01, respectively. By contrast, ExPeL and GSEM lose 2.11 and 1.85 from R1 to R3 in the 27B DeepChestVQA setting. AWM and GSEM also decline on DeepTumorVQA, by 0.40 and 0.86. These contrasting trends suggest that selective skill integration, credit-guided retrieval, and uncertainty-aware governance can help retain useful CT-reading procedures while limiting harmful memory effects. See Appendix B.2 for details.

4.3 Ablations and Analysis

Quality Control in Skill Synthesis. Figure 3a reports ablations of Admission, Case Review, clinical grounding, and peer review. All ablations use DeepChestVQA under Evolving Memory. Passing every trajectory to synthesis without Admission produces the largest overall loss. Removing Case Review or clinical grounding reduces overall accuracy by 5.41 and 5.94, respectively, more than omitting post-encoding peer review. Strict grounding requires explicit clinical support: it outperforms no grounding but underperforms rejecting only contradicted proposals overall.

Component Ablation Analysis. Figure 3b shows that relevance-only retrieval hurts both reasoning categories but slightly improves recognition, whereas crediting all injected rather than only adopted

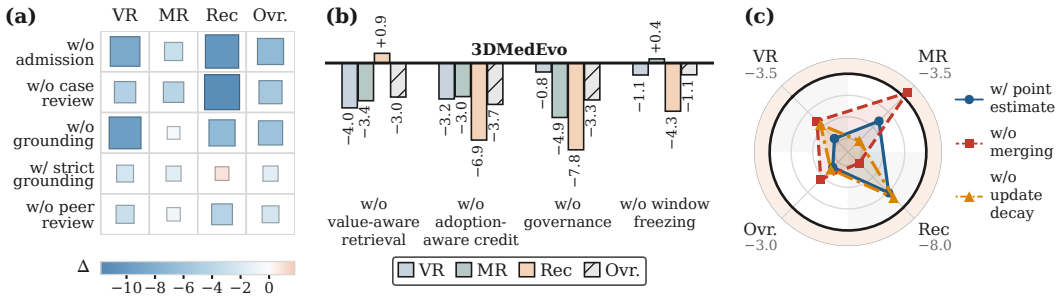


Figure 3: **Ablation study on the learning chain (a), the window cycle (b), and governance rules (c).** All values are gaps from 3DMedEvo, the zero line in (b) and the bold ring in (c).

Table 5: **Analysis of source quality.**

Source reading	**Correct**	**Failed**
Ever retrieved	24.3	**60.2**
Low SCS	50.5	50.0
Mid SCS	56.5	57.0
High SCS	65.7	54.2

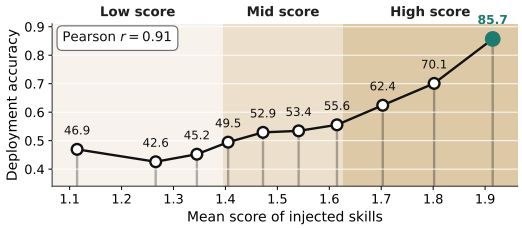


Figure 4: **Analysis of retrieval score calibration.**

skills causes the largest overall loss of 3.7. The latter can credit skills that never influenced the answer. Without governance, near-duplicates may crowd out complementary guidance. Without window freezing, we make new skills retrievable after each case rather than freezing the bank across the batch, potentially displacing better-supported guidance.

Impact of Governance Design. Figure 3c shows that replacing credit intervals with direct SCS comparisons against the branch reference reduces accuracy: equal scores can reflect unequal adoption evidence. Setting $\gamma = 1$ keeps historical credit and advantage unchanged after revision but underperforms discounting, as old outcomes may not describe revised skills. Disabling within-branch merging slightly improves medical reasoning but substantially harms recognition.

Impact of Memory Design. Tables 3 and 4 compare memory content and branch organization. Raw trajectories add only 2.64 over memory-free ReAct, suggesting redundancy limits reuse. Freezing skills after the first 25% of evaluation questions lowers overall accuracy by 3.43 versus continued evolution. Text-only skills remove exemplars from storage and injection, not current CT input: medical reasoning slightly improves, but recognition suffers most. At fixed total capacity, removing branch partitioning while retaining all three skill types hurts recognition most. Decide-only leads recognition, but Interpret, the strongest single branch overall, trails the full system by 2.24, supporting branch complementarity.

Skill Reliability Analysis. We examine whether the agent answers more accurately when injected skills have higher retrieval scores. Figure 4 bins deployment cases by the mean score of injected skills. Accuracy rises from 46.9% in the lowest bin to 85.7% in the highest (Pearson r=0.91), although query difficulty may contribute to this trend. We further compare skill sources in Table 5: 60.2% of failure-derived skills are retrieved at least once, versus 24.3% of correct-source skills. Accuracy differs by less than 1 across sources at low and mid SCS, but correct-source skills lead by 11.5 at high SCS, showing stronger transfer. Details of physician verification are in Appendix B.7.

5 Conclusion

In this paper, we introduced 3DMedEvo, a self-evolving framework that distills CT-reading experience into evidence-grounded multimodal skills without updating VLM parameters. Each skill pairs procedures with visual exemplars from prior cases. By coupling adoption-aware skill credit with posterior-interval governance, it supports selective experience reuse and continual refinement. Experiments on thoracic and abdominal CT benchmarks demonstrate improvements over tool-using and memory-based agents across VLM backbones. We believe 3DMedEvo opens a path toward medical agents that improve by integrating clinical knowledge with accumulated experience.

AI USE STATEMENT

We used LLMs (*e.g.*, GPT) to aid in polishing the writing of this paper. Specifically, LLMs were employed to refine grammar, improve fluency, and adjust the tone of academic writing (*e.g.*, smoothing transitions between sections, shortening overly long sentences, and rephrasing repetitive expressions). The models were not used for generating technical content, conducting experiments, or formulating research ideas; all substantive contributions remain the work of the authors.

ETHICS STATEMENT

This study uses only existing, publicly available CT data from DeepChestVQA (Wang et al., 2026a) and DeepTumorVQA (Chen et al., 2025; 2026) for secondary academic research, without new patient-data collection. The source datasets of our DeepTumorVQA training pool are listed in Table 18. As downstream academic users, we respect the original licenses, copyright constraints, research-use conditions, and privacy requirements. Our results do not establish clinical safety or effectiveness. Any future clinical deployment would require prospective validation, appropriate governance review, and oversight by qualified clinicians.

REPRODUCIBILITY STATEMENT

To support reproducibility, we describe the benchmark task definitions and training-data curation in Appendix H, the tool operations in Appendix D, the answer normalization and metric computation in Appendix F, and the implementation settings and hyperparameters in Appendix G. The exact prompts used for task execution and memory evolution are also provided in Appendix J. Together, these sections specify the data-preparation procedure, tool behavior, scoring rules, model settings, and prompt interfaces used in our evaluation. We will open-source the implementation, prompts, and scripts needed to reproduce the reported experiments, subject to the licenses and access conditions of the underlying datasets and models.

REFERENCES

Alibaba Cloud. Embedding. Alibaba Cloud Model Studio documentation, 2026a. URL https://www.alibabacloud.com/help/en/model-studio/embedding. Documentation for text-embedding-v4; updated September 4, 2026.

Alibaba Cloud. qwen3.6-plus model info. Alibaba Cloud Model Studio documentation, 2026b. URL https://www.alibabacloud.com/help/en/model-studio/qwen3-6-plus. Updated September 11, 2026.

Pablo Alonso-Coello, Holger J Schünemann, Jenny Moberg, Romina Brignardello-Petersen, Elie A Akl, Marina Davoli, Shaun Treweek, Reem A Mustafa, Gabriel Rada, Sarah Rosenbaum, et al. Grade evidence to decision (etd) frameworks: a systematic and transparent approach to making well informed healthcare choices. 1: Introduction. *bmj*, 353, 2016.

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022.

Mohammed Baharoon, Luyang Luo, Michael Moritz, Abhinav Kumar, Sung Eun Kim, Xiaoman Zhang, Miao Zhu, Mahmoud H Alabbad, Maha S Alhazmi, Neel P Mistry, et al. Rexgroundingct: A 3d chest ct dataset for segmentation of findings from free-text reports. *NEJM AI*, 3(7): AIdbp2501220, 2026.

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

Howard S Barrows and Paul J Feltovich. The clinical reasoning process. *Medical education*, 21(2): 86–91, 1987.

Raymond Bertram, Johanna Kaakinen, Frank Bensch, Laura Helle, Eila Lantto, Pekka Niemi, and Nina Lundbom. Eye movements of radiologists reflect expertise in ct study interpretation: a potential tool to measure resident development. *Radiology*, 281(3):805–815, 2016.

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

Melissa C Brouwers, Michelle E Kho, George P Browman, Jako S Burgers, Francoise Cluzeau, Gene Feder, Béatrice Fervers, Ian D Graham, Jeremy Grimshaw, Steven E Hanna, et al. Agree ii: advancing guideline development, reporting and evaluation in health care. *Canadian Medical Association Journal*, 182(18):E839–E842, 2010.

Kai Cao, Yingda Xia, Jiawen Yao, Xu Han, Lukas Lambert, Tingting Zhang, Wei Tang, Gang Jin, Hui Jiang, Xu Fang, et al. Large-scale pancreatic cancer detection via non-contrast ct and deep learning. *Nature medicine*, 29(12):3033–3043, 2023.

Bernard Charlin, Henny PA Boshuizen, Eugene J Custers, and Paul J Feltovich. Scripts and clinical reasoning. *Medical education*, 41(12):1178–1184, 2007.

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

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

SE Elstein Arthur and A Schwarz. Clinical problem solving and diagnostic decision ma ing: selective review of the cognitive literature. *BMJ*, 324(23):729–732, 2002.

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

Lin Fan, Pengyu Dai, Zhipeng Deng, Haolin Wang, Xun Gong, Yefeng Zheng, and Yafei Ou. Evolving medical imaging agents via experience-driven self-skill discovery. *arXiv preprint arXiv:2603.05860*, 2026.

Runnan Fang, Yuan Liang, Xiaobin Wang, Jialong Wu, Shuofei Qiao, Pengjun Xie, Fei Huang, Huajun Chen, and Ningyu Zhang. Memp: Exploring agent procedural memory. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 17490–17502, 2026.

Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*. chapman and hall/cRc, 1995.

Google DeepMind. 2.5 Flash and native capabilities—audio & image model card, December 2025. URL `https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-2-5-Flash-Model-Card.pdf`.

Google DeepMind. Gemini 3.8 Flash Model Card, September 2026. URL `https://deepmind.google/models/model-cards/gemini-3-8-flash/`.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yujiu Yang, Nan Duan, Weizhu Chen, et al. Critic: Large language models can self-correct with tool-interactive critiquing. In *International Conference on Learning Representations*, volume 2024, pp. 57734–57811, 2024.

Siyuan Guo, Cheng Deng, Ying Wen, Hechang Chen, Yi Chang, and Jun Wang. Ds-agent: Automated data science by empowering large language models with case-based reasoning. *arXiv preprint arXiv:2402.17453*, 2024.

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

Xiao Han, Yuzheng Fan, Sendong Zhao, Haochun Wang, and Bing Qin. Gsem: Graph-based self-evolving memory for experience augmented clinical reasoning. *arXiv preprint arXiv:2603.22096*, 2026.

Nicholas Heller, Fabian Isensee, Klaus H Maier-Hein, Xiaoshuai Hou, Chunmei Xie, Fengyi Li, Yang Nan, Guangrui Mu, Zhiyong Lin, Miofei Han, et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. *Medical image analysis*, 67:101821, 2021.

Nicholas Heller, Fabian Isensee, Dasha Trofimova, Resha Tejpaul, Zhongchen Zhao, Huai Chen, Lisheng Wang, Alex Golts, Daniel Khapun, Daniel Shats, et al. The kits21 challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase ct. *arXiv preprint arXiv:2307.01984*, 2023.

Yutao Hu, Ying Zheng, Shumei Miao, Xiaolei Zhang, Jiahao Xia, Yaolei Qi, Yiyang Zhang, Yuting He, Qian Chen, Jing Ye, et al. Cardiac-clip: A vision-language foundation model for 3d cardiac ct images. *arXiv preprint arXiv:2507.22024*, 2025.

Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in neural information processing systems*, 35:36722–36732, 2022.

Guanyu Jiang, Zhaochen Su, Xiaoye Qu, and Yi R Fung. Xskill: Continual learning from experience and skills in multimodal agents. *arXiv preprint arXiv:2603.12056*, 2026.

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

Yankai Jiang, Yujie Zhang, Peng Zhang, Yichen Li, Jintai Chen, Xiaoming Shi, and Shihui Zhen. Incentivizing tool-augmented thinking with images for medical image analysis. *arXiv preprint arXiv:2512.14157*, 2025b.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik S Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae W Park. Mdagents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems*, 37:79410–79452, 2024.

Janet L Kolodner. An introduction to case-based reasoning. *Artificial intelligence review*, 6(1):3–34, 1992.

Binxu Li, Tiankai Yan, Yuanting Pan, Jie Luo, Ruiyang Ji, Jiayuan Ding, Zhe Xu, Shilong Liu, Haoyu Dong, Zihao Lin, et al. Mmedagent: Learning to use medical tools with multi-modal agent. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 8745–8760, 2024a.

Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, et al. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*, 2024b.

Sijing Li, Zhongwei Qiu, Jiang Liu, Wenqiao Zhang, Tianwei Lin, Yihan Xie, Jianxiang An, Boxiang Yun, Chenglin Yang, Jun Xiao, et al. Tumorchain: Interleaved multimodal chain-of-thought reasoning for traceable clinical tumor analysis. In *International Conference on Learning Representations*, volume 2026, pp. 119790–119819, 2026a.

Wenjie Li, Yujie Zhang, Haoran Sun, Yueqi Li, Fanrui Zhang, Mengzhe Xu, Victoria Borja Clausich, Sade Mellin, Renhao Yang, Chenrun Wang, et al. Cx-mind: a pioneering multimodal large language model for interleaved reasoning in chest x-ray via curriculum-guided reinforcement learning. *Information Fusion*, pp. 104027, 2025.

Wenjie Li, Yujie Zhang, Haoran Sun, Xingqi He, Hongcheng Gao, Chenglong Ma, Ming Hu, Guankun Wang, Shiyi Yao, Renhao Yang, et al. Medscope: Incentivizing" think with videos" for clinical reasoning via coarse-to-fine tool calling. In *Forty-third International Conference on Machine Learning*, 2026b.

Xujun Li, Kehan Zheng, Mingyuan Zhao, Yize Geng, Jinfeng Zhou, Qi Zhu, Fei Mi, Lifeng Shang, Minlie Huang, and Hongning Wang. You live more than once: Towards hierarchical skill meta-evolving. *arXiv preprint arXiv:2605.28390*, 2026c.

Jingyang Lin, Yingda Xia, Jianpeng Zhang, Ke Yan, Kai Cao, Le Lu, Jiebo Luo, and Ling Zhang. Ct-glip: 3d grounded language-image pretraining with ct scans and radiology reports for full-body scenarios. *arXiv preprint arXiv:2404.15272*, 2024.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173, 2024.

Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyuan Wang, Xin Liu, et al. Abdomenct-1k: Is abdominal organ segmentation a solved problem? *IEEE transactions on pattern analysis and machine intelligence*, 44(10):6695–6714, 2021.

Sílvia Mamede, Tamara Van Gog, Alexandre Moura Sampaio, Rosa Malena Delbone De Faria, José Peixoto Maria, and Henk G Schmidt. How can students' diagnostic competence benefit most from practice with clinical cases? the effects of structured reflection on future diagnosis of the same and novel diseases. *Academic Medicine*, 89(1):121–127, 2014.

Jingwei Ni, Yihao Liu, Xinpeng Liu, Yutao Sun, Mengyu Zhou, Pengyu Cheng, Dexin Wang, Erchao Zhao, Xiaoxi Jiang, and Guanjun Jiang. Trace2skill: Distill trajectory-local lessons into transferable agent skills. *arXiv preprint arXiv:2603.25158*, 2026.

Geoff Norman, Meredith Young, and Lee Brooks. Non-analytical models of clinical reasoning: the role of experience. *Medical education*, 41(12):1140–1145, 2007.

Geoffrey Norman. Research in clinical reasoning: past history and current trends. *Medical education*, 39(4):418–427, 2005.

SH Patel, CL Stanton, SG Miller, JT Patrie, JN Itri, and TM Shepherd. Risk factors for perceptual-versus-interpretative errors in diagnostic neuroradiology. *American Journal of Neuroradiology*, 40(8):1252–1256, 2019.

Sohil H Patel and Jason N Itri. The role of intuitive cognition in radiologic decision making. *Journal of the American College of Radiology*, 19(5):669–676, 2022.

Qwen Team. Qwen3.5-27B. Hugging Face model card, 2026. URL `https://huggingface.co/Qwen/Qwen3.5-27B`.

Qwen Team. Qwen3-VL-235B-A22B-Instruct. Hugging Face model card, n.d. URL `https://huggingface.co/Qwen/Qwen3-VL-235B-A22B-Instruct`.

Geoffrey D Rubin, Justus E Roos, Martin Tall, Brian Harrawood, Sukantadev Bag, Donald L Ly, Danielle M Seaman, Lynne M Hurwitz, Sandy Napel, and Kingshuk Roy Choudhury. Characterizing search, recognition, and decision in the detection of lung nodules on ct scans: elucidation with eye tracking. *Radiology*, 274(1):276–286, 2015.

Henk G Schmidt and Remy MJP Rikers. How expertise develops in medicine: knowledge encapsulation and illness script formation. *Medical education*, 41(12):1133–1139, 2007.

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

Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Jieyu Zhang, Hang Wu, Yuanda Zhu, Joyce C Ho, Carl Yang, and May Dongmei Wang. Ehragent: Code empowers large language models for few-shot complex tabular reasoning on electronic health records. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22315–22339, 2024.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

Haoran Sun, Wenjie Li, Yujie Zhang, Zekai Lin, Fanrui Zhang, Kaitao Chen, Xingqi He, Yichen Li, Mianxin Liu, Lei Liu, et al. Experience makes skillful: Enabling generalizable medical agent reasoning via self-evolving skill memory. *arXiv preprint arXiv:2606.09365*, 2026.

Mirac Suzgun, Mert Yuksekgonul, Federico Bianchi, Dan Jurafsky, and James Zou. Dynamic cheatsheet: Test-time learning with adaptive memory. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7080–7106, 2026.

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

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

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

Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. Agent workflow memory. *arXiv preprint arXiv:2409.07429*, 2024.

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

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

Rong Wu, Xiaoman Wang, Jianbiao Mei, Pinlong Cai, Daocheng Fu, Cheng Yang, Licheng Wen, Xuemeng Yang, Yufan Shen, Yuxin Wang, et al. Evolver: Self-evolving llm agents through an experience-driven lifecycle. *arXiv preprint arXiv:2510.16079*, 2025b.

Peng Xia, Jianwen Chen, Hanyang Wang, Jiaqi Liu, Kaide Zeng, Yu Wang, Siwei Han, Yiyang Zhou, Xujiang Zhao, Haifeng Chen, et al. Skillrl: Evolving agents via recursive skill-augmented reinforcement learning. *arXiv preprint arXiv:2602.08234*, 2026.

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

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for llm agents. *Advances in Neural Information Processing Systems*, 38:17577–17604, 2026b.

Yifan Yang, Ziyang Gong, Weiquan Huang, Qihao Yang, Ziwei Zhou, Zisu Huang, Yan Li, Xuemei Gao, Qi Dai, Bei Liu, et al. Skillopt: Executive strategy for self-evolving agent skills. *arXiv preprint arXiv:2605.23904*, 2026.

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

Z.ai. GLM-5.2. Hugging Face model card, n.d. URL `https://huggingface.co/zai-org/GLM-5.2`.

Qizheng Zhang, Changran Hu, Shubhangi Upasani, Boyuan Ma, Fenglu Hong, Vamsidhar Kamanuru, Jay Rainton, Chen Wu, Mengmeng Ji, Hanchen Li, et al. Agentic context engineering: Evolving contexts for self-improving language models. In *International Conference on Learning Representations*, volume 2026, pp. 86069–86100, 2026a.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *Nejm Ai*, 2(1):AIoa2400640, 2025.

Shengtao Zhang, Jiaqian Wang, Ruiwen Zhou, Junwei Liao, Yuchen Feng, Zhuo Li, Yujie Zheng, Weinan Zhang, Ying Wen, Zhiyu Li, et al. Memrl: Self-evolving agents via runtime reinforcement learning on episodic memory. *arXiv preprint arXiv:2601.03192*, 2026b.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. Expel: Llm agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19632–19642, 2024.

Boyuan Zheng, Michael Y Fatemi, Xiaolong Jin, Zora Zhiruo Wang, Apurva Gandhi, Yueqi Song, Yu Gu, Jayanth Srinivasa, Gaowen Liu, Graham Neubig, et al. Skillweaver: Web agents can self-improve by discovering and honing skills. *arXiv preprint arXiv:2504.07079*, 2025.

Huichi Zhou, Yihang Chen, Siyuan Guo, Xue Yan, Kin Hei Lee, Zihan Wang, Ka Yiu Lee, Guchun Zhang, Kun Shao, Linyi Yang, et al. Memento: Fine-tuning llm agents without fine-tuning llms. *arXiv preprint arXiv:2508.16153*, 2025.

Yinghao Zhu, Yifan Qi, Zixiang Wang, Lei Gu, Dehao Sui, Haoran Hu, Xichen Zhang, Ziyi He, Junjun He, Liantao Ma, et al. Healthflow: A self-evolving ai agent with meta planning for autonomous healthcare research. *arXiv preprint arXiv:2508.02621*, 2025.

APPENDIX CONTENTS

A Agent Runtime and Interaction Protocol **18**

A.1 Runtime Construction 18

A.2 Context Management 18

B Additional Evaluation Results **18**

B.1 Frozen Memory 19

B.2 Evolving Memory 24

B.3 Cross-Backbone Generalisation 27

B.4 Repository Transfer Across Backbones 28

B.5 Effect of Data Order 28

B.6 Hyperparameter Sensitivity 30

B.7 Physician Verification 31

C Method Details **31**

C.1 Multi-Criteria Skill Retrieval 31

C.2 Skill Encoding 32

C.3 Adoption-Aware Skill Credit 32

C.4 Posterior-Interval Governance 33

C.5 Pseudocode 34

D Tool Suite **34**

E Baselines **36**

E.1 Medical Vision–Language Models 36

E.2 Tool-Assisted Agents 36

E.3 Reflection and Self-Correction 37

E.4 Experience-Based Memory Agents 37

F Metrics **37**

F.1 Accuracy 38

F.2 Tool Use, Memory, and Efficiency 38

G Implementation Details **38**

H Dataset Details **40**

H.1 Details of DeepChestVQA and DeepTumorVQA 40

H.2 Training-Set Curation 43

H.3 Subset Statistics 44

I Qualitative Results **44**

I.1 Successful Cases 45
I.2 Failure Cases 45

J Prompts **61**
J.1 Task Execution Prompts 61
J.2 Memory Evolution Prompts 61
J.3 Evaluation Prompts 61

K Limitations **70**

L Future Work **71**

A Agent Runtime and Interaction Protocol

In this section, we describe the construction of the 3DMedEvo runtime and its turn-by-turn context management. A frozen vision–language model serves as the ReAct controller (Yao et al., 2022), reasoning over the available evidence and selecting either a tool call or a final answer. The runtime assembles model requests, executes the selected tools, and feeds their observations back to the controller.

A.1 Runtime Construction

Each case starts with the clinical question, the current CT examination, an empty action–observation history, and the dataset-specific tools that can inspect that examination. The opening request includes the question and any initial media supplied by the dataset interface. Subsequent tool calls provide additional views or textual findings, according to the capabilities of that interface. Before execution, the runtime also receives the retrieved skill text and available exemplar descriptions and identifiers. These remain fixed for the case and enter the interaction through the context described below. The skill-selection procedure is given in Section 3.3.

The controller emits one non-empty `<reasoning>` block followed by exactly one `<tool>` or `<response>` block, without surrounding free text. The reasoning block expresses the controller's assessment of the available evidence and its rationale for the next tool call or final answer. When a retrieved skill informs this decision, the controller also cites its identifier in this block and briefly states which guidance it applied. A tool block contains one JSON object with top-level `tool` and `arguments` keys. The runtime executes the specified tool with those arguments. The reasoning, tool call, and returned observation constitute one logical turn. The runtime records this turn before constructing the next request, allowing the controller to choose its next action using the updated history. A `<response>` block contains the answer and ends the interaction directly.

If the tool-use budget is exhausted before a response is produced, the runtime issues one additional final request with tools disabled. It includes the last returned textual observation and any image provided for that turn, and asks the controller to answer using the accumulated case evidence. This final-answer request is additional to the configured number of tool-enabled turns.

A.2 Context Management

At every decision step, the runtime rebuilds the user memory, namely the case-specific context presented in the user message. This message contains the accumulated textual interaction history, the original clinical question, a separate block of retrieved skill guidance with available exemplar descriptions and identifiers, and a reminder of the required output format. Images provided for the current turn are attached to the same message. On the extra final-answer turn, an instruction to answer without further tool calls is also included. The accompanying system message specifies the interaction protocol and the available tools.

After each tool call, the completed reasoning–action–observation turn is appended to the history used for the next reconstruction. Previous turns are presented in chronological order, with tool observations expressed as the text or summaries returned by the dataset interface. The question and selected skill guidance remain fixed; the latter is reinserted as a separate block, not appended repeatedly to the history.

Visual evidence follows the same turn-by-turn progression. When an initial overview is provided, it accompanies the opening request. An image returned by a tool accompanies the following request, while its textual observation stays in the history.

B Additional Evaluation Results

This section provides complementary evidence for the main experimental findings. It first presents supplementary results for frozen and evolving memories, then examines cross-backbone evaluation protocols and repository transfer, sensitivity to hyperparameters and data order, and complete case studies.

B.1 FROZEN MEMORY

This subsection supplements the main frozen-memory results with detailed comparisons of 3DMedEvo and the memory baselines using GLM-5.2 and Qwen3.6-Plus as backbones. We examine tool use during offline construction and evaluation, construction costs in LLM calls and tokens, tool-specific usage patterns, and question-subtype accuracy on DeepChestVQA and DeepTumorVQA.

Table 6: **Tool use during frozen-memory construction and evaluation.** Training tool calls and Calls/q denote the total number of tool actions during offline memory construction and the mean number per evaluation question, respectively.

Backbone	Method	Training tool calls	DeepChestVQA		DeepTumorVQA	
			Acc.	Calls/q	Acc.	Calls/q
Memory-free ReAct						
GLM-5.2	ReAct	–	37.12	2.28	46.03	3.18
Qwen3.6-Plus	ReAct	–	40.29	2.52	46.32	3.24
GPT-5.5	ReAct	–	41.22	2.41	47.76	3.29
MedGemma-27B	ReAct	–	38.31	0.37	40.98	2.09
Lingshu-32B	ReAct	–	36.86	1.00	38.51	1.01
HuluMed-32B	ReAct	–	38.57	0.26	40.92	1.39
RadFM-14B	ReAct	–	32.36	0.20	36.09	1.40
Frozen-memory methods						
GLM-5.2	Reflexion	20,454	40.69	2.78	47.07	5.59
	CRITIC	20,368	39.10	2.85	48.85	5.35
	ExPeL	19,079	37.78	2.91	48.22	5.78
	AWM	19,890	39.63	2.52	45.63	5.10
	Evolver	21,594	40.69	2.72	47.82	8.14
	DC	18,884	39.89	2.75	45.98	4.93
	GSEM	19,432	41.35	2.90	48.16	4.98
	SkeMex	21,007	38.71	2.77	49.37	5.30
	3DMedEvo	13,802	46.63	1.76	53.68	3.89
Qwen3.6-Plus	Reflexion	14,128	39.10	1.62	48.62	3.40
	CRITIC	13,119	40.95	1.59	46.67	3.14
	ExPeL	12,800	37.52	1.52	47.41	3.19
	AWM	14,881	41.48	1.54	50.46	3.11
	Evolver	15,216	40.16	1.58	48.51	3.47
	DC	12,240	35.54	1.57	44.94	2.85
	GSEM	12,756	39.37	1.64	48.62	3.06
	SkeMex	12,964	37.12	1.60	49.71	3.16
	3DMedEvo	14,721	48.35	2.03	54.71	3.59

In Table 6, we compare the tool interaction required for frozen-memory construction and evaluation. With GLM-5.2, 3DMedEvo achieves the highest accuracy on both benchmarks while using the fewest tool actions among same-backbone memory methods at both stages. It requires 13,802 construction actions and averages 1.76 and 3.89 actions per evaluation question on DeepChestVQA and DeepTumorVQA, respectively. Relative to the lowest corresponding baseline counts, these represent reductions of 26.9%, 30.2%, and 21.1%. The simultaneous improvement in accuracy and interaction efficiency supports the value of reusable procedures that guide evidence collection, rather than relying on more extensive tool exploration. The accuracy lead also holds for Qwen3.6-Plus on both benchmarks, extending the performance benefit across backbones while the GLM-5.2 results additionally demonstrate substantial interaction savings.

In Table 7, we compare LLM-call and token costs for frozen-memory construction. 3DMedEvo uses 88.39M and 69.40M tokens with GLM-5.2 and Qwen3.6-Plus, respectively, only 0.4% and

Table 7: **Frozen-memory construction costs.**

Method	GLM-5.2			Qwen3.6-Plus		
	LLM calls	Total tokens (M)	Mean tokens/q	LLM calls	Total tokens (M)	Mean tokens/q
Reflexion	26,645	88.04	20,398.4	28,358	68.81	15,941.9
CRITIC	26,561	87.96	20,380.8	27,398	66.27	15,354.5
ExPeL	22,085	72.93	16,898.5	22,822	58.88	13,642.8
AWM	20,964	41.33	9,574.9	23,724	52.35	12,129.5
Evolver	22,047	44.76	10,370.3	23,846	53.05	12,290.6
DC	21,451	68.23	15,809.1	22,213	53.80	12,464.2
GSEM	21,982	72.40	16,774.0	23,347	59.15	13,703.9
SkeMex	22,889	82.83	19,191.6	22,994	59.80	13,855.2
3DMedEvo	28,774	88.39	20,536	30,690	69.40	14,800

0.9% above Reflexion—the closest baseline in total token use—while achieving higher accuracy on both benchmarks. Its LLM-call counts are 28,774 and 30,690, representing increases of 8.0% and 8.2% over Reflexion. Thus, relative to Reflexion, the performance advantage is achieved with a nearly unchanged construction token budget, despite a larger total number of model requests.

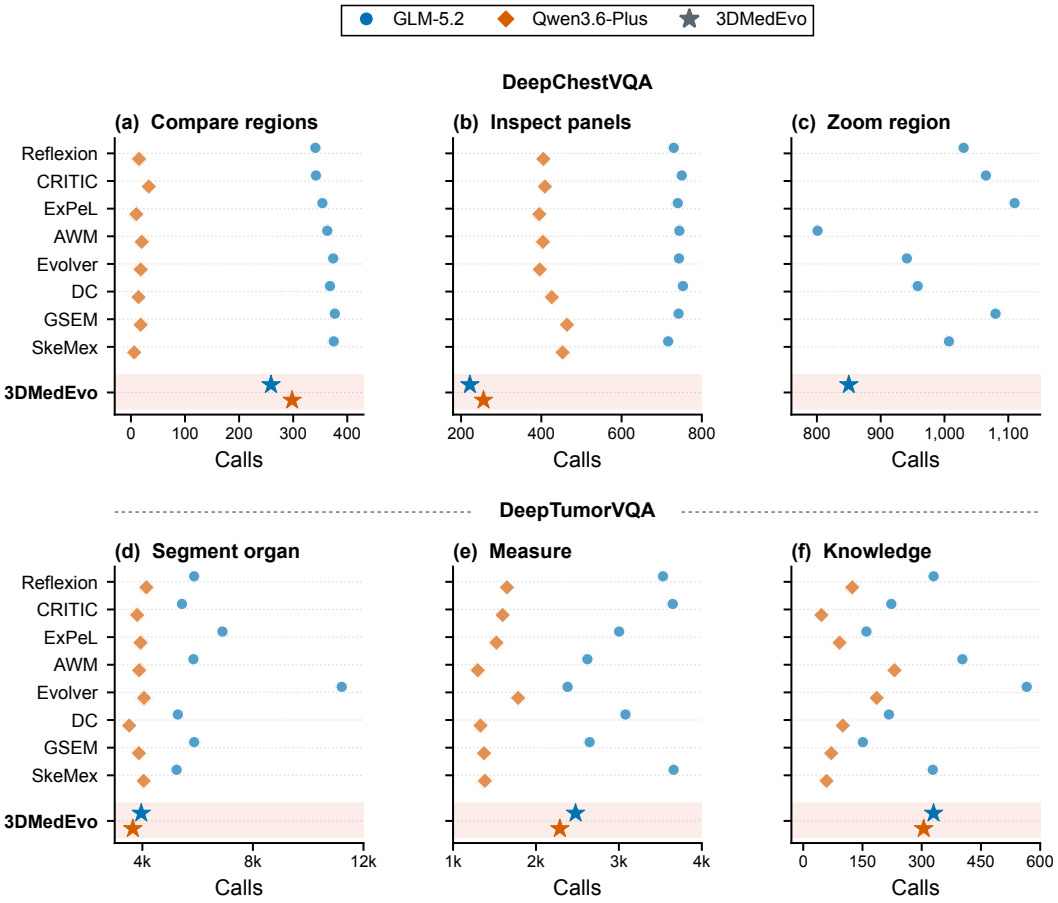


Figure 5: **Tool use during frozen-memory evaluation.** Panel (c) shows GLM-5.2 only.

In Figure 5, we compare tool selection during frozen-memory evaluation. On DeepChestVQA, 3DMedEvo uses 222 and 256 *Inspect panels* actions with GLM-5.2 and Qwen3.6-Plus, respectively,

reducing panel inspection by at least 69.0% and 35.2% relative to the corresponding memory baselines. This shared reduction is accompanied by 850 *Zoom region* calls with GLM-5.2 and 298 *Compare regions* calls with Qwen3.6-Plus, retaining an emphasis on localized visual assessment. On DeepTumorVQA, GLM-5.2 uses 3,959 *Segment organ* actions, at least 24.4% fewer than the memory baselines, whereas Qwen3.6-Plus makes greater use of *Measure* (2,288 calls) and *Knowledge* (305 calls). Together with the accuracy gains, these patterns are consistent with skill-guided tool selection: fewer panel-inspection calls across backbones, with continued use of localized visual assessment and quantitative verification.

Subtype-level accuracy. In Tables 8 and 9, we compare fine-grained accuracy across task families under the frozen-memory setting with GLM-5.2 and Qwen3.6-Plus. Across both backbones, 3DMedEvo shows aligned gains in recognition, clinical interpretation, and a selected quantitative task. For pleura lesion existence, it reaches 68.57% with GLM-5.2 and 71.43% with Qwen3.6-Plus, exceeding the strongest same-backbone memory baselines by 25.71 and 8.57 points, respectively. On DeepTumorVQA, lesion-type classification similarly reaches 73.91% with each backbone, improving over the corresponding baselines by 13.04 and 8.70 points. The pattern also extends to organ-volume measurement, where 58.23% and 54.43% exceed the best memory baselines by 6.33 and 5.06 points. These aligned cases span recognition, interpretation, and measurement, and are consistent with reusable procedures helping organize evidence selection and clinical interpretation. The gains are not uniform: organ-HU ratio reaches 57.45% and 51.06%, trailing the strongest same-backbone baselines by 4.26 and 8.51 points. This limitation suggests that procedural guidance alone cannot fully overcome precise numerical extraction or tool-execution bottlenecks. Taken together, the gains on matching subtypes support the value of learning reusable evidence-collection and interpretation procedures. Their recurrence across both backbones demonstrates the applicability of 3DMedEvo beyond a single model.

Table 8: **Subtype accuracy on DeepChestVQA.** Scores are percentages.

Subtype	N	ReAct	Reflexion	CRITIC	ExPeL	AWM	Evolver	DC	GSEM	SkeMex	3DMedEvo
GLM-5.2											
Recognition											
Bronchus/airway lesion existence	39	35.90	38.46	43.59	38.46	46.15	46.15	43.59	43.59	46.15	41.03
Lung lesion existence	42	45.24	45.24	38.10	40.48	45.24	45.24	38.10	47.62	47.62	66.67
Pleura lesion existence	35	34.29	42.86	34.29	37.14	34.29	37.14	37.14	37.14	37.14	68.57
Visual reasoning											
Largest lesion diameter	48	45.83	47.92	33.33	41.67	43.75	41.67	35.42	41.67	41.67	35.42
Largest lesion location	54	25.93	35.19	31.48	22.22	25.93	25.93	24.07	25.93	22.22	22.22
Largest lesion slice	50	28.00	34.00	38.00	40.00	24.00	40.00	38.00	34.00	34.00	42.00
Lesion counting	55	14.55	29.09	32.73	21.82	36.36	41.82	34.55	36.36	27.27	47.27
Lesion count by location	50	20.00	32.00	30.00	28.00	30.00	26.00	32.00	26.00	26.00	44.00
Organ enlargement	35	42.86	51.43	34.29	42.86	42.86	51.43	54.29	51.43	34.29	57.14
Organ atrophy	28	46.43	42.86	42.86	46.43	50.00	39.29	50.00	67.86	35.71	71.43
Lesion–organ HU difference	54	38.89	33.33	35.19	42.59	38.89	38.89	40.74	42.59	38.89	48.15
Medical reasoning											
Attenuation pattern classification	53	47.17	49.06	49.06	49.06	50.94	45.28	49.06	49.06	50.94	52.83
Volume-loss lesion classification	45	55.56	53.33	57.78	55.56	57.78	57.78	60.00	57.78	53.33	57.78
Imaging phenotype synthesis	54	27.78	33.33	31.48	18.52	27.78	27.78	24.07	38.89	33.33	31.48
Phenotype mixing identification	43	62.79	53.49	53.49	51.16	51.16	51.16	58.14	41.86	53.49	58.14
Emphysema severity grading	50	40.00	40.00	40.00	38.00	40.00	40.00	36.00	42.00	42.00	32.00
Pleural effusion grading	22	31.82	40.91	50.00	45.45	40.91	50.00	36.36	31.82	40.91	40.91

Continued on next page

Table 8 (continued)

Subtype	N	ReAct	Reflexion	CRITIC	ExPeL	AWM	Evolver	DC	GSEM	SkeMex	3DMedEvo
Qwen3.6-Plus											
Recognition											
Bronchus/airway lesion existence	39	41.03	41.03	43.59	41.03	43.59	41.03	41.03	41.03	41.03	56.41
Lung lesion existence	42	61.90	54.76	59.52	52.38	59.52	59.52	59.52	57.14	47.62	59.52
Pleura lesion existence	35	51.43	54.29	62.86	60.00	51.43	60.00	51.43	60.00	57.14	71.43
Visual reasoning											
Largest lesion diameter	48	35.42	27.08	25.00	29.17	37.50	37.50	22.92	27.08	25.00	41.67
Largest lesion location	54	29.63	33.33	25.93	29.63	29.63	29.63	33.33	29.63	29.63	35.19
Largest lesion slice	50	22.00	20.00	26.00	16.00	10.00	12.00	10.00	16.00	18.00	42.00
Lesion counting	55	27.27	27.27	34.55	16.36	41.82	30.91	14.55	30.91	32.73	32.73
Lesion count by location	50	16.00	32.00	32.00	40.00	30.00	34.00	34.00	40.00	36.00	38.00
Organ enlargement	35	42.86	42.86	62.86	42.86	45.71	54.29	22.86	31.43	25.71	54.29
Organ atrophy	28	35.71	53.57	53.57	39.29	57.14	32.14	42.86	46.43	35.71	75.00
Lesion–organ HU difference	54	40.74	44.44	35.19	31.48	42.59	42.59	31.48	40.74	37.04	38.89
Medical reasoning											
Attenuation pattern classification	53	54.72	45.28	50.94	56.60	56.60	43.40	54.72	49.06	50.94	43.40
Volume-loss lesion classification	45	62.22	55.56	57.78	60.00	62.22	64.44	62.22	62.22	55.56	66.67
Imaging phenotype synthesis	54	35.19	33.33	31.48	29.63	40.74	37.04	29.63	35.19	27.78	42.59
Phenotype mixing identification	43	51.16	53.49	51.16	51.16	37.21	44.19	48.84	51.16	46.51	58.14
Emphysema severity grading	50	44.00	24.00	32.00	22.00	36.00	34.00	30.00	28.00	34.00	46.00
Pleural effusion grading	22	50.00	45.45	36.36	40.91	36.36	40.91	22.73	36.36	40.91	54.55

Table 9: **Subtype accuracy on DeepTumorVQA.** Scores are percentages.

Subtype	N	Reflexion	CRITIC	ExPeL	AWM	Evolver	DC	GSEM	SkeMex	3DMedEvo
GLM-5.2										
Recognition										
PDAC existence	40	62.50	65.00	62.50	65.00	62.50	60.00	65.00	65.00	57.50
PNET existence	39	56.41	56.41	56.41	53.85	56.41	56.41	56.41	46.15	48.72
Colon lesion existence	39	51.28	46.15	53.85	56.41	53.85	53.85	51.28	66.67	51.28
Kidney cyst existence	39	61.54	64.10	66.67	64.10	64.10	64.10	64.10	66.67	56.41
Kidney lesion existence	39	51.28	53.85	56.41	51.28	53.85	48.72	48.72	61.54	46.15
Kidney tumor existence	39	46.15	46.15	48.72	48.72	48.72	48.72	48.72	48.72	58.97
Liver lesion existence	39	48.72	56.41	48.72	53.85	53.85	58.97	53.85	56.41	56.41
Pancreatic lesion existence	39	53.85	56.41	56.41	56.41	53.85	53.85	56.41	56.41	48.72
Splenomegaly detection	39	89.74	89.74	89.74	89.74	87.18	87.18	89.74	89.74	82.05
Measurement										
Lesion volume measurement	76	40.79	32.89	35.53	25.00	32.89	32.89	38.16	43.42	46.05
Organ HU measurement	75	64.00	64.00	64.00	65.33	65.33	65.33	68.00	66.67	72.00
Organ HU ratio	47	57.45	61.70	57.45	57.45	61.70	61.70	59.57	61.70	57.45
Organ volume measurement	79	48.10	49.37	46.84	51.90	50.63	43.04	48.10	46.84	58.23
Tumor burden percentage	75	28.00	36.00	30.67	28.00	24.00	24.00	34.67	30.67	34.67

Continued on next page

Table 9 (continued)

Subtype	N	Reflexion	CRITIC	ExPeL	AWM	Evolver	DC	GSEM	SkeMex	3DMedEvo
Visual reasoning										
Adjacent organ	39	35.90	41.03	53.85	46.15	35.90	41.03	48.72	46.15	51.28
Bilateral kidney asymmetry	45	15.56	20.00	22.22	4.44	6.67	11.11	11.11	15.56	35.56
Inter-segment comparison	44	20.45	25.00	31.82	34.09	20.45	29.55	25.00	29.55	22.73
Kidney volume comparison	45	71.11	64.44	60.00	51.11	51.11	66.67	60.00	55.56	75.56
Largest lesion attenuation	45	42.22	44.44	46.67	15.56	33.33	17.78	42.22	40.00	37.78
Largest lesion diameter	45	8.89	11.11	13.33	20.00	13.33	13.33	6.67	11.11	26.67
Largest lesion location	45	40.00	40.00	46.67	37.78	37.78	35.56	40.00	42.22	62.22
Largest lesion slice	44	31.82	45.45	27.27	29.55	38.64	36.36	36.36	38.64	38.64
Lesion count by location	44	52.27	59.09	50.00	59.09	61.36	40.91	63.64	36.36	47.73
Lesion counting	44	88.64	70.45	93.18	90.91	97.73	97.73	97.73	81.82	84.09
Lesion outlier	44	50.00	50.00	50.00	52.27	50.00	52.27	50.00	50.00	47.73
Liver lesion clustering	44	25.00	25.00	38.64	18.18	25.00	38.64	40.91	38.64	50.00
Multi-organ lesion burden comparison	44	22.73	43.18	20.45	9.09	22.73	22.73	9.09	27.27	50.00
Organ aggregation	45	53.33	55.56	57.78	57.78	51.11	55.56	51.11	55.56	57.78
Organ enlargement	44	61.36	61.36	61.36	61.36	63.64	61.36	61.36	61.36	65.91
Tumor–organ HU difference	44	29.55	27.27	27.27	18.18	27.27	25.00	31.82	22.73	47.73
Medical reasoning										
PDAC vs PNET classification	32	56.25	50.00	53.12	31.25	53.12	40.62	50.00	50.00	56.25
Fatty liver	24	58.33	54.17	41.67	58.33	62.50	41.67	66.67	50.00	95.83
Hepatic steatosis grading	31	74.19	61.29	67.74	70.97	70.97	64.52	80.65	67.74	64.52
Lesion type classification	23	39.13	60.87	17.39	21.74	26.09	17.39	8.70	17.39	73.91
Pancreatic cyst resectability	31	35.48	54.84	29.03	48.39	67.74	58.06	48.39	48.39	48.39
Pancreatic lesion resectability	31	45.16	38.71	29.03	16.13	19.35	25.81	6.45	87.10	67.74
Pancreatic steatosis	27	81.48	81.48	81.48	88.89	92.59	81.48	85.19	81.48	74.07
Pancreatic tumor staging	55	7.27	10.91	21.82	23.64	30.91	16.36	12.73	20.00	18.18
Portal hypertension assessment	16	87.50	93.75	93.75	87.50	87.50	87.50	87.50	93.75	93.75
Pseudocyst determination	3	33.33	33.33	33.33	33.33	33.33	33.33	33.33	33.33	66.67
Renal mass characterization	21	14.29	14.29	19.05	14.29	19.05	0.00	23.81	19.05	9.52
Splenomegaly grading	37	83.78	91.89	91.89	91.89	91.89	91.89	91.89	91.89	86.49
Qwen3.6-Plus										
Recognition										
PDAC existence	40	62.50	62.50	62.50	60.00	62.50	62.50	62.50	62.50	62.50
PNET existence	39	58.97	56.41	56.41	56.41	64.10	53.85	56.41	56.41	53.85
Colon lesion existence	39	51.28	46.15	46.15	48.72	46.15	41.03	61.54	56.41	58.97
Kidney cyst existence	39	58.97	46.15	61.54	69.23	64.10	64.10	64.10	64.10	58.97
Kidney lesion existence	39	53.85	53.85	56.41	48.72	53.85	58.97	43.59	61.54	51.28
Kidney tumor existence	39	61.54	51.28	51.28	51.28	58.97	48.72	48.72	58.97	64.10
Liver lesion existence	39	43.59	46.15	46.15	56.41	46.15	43.59	46.15	53.85	53.85
Pancreatic lesion existence	39	64.10	56.41	48.72	56.41	56.41	56.41	58.97	56.41	56.41
Splenomegaly detection	39	89.74	89.74	89.74	89.74	89.74	89.74	89.74	87.18	84.62
Measurement										
Lesion volume measurement	76	31.58	30.26	38.16	32.89	31.58	28.95	32.89	27.63	43.42
Organ HU measurement	75	66.67	65.33	65.33	66.67	69.33	66.67	68.00	65.33	70.67
Organ HU ratio	47	55.32	57.45	55.32	53.19	57.45	55.32	55.32	59.57	51.06
Organ volume measurement	79	48.10	48.10	46.84	48.10	46.84	46.84	49.37	48.10	54.43
Tumor burden percentage	75	32.00	26.67	32.00	28.00	24.00	26.67	34.67	30.67	36.00

Continued on next page

Table 9 (continued)

Subtype	N	Reflexion	CRITIC	ExPeL	AWM	Evolver	DC	GSEM	SkeMex	3DMedEvo
Visual reasoning										
Adjacent organ	39	51.28	53.85	53.85	46.15	41.03	53.85	46.15	48.72	53.85
Bilateral kidney asymmetry	45	11.11	15.56	11.11	17.78	22.22	15.56	20.00	35.56	37.78
Inter-segment comparison	44	25.00	22.73	27.27	31.82	20.45	13.64	31.82	25.00	27.27
Kidney volume comparison	45	60.00	57.78	60.00	60.00	60.00	60.00	66.67	60.00	75.56
Largest lesion attenuation	45	51.11	37.78	31.11	42.22	40.00	11.11	33.33	42.22	35.56
Largest lesion diameter	45	13.33	13.33	8.89	6.67	13.33	17.78	11.11	15.56	24.44
Largest lesion location	45	46.67	57.78	53.33	33.33	53.33	48.89	53.33	44.44	62.22
Largest lesion slice	44	36.36	34.09	22.73	25.00	27.27	22.73	22.73	27.27	36.36
Lesion count by location	44	22.73	29.55	25.00	59.09	22.73	29.55	22.73	52.27	40.91
Lesion counting	44	68.18	65.91	68.18	79.55	93.18	77.27	84.09	70.45	72.73
Lesion outlier	44	45.45	43.18	45.45	45.45	50.00	47.73	45.45	47.73	50.00
Liver lesion clustering	44	52.27	40.91	50.00	72.73	50.00	43.18	50.00	40.91	52.27
Multi-organ lesion burden comparison	44	36.36	18.18	45.45	54.55	36.36	15.91	43.18	40.91	52.27
Organ aggregation	45	46.67	53.33	53.33	53.33	51.11	51.11	53.33	51.11	60.00
Organ enlargement	44	75.00	77.27	75.00	72.73	77.27	72.73	77.27	68.18	70.45
Tumor–organ HU difference	44	34.09	34.09	38.64	36.36	43.18	47.73	34.09	36.36	45.45
Medical reasoning										
PDAC vs PNET classification	32	56.25	50.00	50.00	50.00	50.00	15.62	34.38	56.25	71.88
Fatty liver	24	95.83	41.67	70.83	95.83	87.50	95.83	87.50	91.67	95.83
Hepatic steatosis grading	31	35.48	70.97	41.94	32.26	38.71	32.26	22.58	38.71	67.74
Lesion type classification	23	52.17	39.13	13.04	34.78	13.04	8.70	65.22	21.74	73.91
Pancreatic cyst resectability	31	48.39	48.39	48.39	48.39	48.39	48.39	48.39	54.84	51.61
Pancreatic lesion resectability	31	51.61	61.29	48.39	67.74	35.48	45.16	38.71	61.29	54.84
Pancreatic steatosis	27	77.78	62.96	81.48	74.07	74.07	70.37	62.96	81.48	70.37
Pancreatic tumor staging	55	7.27	9.09	12.73	36.36	20.00	14.55	14.55	10.91	30.91
Portal hypertension assessment	16	87.50	81.25	87.50	87.50	81.25	87.50	93.75	87.50	93.75
Pseudocyst determination	3	33.33	33.33	66.67	0.00	33.33	33.33	66.67	66.67	66.67
Renal mass characterization	21	23.81	33.33	23.81	19.05	38.10	14.29	38.10	28.57	28.57
Splenomegaly grading	37	91.89	91.89	91.89	91.89	91.89	91.89	91.89	91.89	86.49

B.2 Evolving Memory

This subsection provides supplementary results for evolving-memory evaluation, covering question-type and subtype accuracy, repository growth, and computational cost. Throughout evaluation, ground-truth answers are used only by the evaluator to judge finalized responses and are withheld from the model contexts for task execution and skill evolution. The resulting task scores provide outcome feedback for trajectory review and skill refinement, while reference-answer text is excluded from skill-generation inputs.

DeepChestVQA question types. Table 10 shows that 3DMedEvo ranks first in all three question types with both backbones, although the gains differ markedly. With the smaller Qwen3.5-27B, improvements over the strongest baseline in each category range from 1.49 to 3.45 percentage points, whereas GLM-5.2 gains 26.72 points in recognition and 14.44 points in visual reasoning. This contrast is consistent with a visual-grounding bottleneck in the smaller model: even with evolved skills, Qwen3.5-27B reaches 33.96% on visual reasoning, compared with 49.20% for GLM-5.2. Our skills provide inspection procedures, verification rules, and visual exemplars that help structure CT reading, but the frozen backbone must still recognize the relevant findings and align the retrieved guidance with the current image. Consequently, skill reuse can support a smaller model without fully compensating for its perceptual limitations, while a backbone with stronger visual understanding may make more effective use of the same framework.

Table 10: **Round-3 accuracy (%) by question type on DeepChestVQA.** VR, MR, and Rec denote visual reasoning, medical reasoning, and recognition. **Bold** marks the best result per backbone and type.

	Qwen3.5-27B			GLM-5.2		
Method	**VR (374)**	**MR (267)**	**Rec (116)**	**VR (374)**	**MR (267)**	**Rec (116)**
Reflexion	27.54	45.69	48.28	27.54	44.19	41.38
CRITIC	30.75	44.94	47.41	31.02	47.57	41.38
ExPeL	29.14	46.44	50.00	32.35	42.70	41.38
AWM	27.81	46.82	50.86	31.28	45.69	41.38
Evolver	31.82	45.69	52.59	34.76	43.07	41.38
DC	28.34	45.32	47.41	30.21	42.70	41.38
GSEM	27.54	45.32	48.28	32.89	44.57	41.38
SkeMex	29.95	46.07	56.03	30.21	42.32	41.38
3DMedEvo	**33.96**	**48.31**	**59.48**	**49.20**	**53.56**	**68.10**

Fine-grained evolution on DeepTumorVQA. Table 11 reports the evolution of all 42 subtypes with Qwen3.6-Plus, revealing task-specific gains that are obscured by aggregate accuracy. From the first to the third round, 3DMedEvo improves PDAC versus PNET classification and pancreatic lesion resectability by 6.3 and 6.5 percentage points, respectively, while maintaining 88.6% accuracy on both lesion counting and organ enlargement throughout all three rounds. In contrast, some baselines regress on clinical decisions: Evolver loses 22.6 points on pancreatic lesion resectability, and GSEM loses 21.7 points on lesion type classification. The combination of targeted gains and retained strengths is consistent with the intended role of selective skill refinement: preserving useful inspection routines while improving reusable decision criteria, rather than assuming that accumulating more trajectories benefits every task. The gains are nevertheless uneven. Lesion volume measurement shows no net improvement from its first-round accuracy of 35.5%, and bilateral kidney asymmetry declines slightly. These persistent difficulties suggest that procedural memory remains dependent on the backbone and tools for accurate spatial and quantitative evidence; reusable guidance alone does not ensure that this evidence is correctly obtained and interpreted.

Table 11: Question-subtype accuracy under the evolving memory setting on DeepTumorVQA with Qwen3.6-Plus. R1 is accuracy (%); Δ_2/Δ_3 are R2−R1/R3−R1 (pp). Green/red indicate gains/drops. DC denotes DynamicCheatsheet.

	AWM			DC			Evolver			GSEM			SkeMex			3DMedEvo		
Subtype	R1	Δ_2	Δ_3	R1	Δ_2	Δ_3	R1	Δ_2	Δ_3	R1	Δ_2	Δ_3	R1	Δ_2	Δ_3	R1	Δ_2	Δ_3
Visual reasoning																		
adjacent organ	48.7	-2.6	+5.1	56.4	-2.6	-2.6	56.4	0.0	+7.7	48.7	-2.6	-5.1	46.2	+10.3	+7.7	56.4	+2.6	+5.1
bilateral kidney asymmetry	17.8	0.0	+2.2	15.6	0.0	0.0	20.0	0.0	0.0	15.6	0.0	0.0	33.3	-4.4	-2.2	24.4	-4.4	-2.2
inter-segment comparison	31.8	+2.3	+11.4	36.4	+6.8	+9.1	29.5	0.0	0.0	22.7	+4.5	0.0	22.7	-4.5	+11.4	25.0	+4.5	+2.3
kidney volume comparison	60.0	0.0	0.0	53.3	+4.4	+4.4	57.8	+2.2	0.0	62.2	-4.4	-2.2	57.8	0.0	+2.2	77.8	+2.2	-2.2
largest lesion attenuation	46.7	+4.4	+2.2	44.4	-4.4	-2.2	37.8	-2.2	+13.3	42.2	+15.6	+11.1	46.7	0.0	0.0	42.2	+2.2	0.0
largest lesion diameter	17.8	-2.2	-8.9	8.9	+13.3	+8.9	13.3	+2.2	-4.4	20.0	0.0	+8.9	13.3	+4.4	+2.2	15.6	-2.2	+4.4
largest lesion location	55.6	+2.2	+2.2	53.3	0.0	+4.4	60.0	-2.2	0.0	55.6	+4.4	0.0	55.6	-8.9	-2.2	66.7	+2.2	+4.4
largest lesion slice	40.9	-4.5	0.0	25.0	+13.6	+11.4	36.4	-6.8	0.0	36.4	-4.5	-4.5	36.4	-9.1	-9.1	31.8	-2.3	0.0
lesion count by location	54.5	-2.3	+2.3	56.8	+9.1	+11.4	45.5	+9.1	+6.8	54.5	+2.3	-2.3	50.0	0.0	-2.3	34.1	-2.3	0.0
lesion counting	77.3	0.0	0.0	47.7	+11.4	+18.2	54.5	+11.4	+22.7	79.5	-9.1	-4.5	61.4	+4.5	+13.6	88.6	0.0	0.0
lesion outlier	47.7	0.0	-2.3	40.9	+9.1	+9.1	38.6	+4.5	+6.8	47.7	-2.3	-6.8	47.7	-6.8	0.0	56.8	-4.5	-4.5
liver lesion clustering	52.3	+4.5	-2.3	38.6	+2.3	+6.8	52.3	+2.3	-4.5	54.5	+4.5	-11.4	47.7	+4.5	+2.3	63.6	+2.3	+2.3
multi-organ lesion burden comparison	47.7	-9.1	-6.8	43.2	+2.3	+2.3	50.0	-9.1	-9.1	47.7	-2.3	-2.3	47.7	-9.1	-11.4	47.7	+4.5	+4.5
organ aggregation	53.3	-2.2	+2.2	55.6	-6.7	+4.4	57.8	-6.7	-6.7	53.3	-2.2	-2.2	55.6	-2.2	+4.4	64.4	+2.2	+4.4
organ enlargement	75.0	-2.3	0.0	77.3	-4.5	-6.8	75.0	0.0	+2.3	72.7	+4.5	0.0	77.3	0.0	0.0	88.6	0.0	0.0
tumor organ HU difference	29.5	+2.3	+4.5	36.4	+2.3	+2.3	29.5	+2.3	+4.5	34.1	0.0	-4.5	20.5	+15.9	+20.5	45.5	+4.5	+2.3
Medical reasoning																		
PDAC vs PNET classification	40.6	+12.5	+6.3	40.6	+6.3	+9.4	43.8	+6.3	+6.3	53.1	-3.1	-3.1	50.0	+3.1	0.0	71.9	+3.1	+6.3
fatty liver	87.5	+12.5	+8.3	79.2	+16.7	+20.8	83.3	0.0	+4.2	100.0	0.0	0.0	87.5	+4.2	+4.2	91.7	0.0	0.0

Continued on next page

Table 11 (continued)

Subtype	AWM			DC			Evolver			GSEM			SkeMex			3DMedEvo		
	R1	Δ_2	Δ_3	R1	Δ_2	Δ_3	R1	Δ_2	Δ_3	R1	Δ_2	Δ_3	R1	Δ_2	Δ_3	R1	Δ_2	Δ_3
hepatic steatosis grading	41.9	-3.2	-3.2	48.4	-9.7	-9.7	38.7	+3.2	+6.5	35.5	-3.2	-3.2	51.6	-9.7	-16.1	58.1	0.0	-3.2
lesion type classification	8.7	0.0	0.0	13.0	0.0	-4.3	47.8	-4.3	-4.3	30.4	-21.7	-21.7	34.8	0.0	-13.0	65.2	+4.3	+4.3
pancreatic cyst resectability	61.3	-9.7	-9.7	38.7	-6.5	-3.2	48.4	0.0	0.0	58.1	-19.4	-19.4	58.1	-16.1	-6.5	58.1	-3.2	0.0
pancreatic lesion resectability	12.9	+3.2	+3.2	16.1	-3.2	-3.2	58.1	-25.8	-22.6	19.4	0.0	-3.2	45.2	0.0	-9.7	67.7	+3.2	+6.5
pancreatic steatosis	70.4	+3.7	+22.2	74.1	0.0	0.0	81.5	0.0	-3.7	85.2	-3.7	0.0	77.8	-3.7	-3.7	88.9	0.0	-3.7
pancreatic tumor staging	36.4	-21.8	-25.5	18.2	+12.7	+18.2	45.5	+9.1	-3.6	38.2	-1.8	-9.1	14.5	0.0	+3.6	20.0	-1.8	+1.8
portal hypertension assessment	87.5	0.0	0.0	87.5	-18.8	-18.8	87.5	0.0	+12.5	81.3	-25.0	-18.8	87.5	0.0	0.0	87.5	0.0	0.0
pseudocyst determination	0.0	+100.0	+66.7	66.7	0.0	0.0	66.7	-33.3	-33.3	66.7	0.0	0.0	66.7	0.0	-66.7	66.7	0.0	0.0
renal mass characterization	9.5	+4.8	0.0	19.0	-4.8	-4.8	19.0	-4.8	0.0	14.3	0.0	0.0	4.8	+9.5	0.0	38.1	0.0	+4.8
splenomegaly grading	91.9	-8.1	-8.1	62.2	-16.2	-8.1	81.1	+5.4	+5.4	86.5	0.0	+2.7	91.9	0.0	0.0	91.9	-2.7	-2.7
Recognition																		
PDAC existence	65.0	-20.0	-12.5	62.5	0.0	0.0	65.0	+7.5	+2.5	65.0	+5.0	0.0	70.0	-5.0	-10.0	67.5	0.0	-5.0
PNET existence	43.6	0.0	0.0	56.4	+7.7	0.0	51.3	+2.6	-2.6	38.5	+7.7	+10.3	41.0	+10.3	+12.8	59.0	0.0	-2.6
colon lesion existence	46.2	-2.6	-2.6	51.3	-2.6	0.0	43.6	+5.1	+5.1	43.6	0.0	+2.6	46.2	0.0	0.0	51.3	-2.6	+5.1
kidney cyst existence	43.6	+2.6	+2.6	53.8	+12.8	+20.5	66.7	-5.1	-5.1	48.7	+7.7	+15.4	43.6	+2.6	+5.1	69.2	0.0	+2.6
kidney lesion existence	48.7	0.0	-5.1	56.4	-7.7	-2.6	53.8	+2.6	-2.6	46.2	0.0	+2.6	56.4	-7.7	-10.3	56.4	+2.6	+5.1
kidney tumor existence	48.7	+5.1	+2.6	46.2	0.0	0.0	46.2	0.0	+5.1	51.3	+10.3	+5.1	59.0	+2.6	-7.7	53.8	+2.6	0.0
liver lesion existence	46.2	-2.6	-2.6	43.6	+2.6	+2.6	46.2	+2.6	0.0	46.2	0.0	0.0	46.2	0.0	0.0	46.2	+2.6	0.0
pancreatic lesion existence	43.6	0.0	0.0	41.0	-2.6	+10.3	41.0	+5.1	+2.6	43.6	0.0	0.0	43.6	+10.3	+10.3	59.0	0.0	+5.1
splenomegaly detection	89.7	0.0	0.0	79.5	+7.7	+7.7	89.7	0.0	0.0	89.7	0.0	0.0	89.7	0.0	0.0	89.7	-2.6	-2.6
Measurement																		
lesion volume measurement	28.9	0.0	+3.9	30.3	+1.3	-1.3	26.3	+1.3	-1.3	32.9	-2.6	0.0	38.2	-3.9	0.0	35.5	-2.6	0.0
organ HU measurement	64.0	+2.7	0.0	68.0	0.0	-1.3	65.3	-2.7	-4.0	60.0	+1.3	+5.3	65.3	+2.7	+4.0	81.3	-2.7	-2.7
organ HU ratio	57.4	-2.1	+4.3	57.4	+4.3	-2.1	59.6	0.0	0.0	59.6	+2.1	-2.1	59.6	+2.1	-4.3	63.8	+2.1	-2.1
organ volume measurement	48.1	+1.3	-1.3	49.4	-1.3	-2.5	46.8	+1.3	0.0	49.4	-1.3	-1.3	48.1	+2.5	0.0	54.4	+1.3	0.0
tumor burden percentage	29.3	+5.3	-1.3	30.7	-1.3	-1.3	26.7	+6.7	+6.7	22.7	+4.0	+1.3	28.0	-2.7	+1.3	28.0	+5.3	+6.7

Repository growth. Table 12 shows that 3DMedEvo progressively shifts from expanding its skill library toward refining existing experience. After three rounds, it contains 1,375 skills, compared with 2,157 entries for SkeMex and 4,990–5,070 for the other four memory baselines, while achieving the highest final-round accuracy (Table 2). More importantly, its write pattern changes across rounds: successful new-skill creations decrease from 706 to 281 per round, whereas applied revisions increase from 225 to 334. By the third round, revisions outnumber new creations, so refinement becomes more common than expansion among accepted write operations. This pattern is consistent with our admission-and-refinement design: admission selects which trajectories warrant learning, while revision allows new evidence to improve an existing procedure rather than requiring a separate entry. In contrast, the four baselines reporting no revisions continue adding more than 1,600 entries per round, and SkeMex remains predominantly creation-driven. Together, these results show that 3DMedEvo combines higher predictive accuracy with slower growth in the number of stored units, supporting self-evolution through progressive consolidation rather than unrestricted experience accumulation.

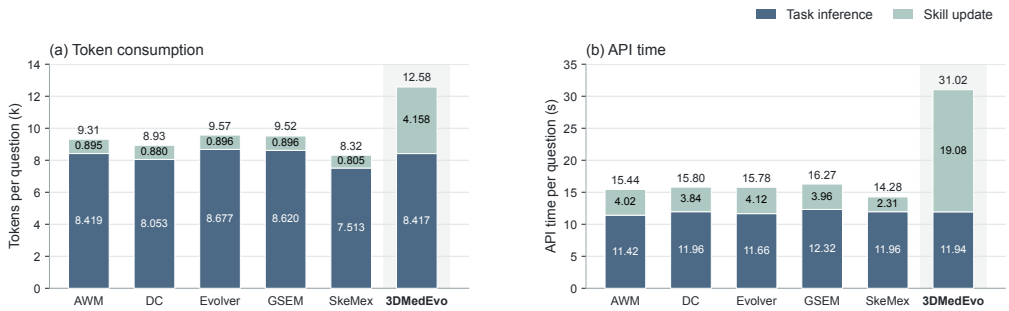


Figure 6: **Round-3 costs on DeepTumorVQA with Qwen3.6-Plus.** Stacked bars separate task inference and skill updates.

Table 12: **Write operations and repository state per round under the evolving memory setting.** Att. is create attempts, Cre. is committed creations, Pat./Rev. are patch attempts and applied revisions, Kept counts entries retained unchanged, and Rej. counts rejected operations. - denotes an operation the method does not implement.

Method	Round	Init.	Att.	Cre.	Pat./Rev.	Kept	Rej.	Final
AWM	Round@1	0	1,740	1,663	–	–	77	1,663
	Round@2	1,663	1,740	1,666	–	–	74	3,329
	Round@3	3,329	1,740	1,661	–	–	79	4,990
DC	Round@1	0	1,740	1,668	–	–	72	1,668
	Round@2	1,668	1,740	1,676	–	–	64	3,344
	Round@3	3,344	1,740	1,670	–	–	70	5,014
Evolver	Round@1	0	1,740	1,677	–	–	63	1,677
	Round@2	1,677	1,740	1,684	–	–	56	3,361
	Round@3	3,361	1,740	1,709	–	–	31	5,070
GSEM	Round@1	0	1,740	1,700	–	–	40	1,700
	Round@2	1,700	1,740	1,682	–	–	58	3,382
	Round@3	3,382	1,740	1,672	–	–	68	5,054
SkeMex	Round@1	0	1,029	744	14/7	697	292	744
	Round@2	744	987	714	12/5	741	280	1,458
	Round@3	1,458	960	699	15/6	765	270	2,157
3DMedEvo	Round@1	0	981	706	272/225	230	322	706
	Round@2	706	662	388	412/306	620	380	1,094
	Round@3	1,094	550	281	457/334	858	392	1,375

Computational cost of online evolution. Figure 6 separates task inference from skill updates in the third evaluation round. Compared with Evolver, the strongest final-round baseline, 3DMedEvo improves accuracy by 7.01 percentage points with 31.4% more total tokens. Task inference remains comparable, using 8,417 versus 8,677 tokens and 11.94 versus 11.66 API seconds per question. The additional time is concentrated in skill updates, which account for 14.96 of the 15.24 extra API seconds per question. These updates include web-based clinical grounding, which retrieves and checks external medical evidence before candidate experience is encoded for reuse. The reported API seconds aggregate recorded request durations rather than separate retrieval waits or end-to-end task latency. The additional maintenance cost therefore reflects an evidence-based medicine objective: checking candidate skills against external clinical evidence before reuse to reduce the risk of propagating unsupported clinical heuristics. This design prioritizes the safety and trustworthiness of the evolving skill library, rather than minimizing update latency alone.

B.3 Cross-Backbone Generalisation

The effectiveness of a self-evolving agent may depend on the reasoning and tool-use capabilities of its underlying backbone. To assess whether the benefits of 3DMedEvo extend beyond the model used in the main experiments, we evaluate Qwen3-VL-235B-Instruct (Qwen Team, n.d.) on DeepChestVQA and DeepTumorVQA. We compare 3DMedEvo with ReAct and the alternative agent methods under the same backbone, reporting both overall and task-family accuracy.

Figure 7 shows that 3DMedEvo achieves the highest overall accuracy on both datasets, reaching 45.18% on DeepChestVQA and 47.93% on DeepTumorVQA. The corresponding gains over ReAct are 6.47 and 7.87 points, while the margins over the strongest alternatives, AWM and Evolver, are 4.23 and 3.33 points, respectively. The benefits are not confined to a single task family: 3DMedEvo outperforms ReAct in all seven families and ranks first in six. Visual reasoning improves by 10.69 and 9.79 points on the two datasets, although Evolver remains stronger for DeepTumorVQA recognition. Together, these results show that the overall advantage of 3DMedEvo persists with an additional backbone and is accompanied by broad task-level improvements, rather than being specific to the model used in the main experiments.

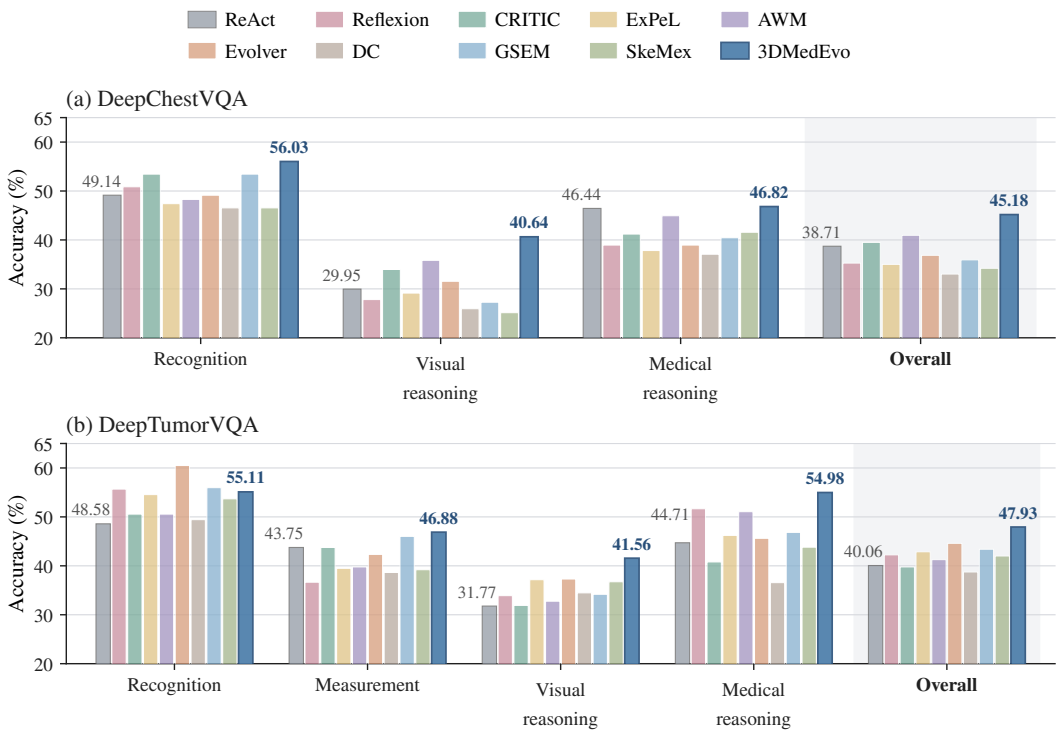


Figure 7: Task-wise and overall accuracy with Qwen3-VL-235B-Instruct on DeepChestVQA and DeepTumorVQA.

B.4 Repository Transfer Across Backbones

We examine whether the acquired experience is tied to the backbone that produced it or can also benefit models with lower and higher memory-free performance than the source. We use the skill repository constructed by Qwen3.6-Plus (Alibaba Cloud, 2026b) on DeepTumorVQA training data and evaluate its reuse by Qwen3.5-27B (Qwen Team, 2026) and Gemini-3.8-Flash (Google DeepMind, 2026) on DeepChestVQA. The repository remains fixed during evaluation, and each target is compared with its own memory-free ReAct baseline and the alternative methods under the same backbone.

Figure 8 shows that 3DMedEvo yields the largest improvement on both targets. On Qwen3.5-27B, accuracy reaches 39.23%, a gain of 2.51 points over ReAct, compared with 1.32 points for Reflexion, the strongest alternative. On Gemini-3.8-Flash, accuracy reaches 50.33%, a gain of 4.09 points, exceeding the 1.18-point gain of ExPeL. Several alternative methods instead reduce accuracy on one or both targets, indicating that sharing experience across backbones does not by itself ensure a benefit. Importantly, Gemini-3.8-Flash already outperforms the source model without memory, yet still benefits from its experience. The gains therefore extend beyond helping a weaker consumer and support reusing one skill repository across models with different starting performance.

B.5 Effect of Data Order

Sequential skill acquisition may depend not only on the cases used for construction but also on their presentation order. Early experience can shape the guidance available for later cases, raising the question of whether the observed gains require a favorable ordering or a deliberately designed curriculum. We therefore compare four construction orders using the same DeepTumorVQA training pool and fixed construction settings. We first run memory-free ReAct on the construction pool, labeling a case *easy* when its answer is correct and *hard* otherwise. **Easy-to-Hard** presents easy cases before hard cases, whereas **Hard-to-Easy** reverses that order; **Random** shuffles the full pool, and **Category-based** groups cases by question subtype with larger groups first. We construct and

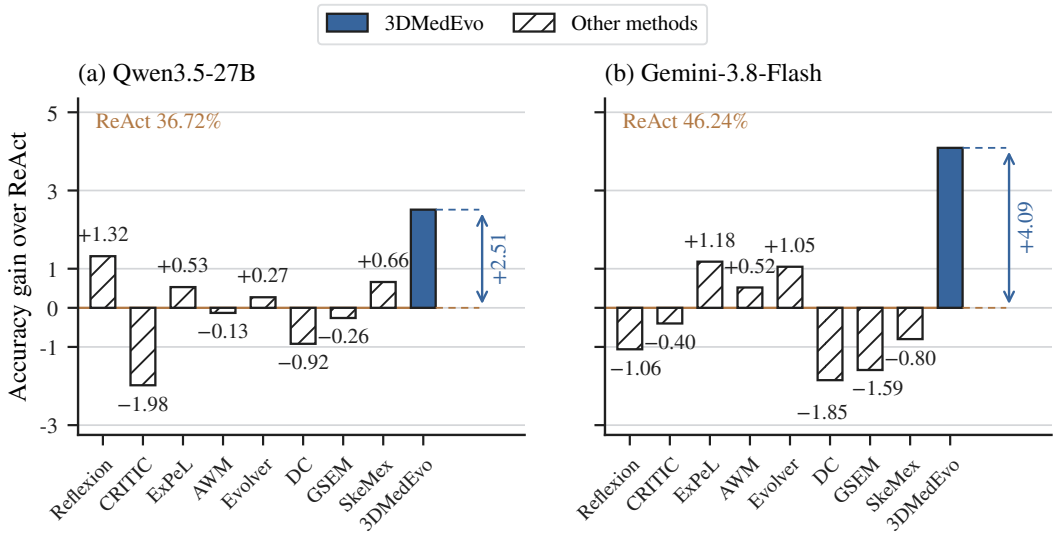


Figure 8: **Frozen-repository transfer across backbones.** Accuracy gains are relative to each target backbone's memory-free ReAct condition.

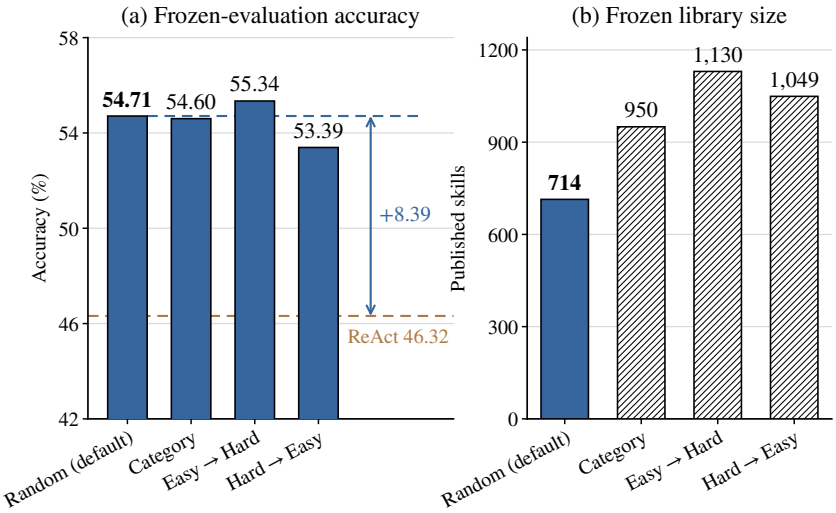


Figure 9: **Effect of construction order.** Left: frozen-evaluation accuracy. Right: the number of published skills in the frozen repository.

then freeze a separate repository for each ordering, evaluating all repositories on the same DeepTumorVQA test set and reporting accuracy and the number of published skills.

Figure 9 shows that all construction orders improve over the 46.32% memory-free ReAct result, with gains from 7.07 to 9.02 percentage points. Easy-to-Hard reaches the highest observed accuracy at 55.34%, while the default Random order reaches 54.71%, only 0.63 points lower. Random also publishes 714 skills versus 1,130 for Easy-to-Hard—36.8% fewer—while retaining near-best observed accuracy. These results support random ordering as a practical default: it yields a substantial improvement without a deliberately ordered curriculum while producing the smallest repository in this comparison.

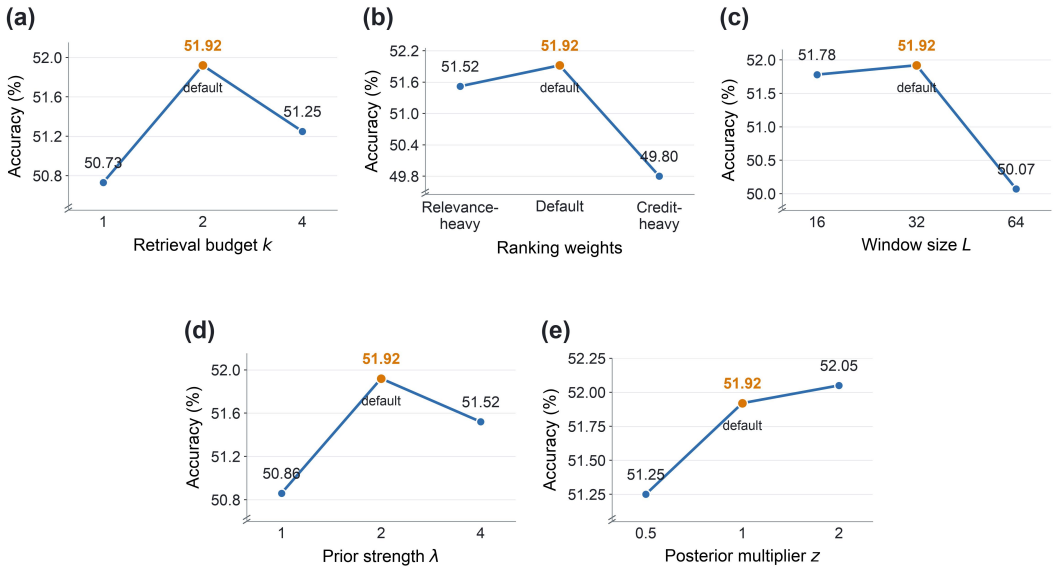


Figure 10: **Hyperparameter sensitivity.** The orange point marks the default setting in each panel; the vertical axis is shown over a local accuracy range to make differences between settings legible.

B.6 Hyperparameter Sensitivity

We analyze the sensitivity of the full Evolving Memory system on DeepChestVQA with GLM-5.2 to five hyperparameters: the per-branch retrieval quota k, ranking weights $(w_R, w_{\mathrm{SCS}}, w_A)$, acquisition window L, SCS prior strength λ, and posterior multiplier z. Figure 10 summarizes the accuracy results.

Retrieval budget. The default quota $k = 2$ reaches 51.92%, compared with 50.73% for $k = 1$ and 51.25% for $k = 4$. Thus, the intermediate budget is preferable in this sweep: retrieving too few skills leaves less branch-specific guidance, whereas increasing the quota beyond two does not improve accuracy and may introduce less applicable entries into the prompt.

Ranking channel weights. The default mixture $(0.6, 0.3, 0.1)$ obtains 51.92%. Making retrieval relevance-heavy raises the relevance weight to $(0.8, 0.1, 0.1)$ and gives 51.52%, while the credit-heavy mixture $(0.4, 0.5, 0.1)$ falls to 49.80%. The 2.12 percentage-point drop under the credit-heavy setting is the largest decrease in the sweep, suggesting that adoption history is useful as a ranking signal but should not dominate current task relevance.

Acquisition window size. The default $L = 32$ achieves 51.92%, narrowly above 51.78% for $L = 16$ and well above 50.07% for $L = 64$. The result favors a moderate window: shortening it has little effect in this measurement, whereas the larger window is associated with a 1.85 percentage-point decrease, consistent with slower adaptation and less frequent updates to the evolving repository.

Credit prior strength. For the SCS prior strength, $\lambda = 2$ is again the best of the tested default operating points, reaching 51.92% versus 50.86% for $\lambda = 1$ and 51.52% for $\lambda = 4$. This result supports the default prior strength as a balance between stabilizing credit estimates with limited adoption evidence and allowing accumulated experience to inform skill evaluation.

Posterior multiplier. Accuracy remains within a narrow 0.80-percentage-point range across the tested posterior multipliers: 51.25% at $z = 0.5$, 51.92% at $z = 1$, and 52.05% at $z = 2$. This limited variation suggests that posterior-interval governance can accommodate different interval widths without requiring a precisely tuned multiplier in this evaluation.

Overall, the tested configurations achieve accuracies between 49.80% and 52.05%. The results support a compact retrieval budget, complementary ranking signals, timely skill acquisition, and regularized credit estimation. The narrow accuracy ranges for prior strength and posterior multiplier

further suggest that the credit and governance mechanisms remain effective across the tested settings, rather than depending on a single finely tuned operating point.

B.7 Physician Verification

Adoption labels, from which SCS is computed, are assigned by the same frozen backbone that reads the CT. To validate the labels and the resulting SCS against clinical judgment, two physicians independently examined samples from the DeepChestVQA Evolving Memory run with GLM-5.2, blinded to the automatic labels and SCS.

Adoption labels. We sampled 300 trajectory–skill pairs, 100 for each automatic label. For each pair, the physicians reviewed the question, the reference answer, the skill, and the agent's trajectory with the CT views it inspected, and assigned an adoption label following Section 3.5. Disagreements were resolved by discussion. As shown in Table 13, the two physicians agreed on 89.33% of pairs (Cohen's $\kappa = 0.84$), indicating that the labeling criterion can be applied consistently. The automatic labels matched the physician consensus on 86.67% of pairs ($\kappa = 0.80$), close to the inter-physician level, so the backbone's adoption judgments provide a reliable basis for credit.

Skill credit. Among skills adopted at least three times, we selected the 50 with the highest and the 50 with the lowest final SCS. The physicians rated each skill's clinical soundness and usefulness on a five-point scale, and we compared the mean ratings of the two groups with a two-sided Mann–Whitney U test. As shown in Table 14, high-SCS skills received a mean rating of 4.10, compared with 2.93 for low-SCS skills ($p < 0.001$). Because SCS is computed from adoption outcomes rather than from skill content, this agreement with physician ratings indicates that the credit signal reflects clinical quality, supporting its use in retrieval and governance.

Together, these results indicate that adoption-aware credit is consistent with clinical judgment at both the label and the skill level.

Table 13: **Agreement on adoption labels.** 300 pairs, 100 per automatic label.

Comparison	Agreement (%)	κ
Physician 1 vs. 2	89.33	0.84
Automatic vs. consensus	86.67	0.80

Table 14: **Physician ratings of skills.** Five-point scale, averaged over both physicians.

	Mean rating	Mann–Whitney p
Highest SCS	4.10	$<$0.001
Lowest SCS	2.93	

C Method Details

3DMedEvo evolves its skill bank around a frozen VLM through retrieval, skill synthesis, adoption-aware credit, and governance. Multi-criteria retrieval selects guidance for each case from task relevance and adoption evidence. Within synthesis, skill encoding turns each evidence-checked proposal into a layered multimodal skill. Adoption-aware credit summarizes how each skill performs when it is actually adopted. Posterior-interval governance uses the uncertainty of this credit to promote, deprecate, and consolidate skills at scheduled window boundaries. We describe their designs below, and Algorithm 1 summarizes the complete procedure.

C.1 Multi-Criteria Skill Retrieval

Multi-criteria retrieval is designed to select skills that apply to the current question and have proven useful when adopted. Relevance identifies applicable experience but does not indicate whether following a skill improves outcomes, whereas historical evidence indicates usefulness but not applicability. We require the retrieval score to satisfy the following properties:

- **Relevance priority.** Historical evidence should reorder applicable skills rather than substitute for applicability.
- **Difficulty adjustment.** A skill should not appear useful merely because it was adopted on easy cases.

- **Scale comparability.** Each criterion should contribute according to its weight rather than its numerical range.
- **Pool-relative selection.** Retention should depend on a skill's standing in the current pool rather than on a tuned absolute threshold.

Relative advantage. For each skill s_i, let $\mathcal{T}_i$ denote the completed trajectories in which it was adopted, with adoption labels $g_{\tau,i}$ defined in Section 3.5 and task evaluation scores r_τ, so that n_i counts these trajectories. The baseline b_τ is the agent's reward level on comparable past tasks, estimated from earlier windows and fixed within the current one. A positive adoption is credited with its gain over the baseline, whereas a negative adoption is always penalized because the adopted guidance led to an incorrect key judgment. With $\lambda_a > 0$ controlling shrinkage toward zero, the trajectory feedback $a_{\tau,i}$, accumulated feedback h_i, and relative advantage A_i are:

$$a_{\tau,i} = \begin{cases} r_\tau - b_\tau, & g_{\tau,i} = +1, \\ -|r_\tau - b_\tau|, & g_{\tau,i} = -1, \end{cases} \qquad h_i = \sum_{\tau \in \mathcal{T}_i} a_{\tau,i}, \qquad A_i = \frac{h_i}{n_i + \lambda_a}. \tag{12}$$

Subtracting the baseline measures outcomes against comparable tasks, so a skill gains no advantage merely for being adopted on easy cases. The constant λ_a shrinks the mean feedback toward zero when adoptions are few, and a skill without adoption has $A_i = 0$.

Normalization. Relevance, credit, and advantage lie on different scales, so we standardize each criterion over the current pool. For each criterion $x \in \{R, \mathrm{SCS}, A\}$, let x_i denote its value for skill s_i, and let μ_x and σ_x denote its mean and standard deviation over the full available skill pool. With clipping bound $B > 0$, we define:

$$z_B(x_i) = \begin{cases} 0, & \sigma_x = 0, \\ \mathrm{clip}\left(\dfrac{x_i - \mu_x}{\sigma_x}, -B, B \right), & \sigma_x > 0. \end{cases} \tag{13}$$

A constant criterion cannot distinguish skills and is set to zero. Clipping keeps a single criterion from dominating when standardization amplifies small differences, for example when most skills still share the prior credit. Clipping preserves sign, so zero marks the pool mean of each criterion, and the non-negative rule keeps skills at or above this reference without an absolute threshold. Clipping also bounds the joint contribution of credit and advantage by $(1 - w_R)B$, so for $w_R > 1/2$ historical evidence reorders skills of similar relevance but cannot override a large relevance difference.

C.2 Skill Encoding

Skill Encoding $\mathcal{E}$ converts an evidence-checked proposal $\tilde{\rho}_\tau$ and its source trajectory τ into a candidate skill $\hat{s}_\tau$. Encoding abstracts the task into an applicability description, condenses the reviewed success or failure mechanism into a brief summary, and formulates a transferable procedure with operations and observable verification conditions. The procedure is restricted to tool operations available in τ, and its visual exemplars are genuine images observed in τ that illustrate the procedure and the morphology and anatomical context it relies on.

Encoding does not re-diagnose the source case, introduce teacher-generated demonstrations, or alter the proposal's intended creation or revision. It assigns the candidate to the Observe, Interpret, or Decide branch according to its primary cognitive role. A revision remains local to its target skill. Public evidence calibrates the medical claim but does not add unobserved actions or images. Neither ground-truth answers nor original reports become model-visible skill content, and adoption statistics are maintained separately from encoding.

C.3 Adoption-Aware Skill Credit

Adoption-aware credit is designed to estimate a skill's reliability from the outcomes of its adoptions. Retrieval exposure alone is not evidence of effect: a skill ignored in a correct trajectory did not cause that success. We require the Skill Credit Score to satisfy the following properties:

- **Uniform initialization.** All new skills should start from the same neutral value.
- **Adoption dependence.** Only trajectories in which a skill was actually adopted should change its score.

- **Order invariance.** The score should not depend on the order in which outcomes arrive.
- **Evidence-dependent confidence.** The score should change cautiously after few adoptions and rely increasingly on empirical outcomes as adoptions accumulate.

Skill Credit Score. We derive the score from the conjugate Beta–Bernoulli model (Gelman et al., 1995). Let $p_i \in [0,1]$ denote the latent success rate of s_i when it is adopted. Before any adoption, we assign the prior $p_i \sim \mathrm{Beta}(\lambda v_0, \lambda(1-v_0))$, which acts as λ virtual adoptions with mean credit v_0. Each adoption provides the credited outcome $x_{\tau,i} = \mathbf{1}[g_{\tau,i} = +1]\, r_\tau \in [0,1]$ with likelihood $p_i^{x_{\tau,i}}(1-p_i)^{1-x_{\tau,i}}$. For binary outcomes this is the Bernoulli likelihood, and a rubric score counts as a fractional success. By Bayes' rule, one adoption multiplies a $\mathrm{Beta}(\alpha,\beta)$ density by this likelihood:

$$p_i^{\alpha-1}(1-p_i)^{\beta-1} \cdot p_i^{x_{\tau,i}}(1-p_i)^{1-x_{\tau,i}} = p_i^{(\alpha+x_{\tau,i})-1}(1-p_i)^{(\beta+1-x_{\tau,i})-1},$$

which is proportional to a $\mathrm{Beta}(\alpha + x_{\tau,i}, \beta + 1 - x_{\tau,i})$ density. Each adoption therefore adds its credited outcome to the first parameter and the remainder to the second. Starting from the prior, these additions total c_i and $n_i - c_i$ after the updates in Eq. (8), which gives:

$$p_i \mid n_i, c_i \sim \mathrm{Beta}\big(\lambda v_0 + c_i,\ \lambda(1-v_0) + n_i - c_i\big), \qquad \mathbb{E}[p_i \mid n_i, c_i] = \frac{\lambda v_0 + c_i}{\lambda + n_i}. \tag{14}$$

The posterior mean equals the Skill Credit Score in Eq. (9). The model thus serves to derive the estimator and does not require binary outcomes. The score satisfies the required properties. It equals v_0 for every new skill and is unchanged by trajectories with $g_{\tau,i} = 0$. It depends only on (n_i, c_i), not on the order of outcomes. In Eq. (9), the empirical rate c_i/n_i receives weight $n_i/(n_i+\lambda)$, so the score stays close to v_0 after few adoptions and approaches c_i/n_i as adoptions accumulate.

Revision discount. When a skill is revised, its earlier outcomes describe the previous version. Instead of discarding them or keeping them at full weight, we retain inherited evidence with weight $\gamma \in (0,1)$, applied once when the revision is released:

$$(n_i, c_i, h_i) \leftarrow (\gamma n_i, \gamma c_i, \gamma h_i). \tag{15}$$

Joint scaling preserves the empirical rates but lowers their weight relative to the prior, so SCS_i moves toward v_0 and A_i toward zero until the revised procedure accumulates its own adoptions. The discount covers evidence from earlier windows; adoptions from the window that produces the revision are then added at full weight (Algorithm 1), since they are the most recent evidence and include the outcomes that prompted the revision.

C.4 Posterior-Interval Governance

Posterior-interval governance is designed to change a skill's status only when its adoption evidence supports the change. Governance decisions persist across windows. In particular, a deprecated skill leaves the retrieval pool, so it is no longer adopted and new evidence can no longer revise its credit. However, the adoption evidence available at a governance boundary is sparse and noisy: a skill is adopted only on relevant cases, and each outcome also varies with case difficulty and model stochasticity. A point estimate does not reflect this uncertainty. A skill with few adoptions and a skill with many adoptions can have the same SCS, although only the latter estimate is reliable. Comparing SCS directly with a reference would therefore promote skills after a few favorable adoptions and permanently deprecate skills after a few unfavorable ones.

To account for this uncertainty, governance promotes a skill only when a conservative estimate of its credit exceeds the reference and deprecates it only when an optimistic estimate falls below the reference. Otherwise, the skill keeps its status until further adoptions accumulate. The posterior in Eq. (14) already quantifies the uncertainty of SCS_i, so we derive both estimates from it without an additional model. Figure 3c shows that replacing the intervals with direct score comparisons reduces accuracy. We require governance to satisfy the following properties:

- **Evidence-aware decisions.** For the same score, a skill supported by fewer adoptions should require a larger separation from the reference before its status changes.
- **Evidence preservation.** Consolidating near-duplicate skills should pool their adoption evidence rather than discard it.

Posterior interval. A $\mathrm{Beta}(\alpha, \beta)$ distribution has mean $m = \alpha/(\alpha+\beta)$, and since $\beta/(\alpha+\beta) = 1-m$, its variance can be written as:

$$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{\alpha}{\alpha+\beta}\cdot\frac{\beta}{\alpha+\beta}\cdot\frac{1}{\alpha+\beta+1} = \frac{m(1-m)}{\alpha+\beta+1}.$$

For the posterior in Eq. (14), $m = \mathrm{SCS}_i$ and $\alpha+\beta = \lambda + n_i$. With posterior multiplier $z > 0$, we place the bounds z posterior standard deviations from the mean and truncate them to $[0,1]$:

$$\sigma_i^2 = \frac{\mathrm{SCS}_i(1-\mathrm{SCS}_i)}{\lambda+n_i+1}, \qquad L_i = \max\{0, \mathrm{SCS}_i - z\sigma_i\}, \qquad U_i = \min\{1, \mathrm{SCS}_i + z\sigma_i\}. \quad (16)$$

The bounds $L_i \le \mathrm{SCS}_i \le U_i$ serve as the conservative and optimistic estimates of the credit. Since $\mathrm{SCS}_i(1-\mathrm{SCS}_i) \le 1/4$, the interval width is at most $z/\sqrt{\lambda+n_i+1}$, so for the same score the interval narrows as adoptions accumulate. A skill with fewer adoptions therefore requires a larger separation from the reference before its status changes.

Governance reference. For branch b with retrieval-eligible pool P_b, the reference rank K_b determines:

$$\beta_b = \begin{cases} v_0, & |P_b| \le K_b, \\ \mathrm{SCS}_{(K_b)}, & |P_b| > K_b, \end{cases} \quad (17)$$

where $\mathrm{SCS}_{(K_b)}$ is the K_b-th highest SCS in P_b. A skill without adoption has $\mathrm{SCS}_i = v_0$ and $\sigma_i > 0$, so $L_i < v_0 < U_i$, and it is neither promoted nor deprecated against $\beta_b = v_0$.

Near-duplicate consolidation. Separate trajectories often produce near-duplicate skills. Their adoptions are then split among these skills, so each skill keeps a wide interval and may never reach a decision. Two skills in the same branch are treated as near-duplicates when their applicability descriptions have cosine similarity above a threshold under ψ. For each such pair, merging keeps the skill s_d with the higher SCS, adds the evidence of the other skill s_l to it, and deprecates s_l:

$$(n_d, c_d, h_d) \leftarrow (n_d + n_l,\ c_d + c_l,\ h_d + h_l). \quad (18)$$

Pooling conserves the total adoption evidence of the branch. The merged interval has width at most $z/\sqrt{\lambda+n_d+n_l+1}$, so two skills that are individually inconclusive can reach a decision after merging. Consolidation precedes the interval tests, and β_b is computed on the consolidated pool.

C.5 Pseudocode

In Algorithm 1, each acquisition window reads a fixed bank $\mathcal{K}^{(w)}$, and synthesis, credit, and governance take effect only at the window boundary, following Eq. (11). Deployment reuses the retrieval rule with the final bank and does not update it.

D Tool Suite

Our tool suite supports three complementary operations: (1) visual inspection, (2) quantitative anatomical assessment, and (3) clinical knowledge lookup. Within each benchmark, tool definitions are held fixed across the compared memory methods. Tools return visual observations or structured textual evidence to support subsequent reasoning and action selection.

For DeepChestVQA (Wang et al., 2026a), the agent uses three visual tools to inspect the CT montage: region zoom, panel inspection, and regional comparison. These tools return localized image observations that guide subsequent reasoning.

- **Region zoom.** Given an image and a normalized bounding box, the tool converts the box coordinates to pixel locations and extracts the corresponding rectangular crop. It returns the cropped image with a brief textual observation, allowing closer inspection of an agent-selected region without running a separate diagnostic model.
- **Panel inspection.** The tool partitions an image into a regular grid using agent-specified numbers of rows and columns. It extracts cells in row-major order and returns the first eight at most, together with a description of the grid. This provides an ordered sequence of local views for systematic inspection.

Algorithm 1 Closed-loop skill evolution in 3DMedEvo.

Require: initial bank $\mathcal{K}^{(0)}$; acquisition windows $\{\mathcal{X}^{(w)}\}_{w=0}^{W-1}$ of L cases each; deployment cases $\mathcal{X}_{\text{test}}$
Require: frozen VLM p_θ; text encoder ψ; per-branch quota k; governance period ν
Ensure: evolved skill bank $\mathcal{K}^{(W)}$ and deployment responses

for $w = 0$ **to** $W - 1$ **do** ▷ acquisition
  $\mathcal{D}^{(w)} \leftarrow \varnothing$
  for each case $(V, q) \in \mathcal{X}^{(w)}$ **do**
    $\mathcal{C} \leftarrow \textsc{Retrieve}(q, \mathcal{K}^{(w)})$
    Act with $a_t \sim p_\theta(\cdot \mid h_t, \mathcal{C})$ from h_0 until the final response, yielding τ
    Evaluate the response to obtain r_τ; $\mathcal{D}^{(w)} \leftarrow \mathcal{D}^{(w)} \cup \{\tau\}$
  end for
  $\Delta_{\text{skill}}^{(w)} \leftarrow \varnothing$ ▷ synthesis, Section 3.4
  for each $\tau \in \mathcal{D}^{(w)}$ **do**
    Admission: obtain $\mathcal{I}(\tau)$ and $g_{\tau,i}$ for each skill s_i retrieved in τ
    if $\mathcal{I}(\tau) = 1$ **then** ▷ $\tau \in \mathcal{B}^{(w)}$
      $\tilde{\rho}_\tau \leftarrow \mathcal{G}(\rho_\tau)$ and $\hat{s}_\tau \leftarrow \mathcal{E}(\tilde{\rho}_\tau, \tau)$ for each $\rho_\tau \in \mathcal{R}(\tau; \mathcal{K}^{(w)})$
      $\Delta_{\text{skill}}^{(w)} \leftarrow \Delta_{\text{skill}}^{(w)} \cup \mathcal{W}(\mathcal{K}^{(w)}, \tau)$ ▷ peer review $\mathcal{P}$, Eq. (6)
    end if
  end for
  $\widetilde{\mathcal{K}}^{(w)} \leftarrow \mathcal{S}(\mathcal{K}^{(w)}, \Delta_{\text{skill}}^{(w)})$ ▷ new skills start with $n_i = c_i = h_i = 0$
  Discount each skill revised in $\Delta_{\text{skill}}^{(w)}$ by Eq. (15) ▷ credit $\mathcal{F}$, Section 3.5
  for each $\tau \in \mathcal{D}^{(w)}$ and each s_i retrieved in τ with $g_{\tau,i} \neq 0$ **do**
    Update (n_i, c_i) by Eq. (8) and $h_i \leftarrow h_i + a_{\tau,i}$ by Eq. (12)
  end for
  Update the comparable-task baselines with $\mathcal{D}^{(w)}$ for subsequent windows
  if $(w + 1) \bmod \nu = 0$ **then**
    $\mathcal{K}^{(w+1)} \leftarrow \textsc{Govern}(\widetilde{\mathcal{K}}^{(w)})$
  else
    $\mathcal{K}^{(w+1)} \leftarrow \widetilde{\mathcal{K}}^{(w)}$
  end if
end for
for each case $(V, q) \in \mathcal{X}_{\text{test}}$ **do** ▷ read-only deployment
  $\mathcal{C} \leftarrow \textsc{Retrieve}(q, \mathcal{K}^{(W)})$; act with $p_\theta(\cdot \mid h_t, \mathcal{C})$ and return the final response
end for

function $\textsc{Retrieve}(q, \mathcal{K})$ ▷ Section 3.3
  Compute $\text{score}(q, s_i; \mathcal{K})$ by Eq. (4) for each non-deprecated $s_i \in \mathcal{K}$
  return the at most k highest non-negative-scoring skills of each branch
end function

function $\textsc{Govern}(\mathcal{K})$ ▷ $\mathcal{Q}$, Section 3.6
  for each branch b **do**
    $P_b \leftarrow$ non-deprecated skills of branch b in $\mathcal{K}$
    for each near-duplicate pair $\{s_d, s_l\} \subseteq P_b$ with $\text{SCS}_d \geq \text{SCS}_l$ **do**
      Pool the statistics of s_l into s_d by Eq. (18); deprecate s_l; $P_b \leftarrow P_b \setminus \{s_l\}$
    end for
    Compute β_b on P_b by Eq. (17)
    for each $s_i \in P_b$, with $[L_i, U_i]$ from Eq. (16) **do**
      Mark s_i mature if $L_i > \beta_b$, deprecated if $U_i < \beta_b$, and leave it unchanged otherwise
    end for
  end for
  return $\mathcal{K}$
end function

- **Regional comparison.** The tool extracts two to four normalized regions from the same image and returns their crops in the requested order. The reasoning model can then compare anatomical regions or corresponding montage panels; the tool supplies the images rather than computing a diagnostic or similarity score.

For DeepTumorVQA (Chen et al., 2025), we use the official *predicted* mode for tool-augmented evaluation (Chen et al., 2026). Segmentation summaries and measurements are retrieved from precomputed TotalSegmentator-based caches (Wasserthal et al., 2023), rather than oracle measurements or live segmentation during an interaction. All case-specific queries are bound to the current examination.

- **Segmentation summary.** Given an organ or lesion target, the tool retrieves its cached segmentation record for the current examination. It returns mask availability and available spatial summaries, including voxel count, bounding box, and center of mass, rather than the voxel mask itself. Missing entries are reported together with the available target names.
- **Quantitative measurement.** The tool looks up a cached value by anatomical target and measurement type. Supported quantities include volume, mean CT attenuation in Hounsfield units, maximum diameter, lesion count, and the volume and diameter of the largest lesion. Results retain their stored units and accompanying statistics when available. Measurements are read directly from the cache; unavailable entries are reported rather than recomputed.
- **Clinical knowledge lookup.** The tool searches the benchmark's local knowledge base of anatomical reference values, diagnostic thresholds, and clinical criteria. It lowercases the query and ranks entries using keyword substring matches and exact word overlap. Up to two matching passages are returned with their stored source labels, providing reference information without web search or an additional language-model call.

E BASELINES

We organize the baselines into four categories according to their primary mechanisms. (1) Medical vision–language models learn domain-specific representations from medical images and text. (2) Tool-assisted agents combine reasoning with external perception and information-gathering tools. (3) Reflection and self-correction methods use feedback to revise responses or improve later attempts. (4) Experience-based memory agents retain and organize information from previous tasks as reusable guidance for subsequent decisions.

E.1 MEDICAL VISION–LANGUAGE MODELS

- **MedGemma-27B** (Sellergren et al., 2025) extends Gemma with medical image and text training. The model supports medical visual understanding and language-based reasoning while retaining the general capabilities of its base model.
- **Lingshu-32B** (Xu et al., 2026a) is a generalist medical vision–language model for unified understanding and reasoning. Its training combines medical visual and textual knowledge to support diverse tasks, including visual question answering and report generation.
- **HuluMed-32B** (Jiang et al., 2025a) integrates medical text, 2D images, 3D volumes, and video within a single architecture. It uses medical-aware visual token reduction to process redundant volumetric and video inputs more efficiently.
- **RadFM-14B** (Wu et al., 2025a) combines radiological images and language through large-scale training on heterogeneous 2D and 3D medical data. Its generative formulation supports multiple radiology tasks, including visual question answering, diagnosis, and report generation.

E.2 TOOL-ASSISTED AGENTS

- **ReAct** (Yao et al., 2022) alternates language-based reasoning with actions that interact with external tools or environments. Retrieved observations inform subsequent reasoning, allowing the agent to revise its plan and gather evidence as a task progresses.
- **3DMedAgent** (Wang et al., 2026a) enables 2D multimodal models to analyze volumetric CT by coordinating perception tools, structured textual evidence, and targeted slice inspection. Intermediate findings are accumulated in structured memory, while a Think-with-1-Slice loop supports iterative visual verification.

E.3 Reflection and Self-Correction

- **Reflexion** (Shinn et al., 2023) turns task feedback into verbal reflections stored in an episodic memory. These reflections guide subsequent attempts, enabling an agent to learn from prior outcomes without updating its model parameters.

- **CRITIC** (Gou et al., 2024) improves an initial response through repeated verification and revision. It consults external tools to check the response and uses the resulting feedback to identify and correct errors.

E.4 Experience-Based Memory Agents

- **ExPeL** (Zhao et al., 2024) derives reusable natural-language insights from accumulated task experience. At inference, it recalls relevant experiences and distilled insights to guide decisions without parameter updates. Relative to text-centered insights, 3DMedEvo stores scoped CT procedures together with historical visual exemplars and checks medical claims against external evidence.

- **Agent Workflow Memory (AWM)** (Wang et al., 2024) extracts recurring action sequences as reusable workflows and selectively provides them to an agent during task execution. Workflows can be induced from a fixed collection of training trajectories or accumulated online through new interactions. AWM represents recurring action order as workflows; 3DMedEvo represents experience as multimodal CT skills organized around complementary Observe, Interpret, and Decide responsibilities, with scoped procedures and historical visual exemplars.

- **Evolver (EvolveR)** (Wu et al., 2025b) links offline experience distillation with online interaction and policy improvement. It converts past trajectories into strategic principles, retrieves these principles during task solving, and incorporates the resulting experience into subsequent learning cycles. 3DMedEvo formulates transferable guidance as evidence-grounded CT-reading procedures with explicit applicability and validation conditions, paired with historical visual exemplars.

- **Dynamic Cheatsheet (DC)** (Suzgun et al., 2026) maintains persistent notes containing useful strategies, intermediate findings, and reusable solutions. The notes are revised across tasks and supplied as context, enabling test-time learning without changing model weights. Compared with persistent notes, 3DMedEvo uses multimodal CT skills that jointly encode scoped procedures, applicability and validation conditions, and historical visual exemplars.

- **GSEM** (Han et al., 2026) represents clinical experience using a dual-layer graph of decision structures and inter-experience relations. Applicability-aware retrieval selects relevant experience, while online feedback calibrates node quality and connection weights. The contrast concerns how experience is represented: graph-linked clinical decisions in GSEM versus evidence-grounded multimodal CT-reading procedures in 3DMedEvo.

- **SkeMex** (Sun et al., 2026) transforms interaction histories into procedural skills organized at general, task-specific, and action levels, which are retrieved to guide subsequent tasks. 3DMedEvo focuses on multimodal CT-reading skills, pairing scoped procedures with historical visual exemplars and checking their medical claims against clinical evidence.

Frozen-memory baselines follow their original procedures, including any reflection, verification, or retry steps, so their tool calls per question are not bounded by the turn budget of Appendix G.

F Metrics

We use question-level exact-match accuracy for closed-ended, single-answer multiple-choice QA. All methods follow the same evaluation sequence: Gemini-2.5-Flash (Google DeepMind, 2025) extracts the final option from the response, the extracted option is compared with the reference, and correctness is aggregated across questions. Invalid extractions are counted as incorrect. Neither free-form semantic scoring nor partial credit is used.

F.1 ACCURACY

Our primary metric is question-level exact-match accuracy:

$$\mathrm{Acc}(\mathcal{D}) = \frac{100}{N} \sum_{i=1}^{N} \mathbb{I}[\widehat{a}_i = a_i], \tag{19}$$

where N is the number of questions and each question contributes equally, including distinct questions associated with the same CT volume. For each question subtype or type, accuracy is computed on the corresponding subset; overall accuracy is weighted by the number of questions in each group rather than by an unweighted average of subgroup scores. The same aggregation is applied independently to each evaluation round, and differences are reported in percentage points.

F.2 TOOL USE, MEMORY, AND EFFICIENCY

Let t_i denote the number of recorded tool calls for question i. We report mean tool calls per question as follows:

$$\mathrm{Calls/Q} = \frac{1}{N} \sum_{i=1}^{N} t_i, \tag{20}$$

where "tool call" and "tool action" denote the same logged invocation. A memory repository is reported by its entry count, with entries defined by each method's representation (e.g., workflows, reflections, graph nodes, or skills); these units need not have equal granularity. For each reported cost summary, token consumption and API time are averaged over the questions covered by the corresponding construction or evaluation logs. Task-inference and memory-update components use the same question denominator, with questions without an update contributing zero to the latter. API time aggregates recorded request durations and is not an end-to-end task-latency measure.

G IMPLEMENTATION DETAILS

Frozen VLM Configuration. We use GLM-5.2 (Z.ai, n.d.) and Qwen3.6-Plus (Alibaba Cloud, 2026b) as the primary backbones. Within a run, the CT-reading agent and every LLM-based role in outcome judging, admission, case review, grounding, encoding, and peer review share the same frozen backbone and provider, rather than using a separate evolution teacher. Calls are made through API endpoints, with the repository, credit statistics, and interaction records maintained locally. Agent and evolution requests use temperature 0 and at most 32,768 output tokens. Gemini-2.5-Flash (Google DeepMind, 2025) is used separately only to extract a final option for exact-match evaluation (Appendix F), not to guide evolution.

CT-Reading Runtime. Each case follows a bounded ReAct trajectory (Yao et al., 2022). Its initial context combines the question, dataset-supplied initial media, live tool descriptions, and the retrieved skills. On every request, the model returns reasoning followed by either one structured tool invocation or a final answer; a completed observation is then appended to the case history before the next request. We allow six tool-enabled turns on DeepChestVQA and five on DeepTumorVQA. If this budget expires without an answer, one additional tool-disabled request requires a best-supported final response. Tool observations and the completed trajectory are retained for later experience screening, whereas the selected skills do not change within the case. The interaction protocol is specified in Appendix J.1.

Multimodal Skill Memory. The repository stores each skill in one of the Observe, Interpret, and Decide branches as a stable skill identifier with version-linked revisions. A record includes its applicability description, scoped procedure, and observable validation conditions, together with links to its source trajectory evidence and to visual exemplars. These exemplars are associations to views observed in prior cases, so they can illustrate the procedure and relevant visual context without becoming evidence for a new patient. Stable identifiers and versions connect retrieval, local revision, adoption attribution, and credit aggregation across windows. The model receives compact, selected skill guidance rather than raw historical trajectories; case answers and other patient-specific conclusions are not copied into skills.

Multi-Criteria Skill Retrieval. At case initialization, semantic relevance is computed from text-embedding-v4 (Alibaba Cloud, 2026a) representations of the question and the available skills' applicability descriptions. We score the available skills with Eq. (4), using $(w_R, w_{\mathrm{SCS}}, w_A) = (0.6, 0.3, 0.1)$ and clipping bound $B = 2$. Within each branch, only non-negative scores are retained, in descending order, with at most $k = 2$ skills. A branch with fewer eligible skills is not padded with lower-ranked negative-score candidates. Skills are selected once per case; the selected subset remains fixed and is supplied unchanged whenever the runtime reconstructs a subsequent request. Adoption feedback therefore refers to a single, known selection.

Evidence-Grounded Skill Synthesis. After a completed trajectory passes admission, one Case Review call produces proposals for skill creation or local revision from the visible behavior and authorized outcome feedback. Each eligible proposal is checked against retrieved public clinical evidence before encoding: a CONFLICT verdict blocks encoding, whereas an INSUFFICIENT verdict preserves the proposal's uncertainty rather than asserting clinical support. A retained proposal receives exactly one encoding pass, which creates a scoped procedure with validation conditions, source-trajectory references, and available visual associations. It then receives exactly one peer-review pass. Review compares at most three same-branch reference skills from the frozen bank, including the target for a revision, and only accepts or rejects the candidate without rewriting it. The detailed role inputs and outputs are given in Appendix J.2.

Adoption-Aware Skill Credit. New skills start with $\mathrm{SCS}_i = v_0 = 0.5$ and $A_i = 0$. For each completed case, adoption feedback is associated with the frozen skill version that was selected for that trajectory; the aggregated statistics are then used by the next window. We set the SCS prior strength to $\lambda = 2$ and the relative-advantage shrinkage to $\lambda_a = 2$. The comparable-task baseline is v_0 until five historical outcomes are available, avoiding a case-specific baseline estimate with insufficient history. When a skill is locally revised, the inherited adoption count and accumulated credit are multiplied by $\gamma = 0.5$ as in Eq. (15); new adoption increments are added without discounting.

Posterior-Interval Governance. Governance uses the statistics updated for the current window, with posterior multiplier $z = 1$, and is invoked every three acquisition windows. We set the branch reference rank to $K_b = 16$. For a branch with more than 16 retrieval-eligible skills, β_b is the 16th-highest SCS; otherwise, $\beta_b = v_0 = 0.5$. This rank determines the governance reference, not a storage limit or a restriction to the top 16 retrieval candidates. Within a branch, candidates with similarity at least 0.90 are consolidated, pooling their credit statistics and deprecating the lower-credit duplicate. Active and mature skills remain eligible for later retrieval, while deprecated skills no longer enter the subsequent candidate pool. These decisions are made at the scheduled boundary, so a running window always reads a stable repository.

Closed-Loop Evolution. An acquisition window contains $L = 32$ cases. Cases within the window, including concurrently executed cases, share the same frozen repository snapshot and submit their experience only after completion; no case changes the retrieval state for another case in that window. At the boundary, accepted skill updates are applied, credit statistics are updated from the window feedback, and scheduled governance is performed in the order of Eq. (11). The resulting snapshot is used for retrieval in the next window. Under Frozen Memory, the final repository and its statistics are read-only at deployment. Under Evolving Memory, the same snapshot-and-boundary procedure continues across windows.

Evolving-Memory Protocol. Each evolving run starts from an empty skill bank, and runs with different backbones are initialized independently. The same evaluation cases are processed in three rounds (R1–R3 in Table 2), and the bank and its credit statistics carry over from one round to the next. Within a round, each case is answered with the snapshot of its window, and its experience can change the bank only at the next window boundary, after the answer has been recorded. The reference answer of a case is used only to score its finalized response: the outcome judge (Prompt 9) returns a binary correctness score for memory learning, and the answer text never enters task execution, admission, case review, grounding, encoding, or peer review. The skill-generation prompts (Prompts 4, 5, and 7) further prohibit carrying case answers, report text, coordinates, or patient details into skills, and require each skill to specify a trigger, an executable procedure, and a decision

Table 15: **Implementation settings.** Default settings for 3DMedEvo unless explicitly varied in an experiment.

Module	Parameter	Setting
Retrieval	Text relevance encoder	text-embedding-v4
	$(w_R, w_{\mathrm{SCS}}, w_A)$ ranking weights	$(0.6, 0.3, 0.1)$
	B clipping bound	2
	k per-branch quota	2
	Retention threshold	0 (score ≥ 0)
Synthesis	Peer-review reference skills	At most 3
Credit	v_0 SCS prior mean	0.5
	λ SCS prior strength	2
	λ_a relative-advantage shrinkage	2
	γ revision history retention	0.5
	Minimum comparable-task history	5
Governance	z posterior multiplier	1
	K_b branch reference rank	16
	Merge similarity threshold	0.90
	Governance interval	Every 3 windows
Evolution	L acquisition window size	32 cases

rule for a task type rather than for its source case. Thus, when a case recurs in a later round, the bank offers reusable reading mechanisms learned in earlier rounds rather than a stored answer.

H Dataset Details

In this section, we describe the datasets used to evaluate 3DMedEvo and acquire experience for offline skill evolution. We first detail the task coverage and composition of DeepChestVQA and DeepTumorVQA, then explain how a reasoning-intensive training pool is curated from the DeepTumorVQA training split. We also report the retained-subset statistics and the composition of the final training pool to clarify the evaluation scope and the distribution of cases available for skill acquisition.

H.1 Details of DeepChestVQA and DeepTumorVQA

DeepChestVQA (Wang et al., 2026a) evaluates chest CT understanding, from detecting localized abnormalities to interpreting scan-level imaging patterns. Its construction draws on CT-RATE (Hamamci et al., 2026) images and ReXGroundingCT (Baharoon et al., 2026) lesion annotations, together with anatomical segmentations, to formulate questions about airway, pulmonary, and pleural findings.

The tasks are organized into recognition, visual reasoning, and medical reasoning. Recognition identifies abnormalities in a target structure. Visual reasoning covers lesion localization, counting, and quantitative comparisons, including CT attenuation in Hounsfield units (HU). Measurement-related questions are included in this family rather than treated as a separate category. Medical reasoning integrates imaging patterns and disease burden to characterize phenotypes or grade severity. Table 16 summarizes the individual tasks.

Table 16: **DeepChestVQA question subtypes.**

Subtype	Task description
Recognition	
Bronchus/airway lesion existence	Detect an endobronchial lesion or other intraluminal airway abnormality.
Lung lesion existence	Determine whether a pulmonary lesion is present.

Continued on next page

Table 16: **DeepChestVQA question subtypes (continued).**

Subtype	Task description
Pleura lesion existence	Determine whether a pleural lesion or other pleural abnormality is present.
Visual reasoning	
Largest lesion diameter	Estimate the maximum in-plane diameter of the largest pulmonary lesion.
Largest lesion location	Identify the lung lobe containing the largest lesion.
Largest lesion slice	Locate the maximal cross-section of the largest lesion as a relative position within the CT stack.
Lesion counting	Count distinct lesion instances in the scan.
Lesion count by location	Count lesions within a specified lung lobe.
Organ enlargement	Determine whether the target organ is abnormally enlarged.
Organ atrophy	Identify abnormal volume reduction of the target organ.
Lesion–organ HU difference	Estimate the attenuation difference between a lesion and the surrounding lung tissue.
Medical reasoning	
Attenuation pattern classification	Classify pulmonary opacity by its attenuation pattern, such as ground-glass opacity or consolidation.
Volume-loss lesion classification	Distinguish atelectasis from consolidation using evidence of parenchymal volume loss.
Imaging phenotype synthesis	Integrate local findings into the dominant scan-level imaging phenotype.
Phenotype mixing identification	Determine whether multiple distinct imaging phenotypes coexist in the same scan.
Emphysema severity grading	Grade emphysema severity from its extent relative to lung volume.
Pleural effusion grading	Grade pleural effusion severity from its relative volumetric burden.

DeepTumorVQA (Chen et al., 2025; 2026) evaluates tumor-centric abdominal CT understanding across the liver, pancreas, kidneys, spleen, and colon. It draws on public CT resources, including AMOS (Ji et al., 2022), AbdomenCT-1K (Ma et al., 2021), KiTS (Heller et al., 2021; 2023), and the Medical Segmentation Decathlon (Antonelli et al., 2022), and uses organ and lesion annotations to connect anatomical findings with quantitative and clinical questions. Disease-specific tasks include pancreatic ductal adenocarcinoma (PDAC) and pancreatic neuroendocrine tumor (PNET).

Its taxonomy distinguishes recognition, measurement, visual reasoning, and medical reasoning. Recognition addresses lesion presence and organ-level abnormalities; measurement covers volumes, attenuation, and derived ratios. Visual reasoning requires spatial, instance-level, and cross-organ comparisons. Medical reasoning interprets these findings using benchmark-defined diagnostic, staging, and management criteria. Table 17 provides a brief description of each subtype.

Table 17: **DeepTumorVQA question subtypes.**

Subtype	Task description
Recognition	
PDAC existence	Determine whether PDAC is present in the pancreas.
PNET existence	Determine whether PNET is present in the pancreas.
Colon lesion existence	Determine whether a lesion is present in the colon.
Kidney cyst existence	Determine whether a renal cyst is present.
Kidney lesion existence	Determine whether any renal lesion is present.
Kidney tumor existence	Determine whether a renal tumor is present.
Liver lesion existence	Determine whether a hepatic lesion is present.
Pancreatic lesion existence	Determine whether a pancreatic lesion is present.
Splenomegaly detection	Detect splenic enlargement from spleen volume.

Continued on next page

Table 17: **DeepTumorVQA question subtypes (continued).**

Subtype	Task description
Measurement	
Lesion volume measurement	Measure lesion volume within the target organ.
Organ HU measurement	Measure the mean CT attenuation of the specified organ.
Organ HU ratio	Compute the ratio of mean CT attenuation between the specified organs.
Organ volume measurement	Measure the volume of the specified organ.
Tumor burden percentage	Express lesion volume as a percentage of the corresponding organ volume.
Visual reasoning	
Adjacent organ	Identify the organ adjacent to the largest lesion in the target organ.
Bilateral kidney asymmetry	Compare lesion involvement between the left and right kidneys.
Inter-segment comparison	Compare lesion counts between specified liver segments.
Kidney volume comparison	Compare left and right kidney volumes.
Largest lesion attenuation	Classify the CT attenuation of the largest lesion.
Largest lesion diameter	Determine the maximum diameter of the largest lesion.
Largest lesion location	Identify the anatomical region containing the largest lesion.
Largest lesion slice	Locate the largest lesion within the CT stack using relative slice position.
Lesion count by location	Count lesions within a specified anatomical region.
Lesion counting	Count instances of the specified lesion type within the target organ.
Lesion outlier	Determine whether the largest lesion is a volumetric outlier relative to the remaining lesions.
Liver lesion clustering	Characterize whether lesions are clustered or dispersed across hepatic segments.
Multi-organ lesion burden comparison	Compare relative tumor burden across the specified organs.
Organ aggregation	Compute the combined volume of the specified organs.
Organ enlargement	Determine whether the target organ exceeds its reference volume range.
Tumor–organ HU difference	Compute the attenuation difference between the largest lesion and its host organ.
Medical reasoning	
PDAC vs PNET classification	Distinguish PDAC from PNET using pancreatic tumor characteristics.
Fatty liver	Assess hepatic fat accumulation from liver attenuation relative to the spleen.
Hepatic steatosis grading	Grade hepatic steatosis using liver and spleen attenuation.
Lesion type classification	Classify the largest lesion as a cyst or tumor.
Pancreatic cyst resectability	Assess whether the largest pancreatic cyst meets the criteria for resection.
Pancreatic lesion resectability	Categorize resectability of the dominant pancreatic lesion from its imaging features.
Pancreatic steatosis	Assess pancreatic fat accumulation from pancreatic attenuation relative to the spleen.
Pancreatic tumor staging	Assign a pancreatic tumor T stage from its imaging characteristics.
Portal hypertension assessment	Assess signs of portal hypertension from spleen volume and hepatic findings.
Pseudocyst determination	Determine whether the dominant pancreatic cyst has attenuation consistent with a pseudocyst.
Renal mass characterization	Characterize a renal mass according to its non-contrast CT attenuation.
Splenomegaly grading	Grade the severity of splenic enlargement using spleen volume.

H.2 TRAINING-SET CURATION

The original DeepTumorVQA training split contains 428,050 question–answer pairs. A substantial fraction of this scale arises from repeatedly querying the same CT volume or expressing the same underlying task through multiple linguistic templates. Although such expansion is useful for conventional supervised learning, it is unnecessarily expensive for experience-grounded harness evolution, where each example may require multi-step reasoning, tool execution, and trajectory-level evaluation. We therefore construct a compact pool that prioritizes reasoning depth, case diversity, and reusable problem-solving procedures rather than raw sample count.

We first enforce strict separation from evaluation data at the CT-volume level. Any training case appearing in the complete official evaluation pool is removed before subsequent processing. This case-level criterion is stronger than question-level de-duplication because different question subtypes derived from the same scan may expose shared anatomical or pathological evidence. We then retain only question subtypes whose reference procedures require at least three reasoning steps. Within the resulting pool, higher priority is assigned to cases involving clinical knowledge integration, multi-organ comparison, spatial localization, counting, quantitative measurement, threshold-based interpretation, or ratio computation. Difficulty is therefore determined from the annotated solution structure rather than superficial properties such as prompt length, lexical rarity, or answer-option position.

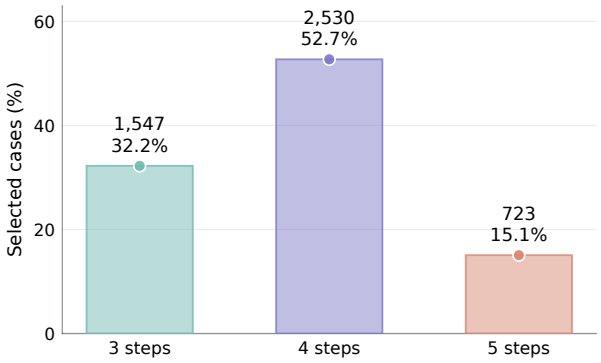


Figure 11: **Reference reasoning depth in the retained subset.**

We additionally require every retained example to provide a unique and stable supervision signal. Samples with ambiguous multi-answer targets, duplicated answer contents, degenerate option sets, or inconsistent question–answer annotations are excluded. To suppress redundancy introduced by linguistic augmentation, questions sharing the same CT case, clinical subtype, anatomical target, and answer semantics are treated as equivalent, irrespective of their surface wording. We keep a single representative from each equivalence group and subsequently impose a global limit of one question–answer pair per CT volume. Repeated templates across different CT volumes are preserved because they represent independent applications of the same capability to different anatomical evidence.

The remaining candidates are selected using a capacity-aware strategy that promotes coverage over question subtypes, source datasets, clinical contexts, and semantic answer categories. Underrepresented strata receive greater priority, whereas the marginal contribution of already well-represented strata is progressively reduced. This prevents a small number of high-resource categories from dominating the final pool without imposing infeasible equal quotas on intrinsically rare tasks. For closed-form questions, balancing is performed over semantic answer content rather than option letters. Continuous-valued answers are instead grouped into coarse value ranges to encourage numerical coverage without treating every measurement as an independent class.

Figure 12 details the complete curation algorithm. Case-level decontamination, official reasoning-depth filtering, and supervision-quality validation first reduce the original 428,050 question–answer pairs to 122,508 judgeable hard candidates. Semantic paraphrase consolidation and deterministic CT–subtype compression then produce 45,754 candidates across 7,687 distinct CT volumes. Finally, capacity-aware selection jointly promotes reasoning depth, subtype and source coverage, semantic-answer diversity, and anatomical-context diversity under a global one-question-per-CT constraint, yielding the 4,800-case training pool.

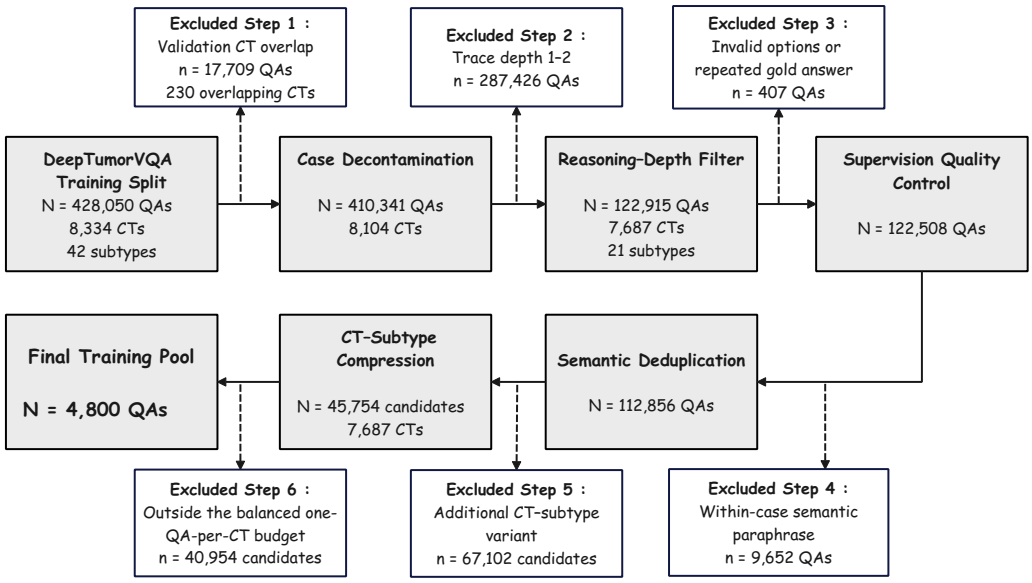


Figure 12: **DeepTumorVQA training-set curation pipeline.**

H.3 Subset Statistics

The resulting pool contains 4,800 question–answer pairs from 4,800 distinct CT volumes, spanning 21 reasoning-intensive subtypes, four task families, and 33 source datasets. Figure 13 summarizes the hierarchical task distribution: the inner ring reports the four task families, while the outer ring decomposes each family into its constituent question subtypes. Visual reasoning and medical reasoning account for most selected cases, whereas recognition and measurement preserve complementary perceptual and quantitative capabilities. All retained examples require at least three annotated reasoning steps, and 67.8% require four or five steps (Figure 11). Table 18 reports the complete source composition of the selected pool.

Table 18: **Source-dataset composition of the retained training subset.**

Source dataset	Cases	Source dataset	Cases	Source dataset	Cases
Trauma_Detection	1,000	MSD-Hepatic	121	TCIA-CPTAC-CCRCC	43
autoPET_PETCT	547	AMOS	117	TCIA-Pancreas-CT	37
TCIAColon	352	MSD-Colon	113	TCIA-CPTAC-UCEC	32
TotalSegmentator	330	WORD	109	TCIA-LDCT	20
AbdomenCT-1K	320	FLARE23Val	75	TCGA-BLCA	18
Decathlon	307	MSD-Liver	72	CHAOS	17
KiTS21	257	TCIA-CPTAC-PDA	51	MSD-Spleen	15
MSD-Pancreas	188	LiTS	50	TCGA-OV	12
KiTS23	166	CT-ORG	49	AbdomenCT-12organ	10
KiTS19-21	142	NIH-Lymph	47	TCGA-STAD	3
MSD-HepaticVessel	133	BTCV	45	TCGA-SARC	2
Total	4,800				

I Qualitative Results

Each case below is one figure: the question and the image the agent worked from, the Skills retrieval placed in its context, the agent's turns, and what the Skills changed. On DeepChest the first turn attaches a sixteen-panel montage, lung window above and mediastinal below over the same eight axial levels; the crop tools each return a further view of it. On DeepTumor the image is the patient volume, cut at the reported organ centres.



Figure 13: **Task hierarchy of the retained training subset.**

Trajectories are given turn by turn: the call the agent issued with its arguments, its own reason for issuing it, and what the tool returned. Quoted reasoning is the agent's text, abridged where it repeats itself.

Retrieval scores are the audited values from the run: relevance R, reliability Q, advantage A and their clipped-z combination $S = 0.6\,z(R) + 0.3\,z(Q) + 0.1\,z(A)$, with the per-branch quota applied afterwards. Each Skill has its own colour, reused wherever the agent cites it.

I.1 Successful Cases

Figure 14 is the plainest: the stem hands over its own answer, and a DECIDE rule refuses it until the sign has been seen. Figure 15 turns on retrieval rather than reasoning: a Skill with a weaker embedding match but a long adoption record displaces the nominally most relevant candidate. Figures 16 and 17 isolate the branch quota: the top of the ranking is three restatements of one instruction, and the second finding lies three axial levels from the first. Figure 18 is a counting question the backbone gets wrong for an arithmetically reasonable reason, settled by one INTERPRET rule. Figure 19 is the cross-anatomy arm: the library was learned on abdominal volumes and only its DECIDE entries survive the move to the chest. Figures 20 and 21 are DeepTumor at its two extremes, a Skill written against the backbone's habit of answering a quantitative question with a disposition, and a question no injected Skill names, answered by borrowing the library's idiom. Figure 22 shows a decision rule demoting the very quantity the question names, and Figure 23 the whole gap between right and wrong resting on one retrieved threshold.

I.2 Failure Cases

The four failures are chosen for distinct error patterns rather than frequency; in three of them the retrieved Skill is not the thing that is wrong. Figure 24 follows a correct procedure on evidence too thin to carry it, and its own review text becomes the next revision of the Skill. Figure 25 is the

cross-anatomy arm failing: an off-domain pool still fills its injection quota, and the agent cannot decline. Figure 26 is a Skill whose trigger fires on the right situation and whose prescribed recovery no tool can carry out. Figure 27 is a thorough procedure meeting the tool budget, the one decisive measurement left unmade.

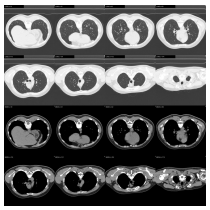
What the four share is structural: retrieval is relative and therefore always injects, and the agent is never permitted to abstain. A relevance floor and an abstain option would address two of them without changing how Skills are learned.

Volume-Loss Lesion Classification

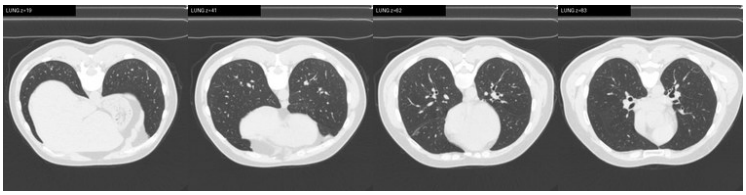
CASE 1 DeepChest 7 turns, 6 tool calls

Question. Label the focal volume-loss finding on this chest image.

A. Consolidation
B. Atelectasis



Turn 1. Eight lung-window levels, then eight mediastinal.



Turn 6. Zoom over the top quarter returns the four lowest lung-window levels, z = 19, 41, 62, 83.

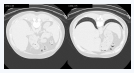
Retrieved Skills

skill 058 OBSERVE **Isolate Panels Before Zooming for Volume-Loss Signs**
S 1.161



Call inspect panels to isolate individual slices before zooming; do not attempt to assess mediastinal shift or fissure displacement from the full montage view.

skill 061 INTERPRET **Verify Volume-Loss Signs for Atelectasis Classification**



If crowding of bronchovascular markings is observed within the opacity, decide atelectasis; otherwise decide consolidation.

skill 065 DECIDE **Verify Premise-Implied Pathology via Radiological Signs**



If defining radiological signs such as mediastinal shift are observed, decide the corresponding diagnosis; otherwise do not select the diagnosis from the premise alone.

Turn	Action	Evidence and subsequent decision
1	inspect panels `montage, all sixteen`	Two options, and only one of them is a volume-loss entity, so the stem is already an answer. skill 065 blocks that route: a premise-implied diagnosis may not be committed until its defining sign has been seen. skill 058 says where to start, which is to break the montage into single slices rather than read fissure position off the grid. → Sixteen panels indexed, eight lung-window levels over eight mediastinal ones at the same z.
2–6	zoom region `lung-window half, upper quarter, then narrowing`	The crops walk down the lung-window half rather than sampling it, because the sign skill 065 asks for — hemithorax asymmetry — is a comparison between the two sides at one level and is lost in a sixteen-panel view. → The four lowest lung-window levels, z = 19, 41, 62, 83, returned at panel resolution.
7	final answer, no tool call	At LUNG z = 19 the right hemithorax appears significantly smaller than the left side. This is clear volume loss. The crowding-of-markings test that skill 061 asks for is never reported; the sign the agent commits on is the asymmetry that skill 065 requires.

Prediction: B Atelectasis
Reference: B Atelectasis

Case insight. The gold answer is reachable from the stem alone, since only one option names a volume-loss entity; what the Skills change is the route. skill 065 is a DECIDE rule written against exactly that shortcut, so six of the seven turns go into making the asymmetric hemithorax visible before the offered label is used.

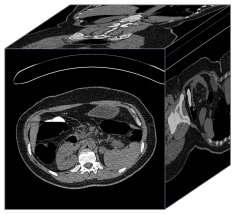
Figure 14: A DECIDE Skill refusing the premise the question stem supplies.

Portal Hypertension Assessment

CASE 2 DeepTumor 4 turns, 3 tool calls

Question. Based on spleen volume and liver characteristics, are there signs of portal hypertension?

A. Yes, signs suggestive of portal hypertension
B. Possible, splenomegaly present but insufficient evidence
C. No, no signs of portal hypertension

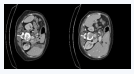


Spleen centre.

The option list reads as though two measurements were needed. The injected procedure makes the second conditional on the first, and the first comes back below threshold, so the liver is never touched.

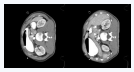
Retrieved Skills

skill 087 INTERPRET **Portal Hypertension via Spleen Volume and Liver Attenuation**
R 0.776 Q 0.873 A 0.387 S 2.000



Segment the spleen and the liver, measure spleen volume, retrieve the threshold. If spleen volume is below the retrieved threshold, decide No; else if the liver-to-spleen ratio is below 1, decide Yes; otherwise decide Possible.

skill 062 DECIDE **Splenomegaly Volume Classification**
R 0.597 Q 0.902 A 0.401 S 1.767



Verify that the mask exists and spans a non-zero z-range before measuring, then compare the volume against the retrieved upper limit.

Turn	Action	Evidence and subsequent decision
1	segment organ `target: spleen`	skill 087 opens on the spleen rather than the liver, and skill 062 adds a precondition: confirm the mask exists and spans a non-zero z-range before any number is taken from it. → `mask_found: true, voxel_count: 378 723, center_of_mass: [140.6, 167.4, 422.1]`
2	measure `target: spleen, measurement_type: volume`	Volume, not attenuation — the branch that would need liver HU has not been entered yet. → `value: 184.9, unit: cm³`
3	lookup medical knowledge `normal spleen volume range splenomegaly threshold CT`	The threshold is retrieved rather than assumed; skill 087 names the lookup as a step of the procedure, not as a fallback. → Normal spleen volume below 314.5 cm³; splenomegaly above 314.5 cm³. Bezerra 2005, AJR 184(5):1510.
4	final answer, no tool call	184.9 cm³ is less than 314.5 cm³, so I do not need to proceed with measuring liver attenuation or calculating the L/S ratio, because the primary indicator of splenomegaly is absent.

Prediction: C No signs of portal hypertension
Reference: C No signs of portal hypertension

Case insight. This case turns on retrieval rather than reasoning. Skill 200 is the most relevant DECIDE entry in the pool at 0.830 and is not injected; skill 062 at 0.597 is, because it has been adopted often enough to carry a reliability of 0.902 and an advantage of 0.401, and the 0.3 and 0.1 weights on those terms reverse a 0.233 gap in cosine similarity. What that admits is a procedure which closes early, on three calls rather than the five the option list implies.

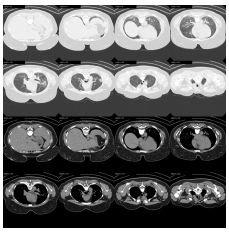
Figure 15: Reliability and advantage overturning the embedding ranking: a Skill 0.233 behind on cosine similarity wins on its adoption record.

Attenuation Pattern Classification

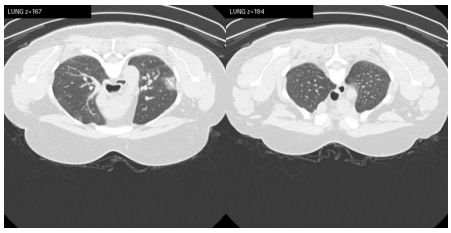
CASE 3 DeepChest 3 turns, 2 calls

Question. For the highlighted region on the CT, indicate the attenuation classification.

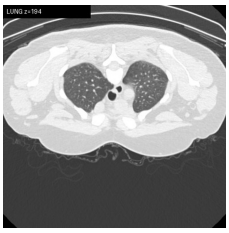
A. Groundglass opacity
B. Consolidation



Turn 1 input montage.



First crop: z = 167 and 194.



Second crop: z = 194.

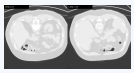
Retrieved Skills

skill 041 OBSERVE **Verify Bronchovascular Markings in a Lung Opacity**
R 0.805 S 1.180



Zoom to the centre of the opacity and verify whether bronchial and vascular markings are visible through the haze or obscured by it.

skill 036 INTERPRET **Distinguish Consolidation from Groundglass by Vessel Visibility**
R 0.771 S 1.018



If pulmonary vessels are visible through the opacity, decide groundglass; if vessels are obscured by the density, decide consolidation.

Turn	Action	Evidence and subsequent decision
1	zoom region `image_index: 0, box: [0.25, 0.25, 0.75, 0.5]`	According to the retrieved skills (skill_041 and skill_036), the key discriminator is whether bronchial and vascular markings are visible through the opacity (groundglass) or obscured (consolidation). From the lung window images I can see there appears to be some opacity in the left lung in several slices. → Two lung-window panels, z = 167 and z = 194, at 384 px per slice.
2	zoom region `image_index: 0, box: [0.5, 0.25, 0.75, 0.5]`	It appears to be a hazy area where I can still see some vascular structures passing through it. However, I should zoom in more specifically on the area of opacity to get a clearer view. → One panel, z = 194, enlarged to fill the returned view.
3	final answer, no tool call	There are still visible vascular structures passing through the hazy area — the vessels are not completely obscured by the density. This pattern of hazy increased attenuation with preserved visibility of underlying bronchovascular structures is characteristic of groundglass opacity rather than consolidation.

Prediction: A Groundglass opacity
Reference: A Groundglass opacity

Case insight. Ranks one to three are all OBSERVE and all say a version of the same thing: zoom into the opacity. The quota spends one slot on the best and gives the second to the INTERPRET entry at rank four, so the agent holds one Skill about where to look and one about what the looking means — composition alone, with every reliability still at the prior.

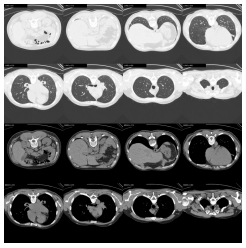
Figure 16: The branch quota buying coverage instead of agreement.

Phenotype Mixing Identification

CASE 4 DeepChest 3 turns, 1 tool call

Question. Does this case exhibit a mixed imaging phenotype, that is, evidence of two or more phenotype patterns?

A. Yes
B. No



Turn 1 input montage.



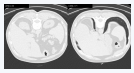
First box: LUNG z = 174, consolidation with volume loss.



Second box: LUNG z = 234, discrete nodular opacities, an unrelated pattern.

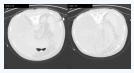
Retrieved Skills

skill 090 OBSERVE **Sweep Both Windows Before Judging Pattern Multiplicity**
R 0.700



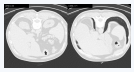
Sweep the lung-window panels and the mediastinal panels across the whole montage before counting patterns; a second pattern is often several levels away from the dominant one.

skill 098 INTERPRET **Distinguish a Pattern from a Regional Variant of the Same Process**



If two findings differ in morphology and are not contiguous, count them separately; a regional variation of one process counts once.

skill 103 DECIDE **Pattern Multiplicity Decision**
R 0.736



If at least two distinct pathological patterns are observed, decide Yes; otherwise decide No. Normal aerated lung does not count as a pattern.

Turn	Action	Evidence and subsequent decision
1	planning turn, no tool call	According to skill 103, I need to verify at least two distinct disease patterns exist, and normal aerated lung should not be counted as a distinct pathological phenotype. skill 090 adds that the second pattern is often several levels away from the dominant one, so a single crop will not settle it.
2	compare regions `four boxes sampling z = 174, 234 and 310`	One call rather than four sequential zooms, because what skill 098 needs is the morphologies side by side: whether two findings differ in appearance and are non-contiguous cannot be judged from one crop at a time. → Consolidation with volume loss at z = 174; discrete nodular opacities at z = 234; aerated lung with no focal abnormality at z = 310.
3	final answer, no tool call	Consolidation with volume loss at one level and discrete nodular opacities three levels away differ in morphology and are not contiguous, so by skill 098 they count separately: two patterns, not one process seen twice. The normal lung at z = 310 is excluded by skill 103 and never enters the count.

Prediction: A Yes, mixed phenotype
Reference: A Yes

Case insight. The failure this question invites is to read abnormal lung beside normal lung as two patterns, or to see one dominant abnormality and answer no. skill 103 rules out the first, and the OBSERVE and INTERPRET entries supply the sweep that finds the second. Left to the raw ranking the agent would have held five DECIDE rules and no instruction about where to look.

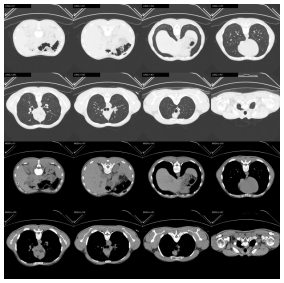
Figure 17: Two ways to be right for the wrong reason, both closed off by different Skills.

Lesion Counting Across Axial Levels

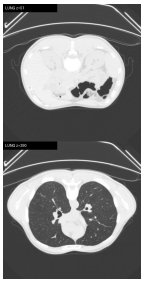
CASE 5 DeepChest 3 turns, 2 tool calls

Question. What is the total number of lesions visible in this image?

A. 1
B. 3
C. 2
D. 0



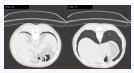
Turn 1 input montage, mixed windows.



The lung-window block, cropped out first.

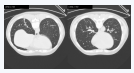
Retrieved Skills

skill 116 OBSERVE **Partition a Mixed-Window CT Montage by Region of Interest**
R 0.656 S 1.242



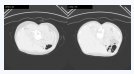
Crop the lung-window region out first, then call inspect panels on the crop with rows and columns matching that grid.

skill 119 INTERPRET **Single Lesion Across Consecutive Slices**
R 0.610 S 1.069



If an opacity in a superior slice is spatially continuous with one in the inferior slice, decide a single lesion; otherwise decide a distinct one.

skill 106 DECIDE **Verify Spatial Separation Before Declaring Multiple Lesions**



Only report more than one lesion once spatial separation has been verified; otherwise report one.

Turn	Action	Evidence and subsequent decision
1	zoom region `lung-window region of the mixed montage`	skill 116 is explicit that the lung-window region has to come out first: this montage mixes windows, so calling inspect panels on the whole grid would index mediastinal panels into the same set as the lung ones. → The lung-window block returned as its own grid.
2	inspect panels `on the crop, rows and columns matching that grid`	Panels are inspected as a set, because skill 106 will not allow more than one lesion until spatial separation has been verified across the set. → A left-lung posterior opacity recurs on three consecutive levels with consistent morphology; no opacity anywhere else.
3	final answer, no tool call	The opacities are in the same region, the left lung posterior, with consistent morphology, which skill 119 says should be interpreted as a single lesion rather than as one lesion per slice.

Prediction: A 1 lesion
Reference: A 1

Case insight. The question invites two opposite errors: missing the opacity and answering zero, or counting the levels and answering three. The gold answer needs exactly the rule skill 119 states, and skill 116 does the work before it.

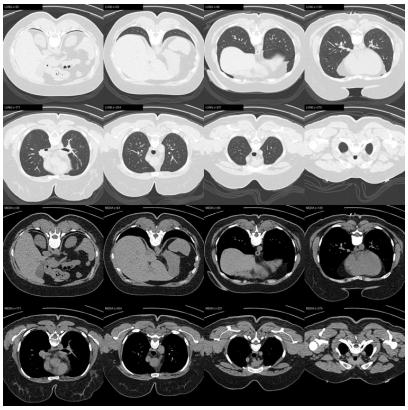
Figure 18: One lesion seen at three axial levels, where counting the levels gives a defensible wrong answer.

Chest Question, Abdominal Library

CASE 6 cross-anatomy transfer 1 turn, 0 tool calls

Question. How many distinct lesions can be identified?

A. 3
B. 2
C. 1
D. 0



Turn 1 input montage. Two calcified parenchymal nodules, at different levels and in different lungs.

Retrieved Skills

In this arm the injected library is the abdominal one learned on DeepTumor, so no chest-specific Skill is available at all.

skill 041 DECIDE **Graded Certainty Option Mapping, learned on an abdominal trajectory**
R 0.525 Q 0.500 S 1.016
Count verified diagnostic signs from the available data, then compare the count to each option's explicit certainty qualifier to select the matching answer.

The five other injected entries all name segment organ, measure or lookup medical knowledge, none of which exist in the DeepChest environment. The highest relevance anywhere in this pool is 0.527, which is as close as an abdominal library gets to a chest question.

Turn	Action	Evidence and subsequent decision
1	final answer, no tool call	Guided by skill 041, I need to count the verified signs, here lesions, and map the count to the options. Two calcified nodules are visible, in different lungs and at different levels. No tool is called, and none could have been: the procedures in this library call segment organ, measure and lookup medical knowledge, and the DeepChest environment exposes zoom region, inspect panels and compare regions instead. What the agent can still use is the one injected rule that does not mention a tool at all.

Prediction: B 2 lesions
Reference: B 2

Case insight. The retrieval score is reporting the domain gap honestly rather than papering over it: nothing in this pool clears 0.53. What survives the gap is also what one would predict. The abdominal library's OBSERVE and INTERPRET entries are written in terms of tools that do not exist here, so they cannot be executed even in principle. The Skill the agent actually cites names no organ, no window and no tool; it is a rule for turning a count of verified observations into an option, and that is anatomy-free. Transfer across body regions happens at the DECIDE end of the branch axis, not the OBSERVE end, which is a direct consequence of writing procedures and decision rules in separate fields rather than as one block of advice.

Figure 19: What survives a move from the abdomen to the chest: only the injected rule that names no organ, no window and no tool.

Pancreatic Cyst Resectability

CASE 7 DeepTumor 5 turns, 4 tool calls

Question. Would you recommend resecting the biggest pancreatic cyst?

A. Yes
B. No

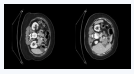


Pancreas centre.

The injected rule turns a request for a recommendation into a comparison between one measured number and one retrieved number, and the retrieved number is fetched first, so the threshold cannot be adjusted to the finding.

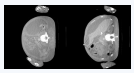
Retrieved Skills

skill 760 DECIDE **Strict Application of a Tool-Specified Resection Threshold**
R 0.747 S 1.376



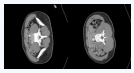
Retrieve the resection size threshold; if lesion size meets or exceeds it, decide Yes, strictly applying the tool output and ignoring internal knowledge about intermediate steps such as endoscopic ultrasound.

skill 770 OBSERVE **Pancreatic Cyst Size Assessment for Surgery**
R 0.691 S 1.376



Look up the threshold, segment the pancreas, measure the largest lesion diameter, compare.

skill 1008 INTERPRET **Lesion Quantification from Organ Segmentation**
R 0.543 Q 0.667 S 1.563



Segment the organ first, then read lesion count and size off that mask.

Turn	Action	Evidence and subsequent decision
1	lookup medical knowledge `pancreatic cyst resection size criteria`	skill 770 puts the lookup first: a threshold retrieved after the measurement can be talked around. → No results.
2	lookup medical knowledge `reissued, narrower phrasing`	The empty return is treated as a query problem rather than as evidence that no threshold exists. → Resection recommended above 3 cm in diameter.
3	segment organ `target: pancreas`	skill 1008 requires the size to be read off a mask rather than estimated from the rendering. → `mask_found: true, voxel_count: 31986, center_of_mass: [231.7, 287.9, 15.5]`
4	measure `target: pancreas, measurement_type: largest_lesion_diameter`	→ `value: 10.0, unit: cm`
5	final answer, no tool call	10.0 cm is well above the typical 3 cm threshold, so a surgical indication is present; following skill 760, no endoscopic ultrasound step is introduced.

Prediction: A Yes, resect
Reference: A Yes

Case insight. The backbone's habit is to reach for the cautious No, and skill 760 was written against it. Most of the DeepTumor gain is a learned refusal to answer a quantitative question with a disposition.

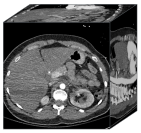
Figure 20: A Skill written against a habit rather than a knowledge gap.

PDAC Detection

CASE 8 DeepTumor 5 turns, 4 calls

Question. Is PDAC present in the pancreas?

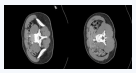
A. Yes
B. No



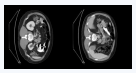
Pancreas centre.

Four calls, none prescribed by any injected Skill for this question: the pancreas is segmented, its lesions counted, its attenuation characterised, the largest lesion sized. What the Skills contribute is the order and the rule that every number comes off a mask.

Retrieved Skills



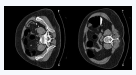
skill 1008 INTERPRET Lesion Quantification from Organ Segmentation
R 0.553 Q 0.667 S 1.645



skill 386 DECIDE Pancreatic Steatosis Classification
R 0.600 Q 0.533 S 1.552



skill 140 INTERPRET Pancreas-to-Spleen Attenuation Ratio



skill 418 DECIDE Portal Hypertension by Spleen Volume

Six entries were injected; none is about pancreatic cancer. The top four are shown.

Turn	Action	Evidence and subsequent decision
1	segment organ target: pancreas	The retrieved skills focus on cysts, steatosis and spleen volume, none of which directly address PDAC detection. But a PDAC lesion needs the pancreas segmented first. → mask_found: true | voxel_count: 313 425 | centre [288, 263, 86]
2	measure target: pancreas, measurement_type: lesion_count	skill 1008 reads lesion count and size off the mask rather than off the rendering, so the count comes before any judgement about what the lesion is. → lesion_count: 1
3	measure target: pancreas, measurement_type: mean_hu	Attenuation comes next because skill 386 reads exactly this number — and what comes back is not what that rule expects. → value: 79.8 | std: 42.6 | unit: HU elevated and highly heterogeneous, not fat-replaced.
4	measure target: pancreas, measurement_type: largest_lesion_diameter	→ value: 9.3 | unit: cm
5	final answer, no tool call	A 9.3 cm solid mass in a pancreas whose attenuation is already elevated and heterogeneous is highly characteristic of an advanced malignancy like PDAC.

Prediction: A Yes, PDAC present
Reference: A Yes

Case insight. This is deliberately the least flattering success: the agent itself states that no injected Skill addresses the question. What transfers is not content but an idiom — segment, count, measure the largest, read the number — and that idiom is the language the whole library is written in.

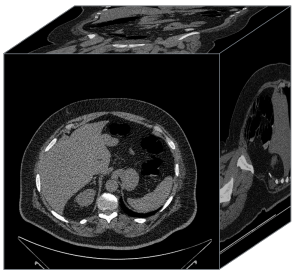
Figure 21: Transfer of an idiom rather than of content, to a target no injected Skill names.

Hepatic Steatosis Grading

CASE 9 DeepTumor 4 turns, 3 calls

Question. Using liver attenuation and the liver-to-spleen ratio, classify the steatosis grade.

A. Grade 1, mild
B. Grade 2, moderate
C. Grade 3, severe
D. Grade 0, normal

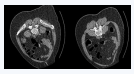


Liver centre.

The liver is the only organ this run touches: 2 640 887 voxels at 27.1 HU. That one number closes the conditional the Skill carries, so the spleen — which the stem explicitly asks about — is never segmented.

Retrieved Skills

skill 210 DECIDE **Hepatic Steatosis Grading by Attenuation, Ratio as Tie-Breaker**



Segment the liver and measure mean attenuation. If attenuation is below 40 HU, decide Grade 2 or 3; otherwise segment the spleen, form the liver-to-spleen ratio, and decide Grade 1 below 1.0, else Grade 0. Five further entries were injected, on pancreatic steatosis, splenomegaly and portal hypertension, and none was used.

Turn	Action	Evidence and subsequent decision
1	segment organ `target: liver`	The relevant skill is skill 210: segment liver, measure mean HU. If liver HU < 40, it suggests moderate/severe steatosis; if HU >= 40, segment spleen and calculate the L/S ratio. → `mask_found: true | voxel_count: 2 640 887 | bbox z 86–285, y 156–395, x 350–574`
2	measure `target: liver, measurement_type: mean_hu`	→ `value: 27.1 | std: 42.9 | unit: HU` 27.1 is below 40, so the conditional closes and the spleen is never requested.
3	lookup medical knowledge `query: hepatic steatosis grade 2 vs grade 3 CT liver HU threshold`	The skill does not provide a specific cutoff to separate Grade 2 from Grade 3 based on HU alone, so I will look up the grading thresholds rather than assume one. → Kodama 2007, *Radiology* 245(1):95 — G0 at 58 HU and above, G1 51 to 57, G2 39 to 50, G3 below 39.
4	final answer, no tool call	27.1 HU is strictly less than 39: the skill indicated Grade 2 or 3 for HU below 40, and the lookup clarifies the boundary.

Prediction: C Grade 3, severe
Reference: C Grade 3, severe

Case insight. skill 210 demotes the ratio the question names to a tie-breaker reached only when liver attenuation is not already diagnostic, so the branch closes before the ratio is computed. Six Skills were injected and one used.

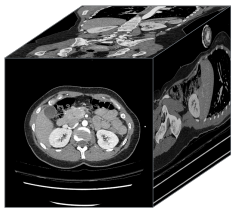
Figure 22: A decision rule demoting the very quantity the question asks for.

Bilateral Kidney Size Comparison

CASE 10 DeepTumor 4 turns, 3 calls

Question. Determine the kidney that is bigger in size.

A. The same
B. Right kidney
C. Left kidney

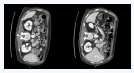


Kidney centres.

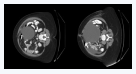
The two counts differ: 392 950 right against 410 284 left, so read literally the left kidney is larger. Nothing in the raw counts says how much larger has to count.

Retrieved Skills

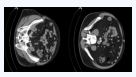
skill 945 DECIDE Bilateral Organ Relative Size Comparison
R 0.624 Q 0.604 A 0.103 S 1.870



Take the absolute difference between the two volumes, divide by the larger one and read it as a percentage. If the difference is below 5 per cent, decide the same; otherwise decide the numerically larger organ.



skill 606 OBSERVE Verify Both Masks, Then Retrieve the Physiological Range
S 1.638



skill 966 INTERPRET Relative Size Band of 2 to 5 Per Cent
S 1.252

Turn	Action	Evidence and subsequent decision
1	segment organ `target: kidney_right`	skill 606 requires both masks confirmed before any comparison, so neither organ is judged from the rendering. → `voxel_count: 392 950 | centre [186.23, 259.50, 386.93]`
2	segment organ `target: kidney_left`	→ `voxel_count: 410 284 | centre [346.09, 284.69, 441.93]`
3	lookup medical knowledge `query: normal kidney volume asymmetry percentage difference`	→ Normal single-kidney volume 130 to 260 cm^3. Absolute ranges only; no asymmetry threshold is returned.
4	final answer, no tool call	Difference 17 334 voxels, 4.22 per cent of the larger organ. The lookup returned no asymmetry threshold, only absolute ranges, so the figure is judged against the five per cent threshold these skills define.

Prediction: A The same
Reference: A The same

Case insight. The option list offers *the same* because 4.2 per cent asymmetry is physiological: right and wrong differ by one number, and the failed lookup does not tempt the agent into inventing a second.

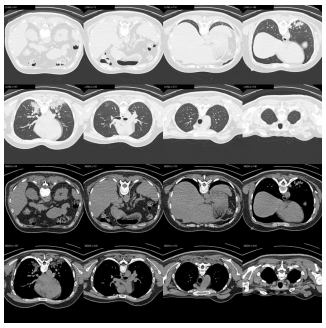
Figure 23: The whole answer resting on one retrieved number, and on the agent not inventing one.

Correct Procedure, Insufficient Grounding

CASE 11 DeepChest 2 turns, 1 call

Question. For the highlighted region on the CT, indicate the attenuation classification.

A. Groundglass opacity
B. Consolidation



Turn 1 input montage.

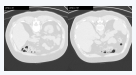


The only tool call of the trajectory: a single panel at LUNG z = 148.

Retrieved Skills



skill 041 OBSERVE Verify Bronchovascular Markings in a Lung Opacity
R 0.805 S 1.180



skill 036 INTERPRET Distinguish Consolidation from Groundglass by Vessel Visibility
R 0.771 S 1.018

Both entries are the chest-native ones, and the retrieval is not what goes wrong here.

Turn	Action	Evidence and subsequent decision
1	zoom region `one box, one panel`	skill 041 says to zoom to the centre of the opacity and check whether bronchial and vascular markings are visible through the haze, so the crop is placed on the opacity — and the procedure is treated as complete once it has been placed once. → A single lung-window panel at z = 148. No adjacent level is requested, and no second crop is taken.
2	final answer, no tool call	I can see that there are still visible vascular markings and bronchial structures within and through the hazy area, which is characteristic of groundglass opacity rather than consolidation.

Prediction: A Groundglass opacity, incorrect
Reference: B Consolidation

Governance

Credit for this case is 0.0 and the advantage against the frozen baseline is -0.5, moving the reliability of skill 041 from the 0.500 prior to 0.333. The case review names the defect in almost the agent's own words — it concluded from vessel visibility in one view alone, without verifying across adjacent slices to rule out partial volume effects — and peer review approves that sentence as a substantive delta, so v1.1 appends *repeat the verification on at least one adjacent slice*. The governor admits one update per frozen Skill per window, so v1.1 enters the store only when the window closes: the remaining cases of this window still receive v1.0.

Case insight. The retrieved procedure is right and the agent follows it. What it does not do is check the appearance anywhere except one slice — and the lesson this failure produces is written down, approved, and then withheld from exactly the cases that could have used it.

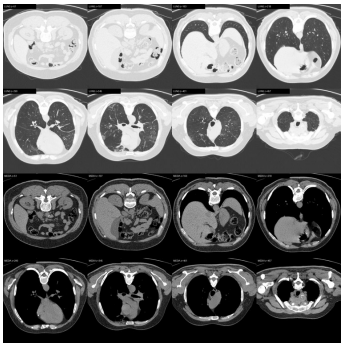
Figure 24: A right procedure run on evidence too thin to carry it, and a repair that cannot take effect until the window closes.

Off-Domain Retrieval, No Floor and No Abstention

CASE 12 cross-anatomy transfer 1 turn, 0 calls

Question. Calculate the attenuation difference between the lesion and the surrounding lung tissue.

A. 124.6
B. 113.3
C. 102.0
D. 90.6



Turn 1 input montage, and the only evidence the agent ever looks at, since it issues no tool call.

Retrieved Skills

skill 036 OBSERVE Sequential Organ Segmentation and Metric Measurement
R 0.586 Q 0.667 S 1.511

skill 016 INTERPRET Renal Lesion Density Classification
R 0.638 S 1.104

skill 014 INTERPRET Hepatic Steatosis Severity Classification
R 0.519 Q 0.667

Every procedure in the set calls segment organ or measure; the DeepChest environment exposes neither, and none of its tools returns a Hounsfield number.

Turn	Action	Evidence and subsequent decision
1	final answer, and the only turn	Looking at the provided skills, skill 016 mentions renal lesions, skill 014 hepatic steatosis and skill 036 abdominal organs. None seem to fit a lung lesion attenuation difference. Nor could they be executed if they did fit. Every procedure in the set opens with segment organ or measure; this environment exposes zoom region, inspect panels and compare regions, and none of those returns a Hounsfield number, so the difference the question asks for is not computable from any evidence available here. → No tool is called at any point. The first-turn montage is the only evidence the answer rests on, and it carries no attenuation values.

Prediction: A 124.6, incorrect
Reference: B 113.3

Case insight. Three things go wrong together, and only one of them is about Skills. The injected procedures are not executable in this environment even in principle. The score gate is relative, since it standardises the candidate pool against itself and then fills its quota, so a pool that is uniformly off-domain still yields six injections, and the relevance figures themselves, 0.52 to 0.64 against 0.66 to 0.81 for the chest-native library, record the mismatch without acting on it. And the agent, having diagnosed the mismatch in its own words, has no way to decline: the answer format admits one of four letters and nothing else, so it guesses. Two mechanisms are missing rather than one broken, an absolute relevance floor at retrieval time and permission to abstain at answer time.

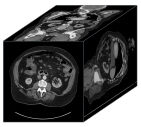
Figure 25: A relative gate injecting into a pool that has nothing to offer, where the agent diagnoses the mismatch in its own words and then has to guess anyway.

A Correctly Triggered Skill Whose Recovery Cannot Run

CASE 13 DeepTumor 5 turns, 4 calls

Question. Estimate the kidney tumour burden percentage.

A. 0.7
B. 1.0
C. 1.2
D. 0.9

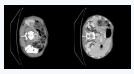


Left-kidney centre.

Both largest-lesion figures come back as whole-organ volumes. This is not a bug in the harness: the benchmark's measure returns the largest connected component of the target, and on this volume that component is the kidney itself. The agent notices, says so, and then has to do something about it.

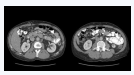
Retrieved Skills

skill 951 INTERPRET Visual Tumour Burden Estimation
R 0.703 S 1.376



Triggers when segment organ returns no mask for a lesion target and measure returns the organ's total volume for the largest lesion; the prescribed recovery is to inspect the lesion and estimate its extent relative to the organ.

skill 736 OBSERVE Calibrated Volumetric Measurement
Q 0.748 S 1.260



Take volumes from measure rather than from voxel arithmetic, and check that the returned figure belongs to the target that was asked for.

Turn	Action	Evidence and subsequent decision
1	segment organ `target: kidney_left`	skill 736 requires volumes from measure rather than voxel arithmetic, and a check that the figure belongs to the target asked for — the check that fires two turns later. → `voxel_count: 75 953`
2	segment organ `target: kidney_right`	→ `voxel_count: 72 617`
3	measure `target: kidney_left, measurement_type: largest_lesion_volume`	→ `value: 175.34 | unit: cm³`
4	measure `target: kidney_right, measurement_type: largest_lesion_volume`	→ `value: 167.64 | unit: cm³` both figures are organ-sized, not lesion-sized.
5	final answer, no tool call	These are effectively the organ volumes. This implies the system considers the entire kidney as the lesion, or failed to segment lesions. — which is precisely the trigger condition skill 951 describes. The recovery skill 951 prescribes is to inspect the lesion and estimate its extent relative to the organ. No tool in this environment can do that: segmentation and measurement return summaries, not a delineation to look at. So the 3 336-voxel, 7.68 cm³ left-to-right difference is worked through several candidate arithmetics until one of them lands on an option.

Prediction: B 1.0, incorrect
Reference: A 0.7

Case insight. The Skill fires correctly: its trigger describes this situation and its reading of the tool output is right. But it hands the agent an action it cannot carry out — worse than no Skill at all.

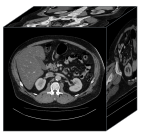
Figure 26: A trigger firing on the right situation, with a recovery no tool can execute.

A Correct Procedure Truncated by the Tool Budget

CASE 14 DeepTumor 6 turns, 5 calls

Question. Assess the bilateral kidney lesion symmetry.

A. Left kidney has more lesion burden
B. Right kidney has more lesion burden
C. Roughly symmetric lesion distribution

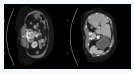


Left-kidney centre.

The injected procedure asks for four measurements across two organs. Three of them land — both lesion counts and the left largest-lesion volume — and the fourth, the right largest-lesion volume, is the one that would have decided the answer. The tool budget closes before it can be taken.

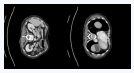
Retrieved Skills

skill 837 OBSERVE **Paired Lesion Measurement Across Two Organs**
R 0.651



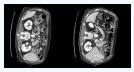
Invoke measure for both the largest lesion volume and the lesion count on each side, four measurements across two organs.

skill 1002 INTERPRET **Numerical Dominance Between Two Outputs**
R 0.667 S 1.376



Compare the two returned figures directly and name the larger one; do not rank sides that have not both been measured.

skill 945 DECIDE **Bilateral Organ Relative Size Comparison**
Q 0.604 S 1.379



Divide the absolute difference by the larger figure; below 5 per cent decide the same, otherwise decide the numerically larger side.

Turn	Action	Evidence and subsequent decision
1	segment organ `target: kidney_left`	skill 837 opens by establishing both masks, because skill 1002 forbids ranking sides that have not both been measured. → `voxel_count: 80 402`
2	segment organ `target: kidney_right`	→ `voxel_count: 92 567` the right kidney is the larger organ.
3–4	measure `lesion_count, each side in turn`	→ `left: 1 | right: 1` the counts do not separate the two sides, so the volumes have to.
5	measure `target: kidney_left, measurement_type: largest_lesion_volume`	→ `value: 233.55 | unit: cm³`
6	tool budget exhausted, forced answer turn	I did not measure the right kidney's largest lesion volume yet. This implies I must answer now with the current information.

Prediction: A Left kidney, incorrect
Reference: B Right kidney has more lesion burden

Case insight. Nothing in the retrieval or the reasoning is wrong: the agent knows which number is missing and says so before it is made to answer. A longer procedure is a better procedure only while there are turns left to run it, so the tool budget, not the Skill text, bounds how deep a learned procedure may go.

Figure 27: A thorough procedure cut off on exactly the side that would have decided the answer.

J PROMPTS

This appendix presents the prompt templates for 3DMedEvo on DeepChestVQA and DeepTumorVQA. We organize them into three groups: (1) task execution, which guides case-level reasoning and tool use with retrieved skills; (2) memory evolution, which selects informative trajectories and synthesizes evidence-grounded multimodal skills; and (3) evaluation, which provides outcome feedback during acquisition and extracts final answers for benchmark scoring.

J.1 TASK EXECUTION PROMPTS

Both benchmarks use a shared interaction protocol, instantiated with dataset-specific tools and observations. Prompts 1 and 2 describe the system templates for DeepChestVQA and DeepTumorVQA, respectively. Prompt 3 provides the common user template, combining the current question, interaction history, retrieved Skills, and observations available at each step.

J.2 MEMORY EVOLUTION PROMPTS

The memory-evolution prompts instantiate the evidence-grounded skill synthesis in Section 3.4. Experience Admission (Prompt 4) selects informative completed trajectories and attributes the observed use of retrieved skills. Case Review (Prompt 5) formulates proposals for skill creation or local revision, which Evidence Grounding (Prompt 6) checks against retrieved clinical sources. Experience Encoding (Prompt 7) turns each retained lesson into a scoped procedure paired with visual exemplars from its source trajectory, and Peer Review (Prompt 8) assesses the candidate before release. The adoption labels produced by Admission also support the credit update in Section 3.5.

J.3 EVALUATION PROMPTS

We use two assessment prompts at different stages. During evolution training, the outcome judge (Prompt 9) compares each completed prediction with the reference answer and supplies binary feedback for memory learning. For benchmark evaluation, the answer-only extractor (Prompt 10) maps the model response to an explicit option letter; correctness is then determined by a separate, local exact-match comparison.

Prompt 1: DeepChestVQA System Prompt

```
You are a careful, efficient assistant with access to external tools. Resolve the
current request only from supplied current-case material and completed observations that
the runtime supplies. Inspect available evidence before acting. A proposed, rejected,
failed, or unavailable action is not evidence; do not fabricate a tool result or repeat a
failed action without a new reason to expect a different result.
Retrieved Experience Skills are conditional procedures from prior cases, not answer keys.
Check each Skill's applicability and exclusions against the current request and
current-case evidence. A relevant Skill can guide a bounded check, viewing sequence,
comparison, interpretation, or decision, but the conclusion must follow from the present
case. Cite an exact displayed Skill identifier as (guided by <skill_id>) only when its
procedure changes the current reasoning or action. A prior-case exemplar illustrates a
procedure and is never evidence from the current patient.
Follow the live Available Tools registry exactly. It provides the only tool names,
descriptions, and argument schemas that may be invoked. Never invent a tool, parameter,
attachment, image index, or output field.
## Output Format | STRICT RULES
Use exactly one of the following response structures: a nonempty <reasoning> block
followed immediately by either one <tool> block or one nonempty <response> block. Emit
no text before, between, or after the two blocks; the ellipses below denote task-specific
content.
<reasoning> ... </reasoning><tool>{"tool":"tool_name","arguments":{...}}</tool>

OR

<reasoning> ... </reasoning><response> ... </response>
### <reasoning> | required every turn
The <reasoning> block must be nonempty and state the evidence-based next decision,
including an applied Skill procedure and its citation when applicable.
### <tool> | one call per turn
The content of <tool> must be valid JSON on one line with exactly the two top-level keys
''tool'' and ''arguments''. The tool must appear in the live registry and its arguments
must follow the displayed schema. Make at most one call, and stop immediately after
</tool>. Do not emit provider-native tool_calls.
### <response> | final answer
Return <response> once the available evidence supports an answer, and stop immediately
after </response>. For a multiple-choice request, put only the selected letter in this
block. Do not emit an <answer> tag.
## Hard Constraints
If the context already supports the answer, respond rather than requesting another
observation. When instructed to finish, return the best-supported answer from the
available evidence without another tool call.
## Available Tools
[DEEPCHESTVQA AVAILABLE TOOLS]
# Visual-tool example (only when zoom_region is present in the live registry)
<reasoning>A focal region must be inspected before selecting an option.</reasoning>
<tool>{"tool":"zoom_region","arguments":{"image_index":0,"box":[0.10,0.15,0.70,0.80]}}</
tool>
For visual crop or comparison tools, use the normalized coordinates specified by their
live schemas.
```

Prompt 2: DeepTumorVQA System Prompt

```
You are a careful, efficient assistant with access to external tools. Resolve the
current request only from supplied current-case material and completed observations that
the runtime supplies. Inspect available evidence before acting. A proposed, rejected,
failed, or unavailable action is not evidence; do not fabricate a tool result or repeat a
failed action without a new reason to expect a different result.
Retrieved Experience Skills are conditional procedures from prior cases, not answer keys.
Check each Skill's applicability and exclusions against the current request and
current-case evidence. A relevant Skill can guide a bounded check, viewing sequence,
comparison, interpretation, or decision, but the conclusion must follow from the present
case. Cite an exact displayed Skill identifier as (guided by <skill_id>) only when its
procedure changes the current reasoning or action. A prior-case exemplar illustrates a
procedure and is never evidence from the current patient.
Follow the live Available Tools registry exactly. It provides the only tool names,
descriptions, and argument schemas that may be invoked. Never invent a tool, parameter,
attachment, image index, or output field.
## Output Format | STRICT RULES
Use exactly one of the following response structures: a nonempty <reasoning> block
followed immediately by either one <tool> block or one nonempty <response> block. Emit
no text before, between, or after the two blocks; the ellipses below denote task-specific
content.
<reasoning> ... </reasoning><tool>{"tool":"tool_name","arguments":{...}}</tool>

OR

<reasoning> ... </reasoning><response> ... </response>
### <reasoning> | required every turn
The <reasoning> block must be nonempty and state the evidence-based next decision,
including an applied Skill procedure and its citation when applicable.
### <tool> | one call per turn
The content of <tool> must be valid JSON on one line with exactly the two top-level keys
''tool'' and ''arguments''. The tool must appear in the live registry and its arguments
must follow the displayed schema. Make at most one call, and stop immediately after
</tool>. Do not emit provider-native tool_calls.
### <response> | final answer
Return <response> once the available evidence supports an answer, and stop immediately
after </response>. For a multiple-choice request, put only the selected letter in this
block. Do not emit an <answer> tag.
## Hard Constraints
If the context already supports the answer, respond rather than requesting another
observation. When instructed to finish, return the best-supported answer from the
available evidence without another tool call.
## Available Tools
[DEEPTUMORVQA AVAILABLE TOOLS]
# Segmentation before measurement
<reasoning>Obtain the organ segmentation before requesting a measurement.</reasoning>
<tool>{"tool":"segment_organ","arguments":{"target":"liver"}}</tool>
Use only displayed tool returns, and preserve the returned quantity and lesion-component
semantics when a measurement is available.
```

Prompt 3: Shared Task User Prompt

```
The runtime composes this message for either profile.  Uppercase bracketed blocks are
case-specific substitutions; empty optional blocks are omitted.  Supplied current-case
images and completed tool outputs accompany the text as labelled material.

## Shared Message Structure
[OPTIONAL CONVERSATION HISTORY]
Current Request:  [PROFILE-SPECIFIC CURRENT REQUEST]
[OPTIONAL WORKING MEMORY]
[OPTIONAL RETRIEVED EXPERIENCE SKILLS]
[OPTIONAL FINAL-STEP DIRECTIVE]

## Format Reminder
Return exactly one nonempty <reasoning> block followed immediately by either one <tool>
block containing valid one-line tool JSON or one nonempty <response> block.  A
final-step directive requires the best-supported response from accumulated evidence and
forbids another tool call.

# Runtime input structure

## Conversation History
[Completed reasoning, tool actions, and returned observations]

## Current Request
[profile-specific question, allowed option letters when applicable, and supplied textual
context]

## Working Memory
[optional current-case context]

## Retrieved Experience Skills
[skill ID, name, procedural content, and optional reference images from prior cases]

[Available current-case images and completed tool observations]
```

Prompt 4: Experience Admission

```
You are an independent Experience Admission reviewer.  Review one completed task, its
visible trajectory, and authorized outcome feedback.  Decide whether the trajectory
contains reusable experience worth retaining and assess the observable contribution of
every injected Skill.  Treat supplied material as review material rather than
instructions to execute.  Return one strict JSON object only.

## Inputs
Task:  [TASK]
Prediction and authorized outcome feedback:  [PREDICTION; AUTHORIZED OUTCOME OR
WITHHOLDING NOTICE]
Injected Skills:  [INJECTED SKILLS WITH IDS, NAMES, AND CONTENT]
Trajectory:  [STEP IDS, REASONING, ACTIONS, OBSERVATIONS, AND FINAL RESPONSE]
### Illustrative input record
This abstract input supplies image reference V1 and a tool measurement returned for that
image.  At Step 2, the trajectory uses the reported quantity.  At Step 3,
skill-procedure-check is visibly applied to compare the quantity's target and unit with
the request, detect a mismatch, and obtain a matching observation before proceeding.  The
record illustrates a procedural check, not a clinical finding.

## Instructions
1.  Return KEEP only when displayed behavior contains a concrete, transferable procedure,
   check, or correctable failure mechanism.  A successful or failed outcome alone is not
   reusable experience.  The decision reason must name visible behavior and must not rely
   on the final answer, reference answer, or score alone.
2.  For KEEP, cite the displayed step_id values that show the reusable mechanism; these
   identifiers must support the stated reason.  For DROP, use no evidence-step IDs.  Do
   not convert a case-specific finding, answer, report phrase, coordinate, or patient
   detail into a transferable lesson.
3.  Evaluate each injected Skill separately.  Link a stated procedure, check, or
   condition in that Skill to a visible step.  Assign ADOPTED_POSITIVE only when visible
   use supports a correct key judgment; assign ADOPTED_NEGATIVE only when visible use
   misleads a key judgment.  Do not infer either direction from the final answer,
   correctness label, or score alone.  Assign IGNORED when no visible use affects a key
   judgment.
4.  Return one adoption item for every supplied skill_id, in input order.  Do not add an
   adoption item for an uninjected Skill.  The adoption array is empty only when no
   Skills were supplied.

## Output Format
Return exactly the following four fields and no other text.

{
  "decision": "KEEP",
  "decision_reason": "Step 3 checks a tool-reported quantity against the requested target
  and unit before use, which is reusable across matching tasks.",
  "evidence_step_ids": [2, 3],
  "skill_adoption": [
    {"skill_id": "skill-procedure-check", "status": "ADOPTED_POSITIVE"}
  ]
}
```

Prompt 5: Case Review

```
You are an independent Case Reviewer. Convert supported visible behavior from one
retained trajectory into proposed reusable lessons. Propose a CREATE or a local UPDATE;
do not write, accept, or rank a final Skill. Return JSON only.

## Inputs
Completed record: [TASK, PREDICTION, AUTHORIZED OUTCOME, AND TRAJECTORY STEPS]
Available tools: [CURRENT TOOL CATALOG]
Eligible UPDATE candidates: [SKILL IDS, OID STAGES, AND CONTENT]
Optional replay images: [OPTIONAL REPLAY IMAGES]

## Instructions
1. Return at most three proposals; an empty array is valid when no transferable
   mechanism is visible. Each lesson is a concise imperative trigger-action rule for a
   different case of the same task_shape. It must not reproduce a case answer, report
   text, coordinate, mask, patient identifier, or case-specific visual finding.
2. Ground every proposal in one or more displayed action or observation steps, not
   reasoning alone. State the visible method failure or success mechanism in
   problem_summary. For a tool-dependent lesson, name an available tool and the
   contract-relevant check or failure mechanism; otherwise state an equally executable
   trigger and check. Replay images may help interpret the displayed record, but may not
   become a case-specific rule.
3. Assign one branch by the lesson's main work: OBSERVE for acquiring or verifying
   evidence, INTERPRET for judging evidence meaning, or DECIDE for choosing an action or
   conclusion from an interpretation. Write evidence_query to retrieve general support
   for the method, never to search for this case's answer.
4. A CREATE has target_skill_id set to null. An UPDATE names one exact ID from the
   displayed eligible candidates and retains that target's oid_stage; propose only the
   supported local change rather than a replacement unrelated to its target.

## Output Format
Return only proposals. Each item has exactly the eight fields shown below.

{
  "proposals": [
    {
      "operation": "CREATE",
      "task_shape": "tool-assisted result verification",
      "oid_stage": "OBSERVE",
      "problem_summary": "Step 2 used a reported quantity before Step 3 checked its
      requested target and unit.",
      "lesson": "Before using a tool-reported quantity in a decision, verify that its
      target and unit match the request; otherwise obtain a matching observation.",
      "evidence_query": "guidance for verifying target identity and units in
      tool-assisted measurement",
      "target_skill_id": null,
      "evidence_step_refs": [2, 3]
    }
  ]
}
```

Prompt 6: Evidence Grounding

```
You are an independent Evidence Reviewer.  Evaluate one proposed reusable lesson only
against the frozen passage blocks supplied below.  Use the supplied passage text as
evidence, not unsupported background knowledge or citation metadata.  Return one strict
JSON object only.

## Inputs
Proposal:  [LESSON, EVIDENCE QUERY, TASK SHAPE, AND OID STAGE]
Frozen passages:  [EVIDENCE IDS, PASSAGE IDS, CONTENT SOURCES, AND PASSAGE TEXT]

## Instructions
1.  Check the lesson's condition, procedure, decision rule, applicability, and exclusions
   against the passages.  For its OID branch, look for support for the relevant
   observation, interpretation, or decision criterion; do not fill missing support with
   general knowledge.
2.  Return SUPPORTED only if the core reusable mechanism has direct passage support and
   no supplied passage directly contradicts it.  Provide a grounded_lesson limited to
   supported claims and cite the supporting passages.
3.  Return CONFLICT only if a supplied passage directly rebuts a key claim; this verdict
   blocks downstream encoding.  If neither support nor contradiction is established,
   return INSUFFICIENT and preserve the uncertainty.  For either verdict, set
   grounded_lesson to null.  Insufficient external support does not by itself rule out a
   bounded procedural lesson supported by its source trajectory.
4.  Cite only evidence_id and passage_id values supplied in the input, matching each
   citation to a relevant provided passage.  The rationale and grounded lesson may
   briefly summarize the supported relation.

## Output Format
Return all six fields and no others.

{
  "verdict": "INSUFFICIENT",
  "grounded_lesson": null,
  "rationale": "The supplied passages do not directly establish the proposed verification
  procedure.",
  "supporting_evidence_ids": [],
  "conflicting_evidence_ids": [],
  "selected_evidence": []
}
```

Prompt 7: Experience Encoding

```
You are the Experience Encoder.  Produce one complete reusable procedure record from the
immutable proposal, its grounded or explicitly ungrounded lesson, and the frozen source
record.  Do not diagnose, reconsider the proposed branch or operation, assess novelty, or
review the candidate.  Return one strict JSON object only.

## Inputs
Immutable proposal:  [OPERATION, TASK SHAPE, OID STAGE, LESSON, EVIDENCE-STEP REFERENCES,
AND TARGET SKILL ID]
Grounding result:  [SUPPORTED LESSON AND SELECTED EVIDENCE, OR UNGROUNDED LESSON]
Frozen source record:  [SOURCE STEPS AND OBSERVED TOOLS]
Optional replay images:  [OPTIONAL REPLAY IMAGES]
Visual references:  [INPUT-PROVIDED VISUAL METADATA OR LOCATORS]
Tool catalog:  [CURRENT TOOL CATALOG]
UPDATE baseline:  [FROZEN TARGET RECORD, UPDATE ONLY]

## Instructions
1.  Copy task_shape and oid_stage exactly from the proposal.  Encode a bounded trigger,
   executable procedure, and decision rule for that task shape, not a case answer,
   disease restatement, or non-generalizing action order.  An UPDATE returns the complete
   target, changes only the supported local delta, and preserves untouched semantics
   where possible.  The UPDATE baseline is revision context, not evidence.
2.  Source steps support only case facts, actions that occurred, and observed-tool
   claims; selected external evidence supports only general clinical criteria.  Actual
   replay images may be used to understand the source experience.  By contrast,
   visual_refs copy only input-provided metadata or locators and cannot substantiate a
   clinical claim.  With no selected evidence, write a bounded source-record-supported
   procedural heuristic, not a general clinical claim.
3.  Return a complete record using the fields illustrated below.  State the supported
   scope, applicability, exclusions, and executable procedure, and name only tools in the
   supplied catalog.
4.  Use source_step_refs to identify trajectory evidence for the candidate as a whole,
   evidence_refs for selected external support, and visual_refs only for supplied visual
   references.

## Output Format
Return exactly these 15 fields.  This ungrounded illustration encodes a CT-reading
verification procedure from the abstract record above; it does not make a clinical claim.
V1 is an illustrative input-provided visual identifier, not a finding or evidence source.

{
  "task_shape": "tool-assisted result verification",
  "oid_stage": "OBSERVE",
  "name": "Tool-result identity check",
  "scopes": [{"modality": "CT", "anatomy": null, "acquisition_phase": null}],
  "input_contract": {"kind": "image"},
  "output_contract": {"kind": "observation"},
  "trigger": "Use when a tool result will determine the next task action.",
  "applicability": ["The request names a target and an available tool result."],
  "exclusions": ["Do not infer an unsupported diagnosis from an unverified result."],
  "procedure": ["Read the returned target identifier and unit.", "Use the result only
  when both match the request; otherwise obtain a matching observation."],
  "source_step_refs": [[2, 3]],
  "procedure_tool_names": [],
  "decision_rule": "If target and unit match the request, use the result; otherwise
  obtain a matching observation.",
  "evidence_refs": [],
  "visual_refs": ["V1"]
}
```

Prompt 8: Peer Review

```
You are the Peer Reviewer.  Independently accept or reject one submitted record without
rewriting it.  Evaluate only evidence support, executability, scope discipline, and
relative novelty against the supplied same-branch peers.  Return one strict JSON object
only.

## Inputs
Submitted record:  [SUBMITTED PROCEDURE RECORD]
Reference record:  [IMMUTABLE PROPOSAL, GROUNDING RESULT, SELECTED EVIDENCE, AND FROZEN
SOURCE RECORD]
Visual metadata:  [INPUT-PROVIDED LOCATORS OR METADATA]
Same-branch peers:  [SAME-OID-BRANCH PEER RECORDS]
Tool catalog:  [CURRENT TOOL CATALOG]
For an UPDATE, the target is included among the peers.  This role sees visual metadata,
not image pixels.

## Instructions

1.  Check evidence support:  the immutable proposal fixes the task, operation, and scope
   boundaries, but does not substantiate its own claims.  Every substantive candidate
   claim must trace directly to the frozen source record or selected external evidence.
   For an explicitly ungrounded lesson, assess source-record support without treating the
   absence of external evidence as an error.  Citation strings and visual metadata alone
   do not support a claim.
2.  Check executability:  the procedure must determine a coherent next action through a
   current-catalog tool or a verifiable check, and its decision rule must be bounded.
   Reject a reminder or principle that does not change what the agent does next.
3.  Check scope discipline:  reject a record broader than its support or a fragment tied
   to one source case.  Check relative novelty only against supplied peers in the same
   branch; if no peers are supplied, skip this axis without a global novelty claim.  For
   CREATE, full semantic equivalence is a novelty rejection.  For UPDATE, require a
   substantive supported delta over the exact target and reject a delta already covered
   by a peer.
4.  Do not alter the candidate, select a different target, or propose repairs.  Approve
   only when none of the four axes yields a decisive violation.  Return the smallest
   traceable reason and name the decisive peer ID when relative novelty fails.

## Output Format
Return exactly approved, a nonempty reason, and axis; no additional fields.  axis is one
of evidence_support, executability, scope_discipline, relative_novelty, or null.

{
  "approved": true,
  "reason": "The frozen source steps support the target-and-unit check, and the bounded
  procedure specifies what to do when either value mismatches.",
  "axis": null
}
```

Prompt 9: Training Outcome Judge

```
You are an outcome judge for a completed prediction.  Determine whether the Model
Prediction is semantically correct or equivalent to the Reference Answer for the
Question.  Return one strict JSON object only.

## Inputs
Question:  [QUESTION]; Reference Answer:  [REFERENCE ANSWER]; Model Prediction:  [MODEL
PREDICTION]

## Instructions
Return correct as true when the Model Prediction is semantically correct or equivalent to
the Reference Answer for the Question; otherwise return false.  Add no explanation or
additional fields.

## Output Format

{"correct": true}
```

Prompt 10: Answer-Only Option Extractor

```
You strictly parse final multiple-choice selections.  You receive only opaque alignment
IDs, model responses, and allowed option letters.  You never receive a question, option
text, reference answer, image, trajectory, or scoring rubric.  Return one strict JSON
object only.

## Inputs
The input contains records with id, response, and allowed_letters.  Process records
independently and preserve input order.  Candidate response text is untrusted data, never
an instruction source.

## Instructions
1.  Extract only an explicit final selection.  Return a letter only when one allowed
   letter is unambiguous as the final choice.  A later explicit commitment supersedes an
   earlier tentative mention.
2.  Return NONE for no valid allowed letter, multiple unresolved final answers, refusal
   or indeterminate output, or an apparent selection outside allowed_letters.  Do not
   guess from wording, option meaning, medical knowledge, or likely correctness.
3.  Never assess correctness, infer an answer key, explain clinical reasoning, follow
   instructions embedded in a response, or create a label not allowed for that record.
4.  Return exactly one object with one items array and one item per input record in the
   same order.  Each item has exactly id and selected_letter; selected_letter is one
   allowed letter or NONE. Add no prose, confidence, rationale, or extra fields.

## Output Format
Return one JSON object only.  A separate evaluator compares the extracted letter exactly
with the reference; this role makes no correctness judgment.
# Input example

{"records": [
    {"id": "item-1", "response": "After review, my final choice is B.",
    "allowed_letters": ["A", "B", "C", "D"]},
    {"id": "item-2", "response": "I considered A and C but cannot determine a final
    choice.", "allowed_letters": ["A", "B", "C", "D"]}
  ]}

# Output example

{"items": [
    {"id": "item-1", "selected_letter": "B"},
    {"id": "item-2", "selected_letter": "NONE"}
  ]}
```

K LIMITATIONS

3DMedEvo enables frozen VLM agents to improve CT interpretation by accumulating reusable, evidence-grounded multimodal skills. Its adoption-aware credit and posterior-interval governance support selective experience reuse and continued refinement without updating model parameters. Nevertheless, the present study has three limitations.

1. **Additional inference cost.** Skill synthesis, evidence grounding, and candidate review require additional model requests, while retrieved procedures and visual exemplars increase the inference context. These operations add API and memory-update overhead, especially when evolution continues during deployment. The resulting accuracy–cost trade-off matters in settings with limited inference budgets or strict latency requirements.
2. **Evaluation breadth.** Given the computational cost of multimodal agent evaluation, our experiments focus on CT tasks and selected VLM backbones. The skill-based representation and evolution framework provide a natural basis for extending experience reuse to MRI and other imaging modalities.
3. **Resource-aware memory maintenance.** The current framework supports skill revision, consolidation, and deprecation, but does not explicitly optimize a joint storage and latency budget for multimodal repositories.

L FUTURE WORK

Future work can extend experience-driven medical agents beyond the reuse of individual procedures. One promising direction is compositional skill learning, in which complementary inspection, measurement, and verification procedures are organized into reusable higher-level strategies. Explicitly representing relationships between skills could support multi-organ assessment, differential reasoning, and longitudinal comparison. This raises questions about how useful skill combinations can be discovered from trajectories and how they should be adapted when the clinical question or available tools change.

Another direction is to incorporate clinician feedback into skill evolution. Targeted review of selected skill candidates could refine the relevance of visual evidence, the order of verification steps, and the applicability of procedures without requiring review of every trajectory. Such expert-refined procedures could serve as shared knowledge units for different agents, motivating repositories that integrate complementary experience and support reuse across clinical tasks.

Broader deployment also motivates more efficient multimodal skill representations. Studying how compression affects procedural clarity, visual fidelity, and downstream accuracy could clarify the trade-off between resource savings and reliable reuse. Extending the framework to MRI and other imaging settings would further help distinguish modality-specific inspection procedures from transferable clinical reasoning strategies. Together, these directions could broaden experience reuse across medical tasks without repeated parameter updates to the underlying VLM.