REVISION3D: ATTRIBUTION-GUIDED RECURSIVE SELF-IMPROVEMENT FOR 3D MEDICAL PERCEPTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recursive self-improvement (RSI) offers a promising path for overcoming the limited visual capability of current medical imaging agents. Yet applying RSI to volumetric imaging remains difficult: failures can arise from acquisition, perception, training recipe, or downstream inference, while self-generated feedback and logged trajectories provide little guidance on which component should change. We introduce **ReVision3D**, an RSI system that leverages 3D volumes with spatially grounded annotations to determine where visual evidence is lost and recursively improve the corresponding visual capability. A frozen language-model designer proposes revisions to acquisition, perception, training, or inference, while the verifier and system-level objective remain fixed. Our key insight is that an annotated volume forms an **exact replay world for view rendering and spatial verification**: unvisited views can be rendered on demand, and localized predictions can be checked directly against reference masks. This grounded feedback directs targeted revision, while only changes that improve beyond measured seed noise are retained. Each accepted change triggers renewed attribution, allowing the dominant bottleneck to shift across rounds. On abdominal CT, attribution identifies perception as the dominant remaining limitation. Revising that level enables ReVision3D to achieve 79% liver recall and 83% kidney recall at under 0.4 false positives per patient, outperforming the evaluated frozen multimodal foundation models, with the largest gains on small lesions.

1 INTRODUCTION

Multimodal foundation models enable medical imaging agents to reason over images, reports, and clinical context, yet fluent reasoning does not guarantee reliable visual understanding. Reading a three-dimensional scan requires acquiring relevant evidence, recognizing subtle findings when visible, and reconciling observations across the volume. Evidence that is never rendered or perceived cannot be recovered by better downstream reasoning. This limitation is particularly important in volumetric imaging, where a model typically observes only a small fraction of the available image space. Recursive self-improvement (RSI) offers a natural path toward improving how these systems see. Recent efforts suggest a promising direction toward medical agents that improve from diagnostic experience and progressively acquire new capabilities (Li et al., 2024; Fan et al., 2026; Shen et al., 2026). Yet self-improvement of the **visual reading process itself** raises a different question: **what part of the visual system should improve?** A missed finding may result from failed acquisition, weak perception, inconsistent detection, or downstream aggregation, and each failure mode calls for a different intervention. Improving the wrong component can therefore leave the dominant bottleneck unchanged.

Conventional self-improvement provides limited feedback for answering this question. Logged trajectories reveal only what the agent inspected, so they cannot determine whether an unvisited slice or crop would have exposed a missed finding. Moreover, self-generated judgments likewise do not independently establish whether visual evidence was localized correctly. Effective visual RSI therefore requires both **attribution**, to identify where evidence is lost, and **external verification**, to determine whether a proposed revision genuinely improves visual capability. Volumetric medical imaging provides unusual structure for both. Given a scan and reference masks, previously unvisited views can be rendered on demand, their lesion coverage computed directly from geometry, and localized predictions mapped into a shared 3D coordinate system for verification. We therefore treat

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.

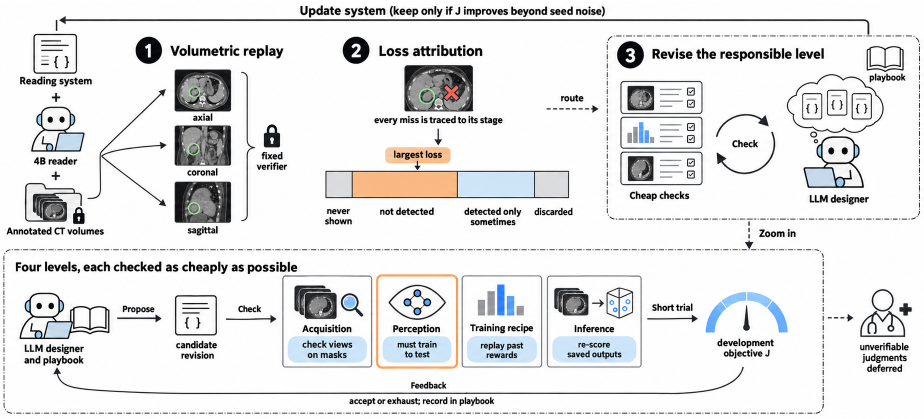


Figure 1: **Overview of ReVision3D. (1) Volumetric replay** renders alternative CT views and verifies localized predictions using a fixed mask-grounded verifier. **(2) Loss attribution** traces each missed lesion to the stage where evidence is lost. **(3) Revision** targets acquisition, perception, training recipe, or inference using a frozen LLM designer and replay-based checks where available. Changes are retained only when the fixed development objective J improves beyond measured seed noise, after which attribution is recomputed.

an annotated scan as an **exact replay world for rendering and localization**: a reusable environment for counterfactual view evaluation and grounded perception feedback without relying on the reader to judge itself.

Building on this formulation, we introduce **ReVision3D**, a framework for **attribution-guided recursive self-improvement** of medical visual capability (Figure 1). ReVision3D traces missed lesions through the visual pipeline and directs a frozen language-model designer toward one of four intervention levels: **acquisition**, **perception**, **training recipe**, or **inference**. Candidate revisions are screened through replay whenever possible, while perception changes require actual training. All candidates are then evaluated under a fixed verifier and system-level objective, and retained only when their gains exceed measured variability. Each accepted change triggers renewed attribution, allowing the dominant bottleneck to shift across rounds. We study this process on abdominal CT lesion localization, a controlled and spatially verifiable instance of medical visual perception. Recipe-level revisions improve how existing detections are used but do not move the visual ceiling; attribution instead identifies perception as the dominant remaining limitation. Revising the visual representation substantially improves recall for both liver lesions and kidney tumors, including small lesions, while inference over successive readers recovers complementary evidence unavailable to any single reader. The loop also rejects revisions that remain within seed variability rather than forcing monotonic progress. Our contributions are threefold:

- **Volumetric replay.** We formulate annotated volumes as exact replay environments for rendering previously unvisited views and grounding localized predictions in a shared 3D coordinate system.
- **Attribution-guided RSI.** We attribute failures across acquisition, perception, training, and inference and direct improvement toward the corresponding intervention space while keeping the verifier, system-level objective, and acceptance rule fixed.
- **Bounded visual self-improvement.** On abdominal CT, we show that different interventions move different failure modes, that perception adaptation improves liver and kidney recall after recipe-level progress saturates, and that the loop can reject unsupported changes and expose limitations outside its current intervention space.

2 Related Work

Medical imaging agents and grounded perception. Volumetric medical agents extend multimodal foundation models to CT through slice rendering and segmentation or measurement tools, often

leaving the image reader itself fixed (Wang et al., 2026; Mao et al., 2026). Complementary work trains medical readers with scan-grounded rewards (Lin et al., 2026; Li et al., 2026a;b), while active-perception methods learn cropping or zooming policies to acquire informative observations (Zheng et al., 2026b; Liu et al., 2026). ReVision3D connects these directions by treating acquisition, perception, training, and inference as distinct intervention spaces and using volumetric replay to evaluate counterfactual views without reader inference.

Recursive self-improvement and self-evolving agents. Prior work on recursive improvement spans executable artifact search, where language models iteratively propose and evaluate programs or algorithms against external objectives (Romera-Paredes et al., 2024; Shojaee et al., 2025; Novikov et al., 2025), and agent-level self-evolution, where accumulated experience is used to refine memory, skills, tools, or execution policies (Li et al., 2024; Fan et al., 2026; Shen et al., 2026). Dream-RSI further uses recorded discovery trees as replay environments for improving exploration policies (Zheng et al., 2026a). Recent medical systems extend this direction by turning diagnostic experience into reusable clinical capabilities and persistent tools (Fan et al., 2026). ReVision3D differs in both the *object of improvement* and the *source of feedback*: it recursively revises the visual reading process itself, while an annotated 3D volume provides counterfactual observations that need not appear in any logged trajectory. Volumetric replay and loss attribution then determine whether the next revision should target acquisition, perception, training, or inference under a fixed external verifier.

3 REVISION3D: RECURSIVE SELF-IMPROVEMENT THROUGH VOLUMETRIC REPLAY

Motivation: annotated scans as replay worlds. A stored medical volume contains more information than the views an agent happened to inspect. Given the acquisition geometry, one can render any admissible slice or crop after the fact and determine which annotated lesions that view exposes. This resembles model-based reinforcement learning, where simulated experience evaluates policies without repeated interaction (Hafner et al., 2023), and replay-based RSI, where recorded discovery trees support counterfactual search over previously explored branches (Zheng et al., 2026a). A volume with reference masks provides a stronger object for the spatially verifiable part of medical perception: rendering is known rather than learned, unvisited views remain available, and localized predictions can be checked against fixed geometry. It therefore supports two forms of replay needed for visual self-improvement: counterfactual replay of **where to look** and grounded replay of **what was localized**.

As shown in Figure 1, ReVision3D alternates between **attributing failures** in replay and **revising** the medical reading system. The system $\Sigma = (\psi, \varphi, \theta, s)$ contains a view-selection program ψ, a perception configuration φ, reader weights θ, and an inference program s. Each revision cycle consists of three steps. Replay first attributes each missed lesion to the stage where evidence is lost. A frozen language-model designer then revises the corresponding component (ψ, φ, s, or the training recipe η that produces θ). Finally, a fixed evaluator determines whether the revision is retained. Only Σ is allowed to change. The verifier, reference masks, system objective, and acceptance rule remain fixed, and the designer cannot inspect evaluation examples or modify what scores it.

Formalizing volumetric replay. A replay world is $w = (x, m)$, where x is a CT volume and m its reference lesion masks. A view a (a plane and position, or a zoomed crop) is rendered deterministically as $I_a = \mathcal{R}(x, a)$, together with a known map ϕ_a from image to physical coordinates. When the reader reports a location $\mathbf{p}_i$ on view a_i, the finding is lifted to $\mathbf{z}_i = \phi_{a_i}(\mathbf{p}_i)$ in a shared 3D frame, where findings from different slices and planes are reconciled into lesion-level candidates. Deployment and replay use the same read interface, while replay additionally renders views that were never requested and evaluates localized findings against m. The replay world is therefore exact for rendering, geometric exposure, and spatial verification, but it does not assume the reader's response to an unvisited view without actually running the reader.

This structure provides two complementary forms of grounded feedback: *counterfactual exposure*, which asks whether a view would show a lesion geometrically, and *grounded verification*, which asks whether a reported location corresponds to a reference lesion. We use the same fixed spatial tolerance $\tau = 5\,\text{mm}$ for both.

Counterfactual exposure. For a view set A, let $\mathrm{Vis}(A)$ denote the lesions geometrically exposed by at least one view in A. A lesion is considered exposed when a view plane intersects its extent or passes within τ of it. Because $\mathrm{Vis}(A)$ depends only on scan geometry and reference masks, a view-selection program can be evaluated for the lesions it would expose without invoking the reader. Exposure and detection are therefore distinct: a lesion may be geometrically covered by a rendered view yet still be missed by the reader.

Grounded verification. For prediction grounding, each lifted location $\mathbf{z}_i$ is matched to the nearest reference lesion within the same tolerance τ, yielding distinct hits $H(y)$, duplicates $D(y)$, and ungrounded findings $F(y)$. For a parseable output y, the training reward is

$$R_\lambda(y; w, A) = \underbrace{|H(y)|}_{\text{lesions found}} - \underbrace{\lambda_d D(y)}_{\text{duplicates}} - \underbrace{\lambda_f F(y)}_{\text{false findings}} + \underbrace{b_0\, \mathbf{1}[y = \varnothing \wedge \mathrm{Vis}(A) = \varnothing]}_{\text{correct silence}}, \tag{1}$$

with $b_0 = 0.5$ and $\lambda = (\lambda_d, \lambda_f) \in [0.1, 1]^2$. The verifier is fixed and a training recipe may tune only the bounded weights λ, which shift the precision-recall trade-off but cannot redefine a correct localization. Reference masks are used only by replay and verification and are never shown to the reader.

System objective. Grounded verification defines whether a predicted location is correct, and R_λ uses that signal to train the reader. System promotion, however, is determined by a separate fixed objective. Thresholding the inference score produces operating points $(\mathrm{Rec}_k, \mathrm{FP}_k)$. For false-positive budgets $\mathcal{B} = \{0.25, 0.5, 1, 2\}$ per patient, we define

$$J(\Sigma) = \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \max_{k:\, \mathrm{FP}_k \le b} \mathrm{Rec}_k, \tag{2}$$

macro-averaged over organs. Unlike the training reward, J has no editable parameters: every candidate revision is judged by the same criterion. Development patients determine promotion. Final multi-organ evaluation occurs only after all choices are frozen. Appendix B documents overlap with earlier component-study test cohorts.

Failure attribution. Replay lets us trace each missed lesion through the stages of the deployed visual pipeline. For lesion j, let E_j indicate that it is geometrically exposed by the deployed views (or is hit by any recorded read), $\widehat{Q}_j^G$ that it is detected either by the deployed read or by at least one of G additional sampled reads of the same views, D_j that it is detected by the deployed read, and K_j^b that the resulting candidate is retained at false-positive budget b. By construction, $K_j^b \le D_j \le \widehat{Q}_j^G \le E_j$, so

$$1 - \mathrm{Rec}_b = \underbrace{\mathbb{E}[1 - E_j]}_{\text{not shown}} + \underbrace{\mathbb{E}[E_j - \widehat{Q}_j^G]}_{\text{shown, never detected}} + \underbrace{\mathbb{E}[\widehat{Q}_j^G - D_j]}_{\text{detected only when sampled}} + \underbrace{\mathbb{E}[D_j - K_j^b]}_{\text{detected, then discarded}}. \tag{3}$$

The decomposition itself is an exact, non-negative identity. The distinction between the two middle terms is sample-based: with finite G, failure to observe a hit cannot establish that a lesion is truly unreachable by the reader. On development and test patients we therefore do not use additional sampling and report these terms jointly as the *never-detected* share, $\mathbb{E}[E_j - D_j]$. Attribution is used to identify which intervention space remains limiting and should not be interpreted as a causal claim.

Four levels of revision. Each term in Eq. (3) maps to a distinct intervention. *Not shown → acquisition:* ψ requests at most B additional views, and candidate programs are screened using geometric exposure Vis without invoking the reader. *Shown but not detected → perception:* φ selects the adapted vision modules, LoRA rank (Hu et al., 2021), and input resolution. Candidates require actual training with a supervised warm start followed by Eq. (1), while pretrained weights remain frozen. *Detected inconsistently → training recipe:* group-relative optimization (Shao et al., 2024) learns only from prompts with non-identical sampled rewards. A **learnability census** $\mathcal{C}_\theta$ stores these rewards, and for recipe distribution q_η replay estimates

$$\widehat{S}(\eta) = \sum_u q_\eta(u) \left[1 - \sum_v \widehat{p}_u(v)^G \right], \tag{4}$$

the expected fraction of non-degenerate groups, used only as a screening proxy. *Discarded → inference:* s scores merged 3D candidates from label-free features such as cross-slice support and per-reader contributions. Cached outputs allow candidate programs to be evaluated directly on J without retraining.

Recursive improvement. At round t, attribution measures the incumbent's not-shown, never-detected, and discarded losses. The never-detected term contains two qualitatively different cases: (i) lesions that the current reader can detect on some sampled reads, which may still benefit from training, and (ii) lesions that are never detected, which more strongly suggests a perception bottleneck. We estimate this split using the learnability census. Let $r_t = n_{\text{front}}/(n_{\text{front}} + n_{\text{never}})$ denote the fraction of lesion-bearing prompts that lie on the learning frontier among those not always detected. We then define

$$\widehat{L}_{\text{perc}} = (1 - r_t)L_{\text{undet}}, \qquad \widehat{L}_{\text{pol}} = r_t L_{\text{undet}}, \qquad \ell_t = \arg\max_{\ell \notin \mathcal{X}_t} \widehat{L}_\ell, \tag{5}$$

where $\mathcal{X}_t$ contains exhausted intervention levels. Thus, a larger learning-frontier share assigns more of the undetected loss to training, whereas a larger never-hit share directs the search toward perception. This split is a routing heuristic rather than an exact decomposition.

The designer then proposes revisions only within level ℓ_t, using replay screens where available. Screens rank candidates but never promote them. Training revisions are evaluated on matched development patients and seeds ξ and accepted only if $\overline{\Delta J} > \nu$ and $\min_\xi \Delta J^{(\xi)} > 0$, where ν is the incumbent's seed range. Inference revisions require $P(\Delta J > 0) \geq 0.9$ under a paired bootstrap. An accepted revision updates Σ and triggers new attribution, otherwise the level is eventually marked exhausted. Across rounds, an evidence-backed **playbook** carries successful and failed hypotheses forward. The routing rule in Eq. (5) was formulated retrospectively and did not schedule the experiments reported here.

4 Experimental Design

We evaluate ReVision3D on abdominal CT lesion localization using LiTS (Bilic et al., 2023) for liver lesions and KiTS23 (Heller et al., 2023; Myronenko et al., 2023) for kidney tumors. A prediction counts as a hit if its localized point lies within 5 mm of a reference lesion mask and the primary analysis includes lesions of at least 1 mL. The reader is Qwen3-VL-4B-Instruct (Bai et al., 2025), and the default designer is Claude Opus 5. All model, program, and threshold selections are performed with development patients. Final multi-organ evaluation occurs after all choices are frozen. Appendix B documents overlap with earlier component-study test cohorts and gives full splits, rendering details, and optimization settings.

Our primary baselines are thirteen frozen frontier and open vision–language models. Each receives the same rendered images, captions, prompt, and output grammar as our reader on a uniform survey of eight axial slices. Eight slices match the observation budget of our budgeted acquisition study and geometrically expose all liver lesions ≥ 1 mL under the 5 mm tolerance, even lesion-guided oracle slices raise no frontier reader above 0.43 recall (Appendix C). ReVision3D uses a *dense read*: axial slices at 5 mm spacing are processed one per prompt and neighboring findings are merged into 3D candidates. A candidate's *support* is the number of slices reporting it, so varying the support threshold traces a recall–false-positive operating curve. To separate protocol from perception, we also evaluate the strongest matched frontier baseline, Gemini 3.8 Flash, under the same dense read and compare our reader variants under an identical protocol.

For the survey baselines we report recall at ≤ 0.5 false positives per patient, overall and by lesion volume (1–5, 5–20, 20–100, and > 100 mL). System revisions are promoted using J from Eq. (2). Confidence intervals use patient-level bootstrap resampling, and paired systems are compared with paired bootstraps.

Table 1: **Recall by lesion volume on the liver and kidney test patients.** Recall at ≤0.5 false positives per patient. Frontier models read 8 uniform slices, ours every slice (same-protocol comparison: Appendix Table A3). Bold: best in column; underlined: best frontier model.

	Liver: 31 patients, 111 lesions					Kidney: 70 patients, 81 lesions				
	Lesion volume (mL)					Lesion volume (mL)				
Reader	1–5	5–20	20–100	>100	All	1–5	5–20	20–100	>100	All
Number of lesions	52	31	17	11	111	14	24	25	18	81
Mistral Large 3	0.06	0.00	0.06	0.82	0.12	0.00	0.04	0.00	0.61	0.15
Qwen3.5-397B	0.00	0.10	0.06	0.36	0.07	0.00	0.08	0.56	0.94	0.41
Gemini 3.1 Pro	0.12	0.42	0.71	0.73	0.35	0.21	0.17	0.68	**1.00**	0.52
Gemma 4 31B	0.02	0.10	0.41	0.82	0.18	0.14	0.29	0.72	0.94	0.54
GPT-5.5	0.04	0.13	0.24	0.64	0.15	0.00	0.17	0.60	0.83	0.42
Claude Opus 5	0.00	0.10	0.18	0.45	0.10	0.00	0.00	0.28	0.78	0.26
DeepSeek V4.1 Flash	0.00	0.00	0.00	0.55	0.05	0.07	0.04	0.24	0.67	0.25
Qwen3.8-27B	0.00	0.03	0.24	0.55	0.10	0.00	0.04	0.36	0.83	0.31
Qwen3.8-Max	0.13	0.35	0.47	**0.91**	0.32	0.14	0.17	0.52	**1.00**	0.46
GLM-5.3 Flash	0.04	0.10	0.29	0.64	0.15	0.07	0.21	0.32	0.89	0.37
Qwen3.8-Flash	0.02	0.10	0.18	0.45	0.11	0.00	0.21	0.48	0.94	0.42
Gemini 3.8 Flash	0.08	0.48	0.65	**0.91**	0.36	0.21	0.46	0.68	**1.00**	0.60
GPT-6 Astra	0.13	0.29	0.65	0.82	0.32	0.14	0.58	0.92	**1.00**	0.70
Ours: the same 4B reader, dense read (every slice); only the perception configuration differs										
Frozen vision encoder	0.06	0.13	0.41	0.64	0.19	0.14	0.29	0.88	**1.00**	0.60
Adapters, 4 blocks (human-chosen)	0.12	0.19	0.53	0.64	0.25	0.29	0.50	0.92	**1.00**	0.70
ReVision3D (designer's adapters)	**0.69**	**0.84**	**1.00**	0.82	**0.79**	**0.57**	**0.71**	**0.96**	**1.00**	**0.83**

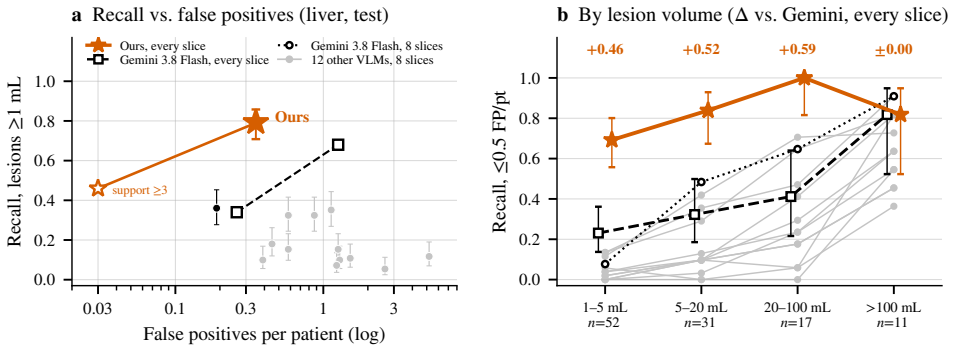


Figure 2: **ReVision3D against frozen frontier models**, 31 liver test patients (111 lesions ≥1 mL). (a) Recall against false positives per patient; dense readers at support (slices reporting a finding) ≥3 (left point) and ≥2 (right point). (b) Recall by lesion volume at ≤0.5 false positives per patient; numbers above the bins are differences in recall from Gemini 3.8 Flash reading every slice under our protocol. Wilson 95% intervals over lesions.

5 Results

5.1 Comparison with Frontier Models

Main results. Table 1 and Figure 2 summarize held-out performance. At ≤ 0.5 false positives (FP) per patient, ReVision3D detects 0.79 of liver lesions (88/111), compared with 0.05–0.36 across thirteen frozen frontier readers. Gemini 3.8 Flash is the strongest frozen reader on liver, reaching 0.36 recall. Kidney tumors are generally easier for the evaluated models, with frozen-reader recall ranging from 0.15 to 0.70. ReVision3D reaches 0.83, compared with 0.70 for the strongest frozen baseline, GPT-6 Astra. The largest gap appears for small lesions. For 1–5 mL lesions, ReVision3D reaches 0.69 recall on liver and 0.57 on kidney, whereas no frozen frontier reader exceeds 0.13 and 0.21, respectively. The advantage narrows for large lesions, where several frozen readers already approach ceiling performance.

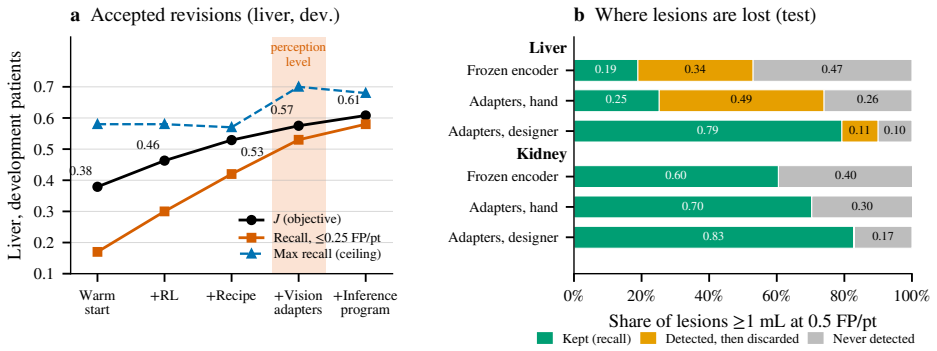


Figure 3: **Attribution shows where improvement stops and which level moves it.** (a) The liver system after each accepted revision (20 development patients, dense read): objective J, recall at $\leq$0.25 false positives per patient, and the maximum reachable recall. (b) Where lesions $\geq$1 mL are lost at 0.5 false positives per patient (Eq. (3)) for the three perception configurations in the single test evaluation

Same reading protocol. Reading every slice does not close the gap by itself. Under the same dense protocol, Gemini 3.8 Flash detects 0.34 of liver lesions at $\leq$ 0.5 FP per patient and reaches 0.68 recall only at 1.26 FP per patient (Figure 2a). Its system objective is $J = 0.34$, compared with 0.71 for ReVision3D (paired difference $+0.37$ [$+0.15$, $+0.47$]). On kidney, J increases from 0.53 to 0.80 ($+0.27$ [$+0.17$, $+0.37$]). Thus, additional slices alone do not explain the improvement: at a comparable false-positive rate, the frozen reader performs similarly under dense and survey reading (0.34 versus 0.36 recall on liver, Appendix Table A3).

Where the gain comes from. The last three rows of Table 1 hold the reader family, training data, and dense protocol fixed while changing only the perception configuration. With the vision encoder frozen, our reader detects 0.19 of liver lesions, adapting four vision blocks raises recall to 0.25, while the designer-selected configuration reaches 0.79. For 1-5 mL lesions, recall rises from 0.06 to 0.12 to 0.69 on liver and from 0.14 to 0.29 to 0.57 on kidney. This controlled comparison isolates perception as the largest source of improvement, consistent with the attribution analysis in Section 5.2.

5.2 Recursive Improvement Trajectory

Improvement trajectory. Figure 3a follows the liver system across successive accepted revisions. Reinforcement learning with the mask-grounded reward raises J from 0.379 to 0.463, and the learnability-aware recipe further raises it to 0.529. These gains primarily improve precision: recall at 0.25 false positives per patient increases from 0.17 to 0.42, while the maximum reachable recall remains nearly unchanged at 0.57–0.58. Recipe-level progress then saturates. Across two recipe searches, none of fourteen designer proposals exceeds its incumbent by more than the measured noise floor after full training (Appendix E).

The bottleneck shifts to perception. Recipe-level improvement leaves a large never-detected term. With the frozen encoder, the share is 0.42/0.45 on liver/kidney development and 0.47/0.40 on test (Figure 3b). Prompts whose sampled outputs never hit these lesions provide no group-relative learning signal. Applied retrospectively, the routing rule in Eq. (5) therefore assigns the dominant remaining loss to perception. Adapting the vision encoder raises the reachable recall ceiling from 0.57 to 0.70 and reduces the never-detected share to 0.25/0.14 on development and 0.10/0.17 on test. The subsequent inference revision combines complementary evidence from successive readers, reaching 0.58 recall at 0.25 false positives per patient and $J = 0.608$.

5.3 Perception Level and Test Evaluation

Perception search. After recipe-level progress saturates, attribution indicates that the dominant remaining loss lies in perception. The designer therefore searches the perception configuration (Figure 4, Appendix Table A8). Doubling the input resolution produces only a small gain, whereas

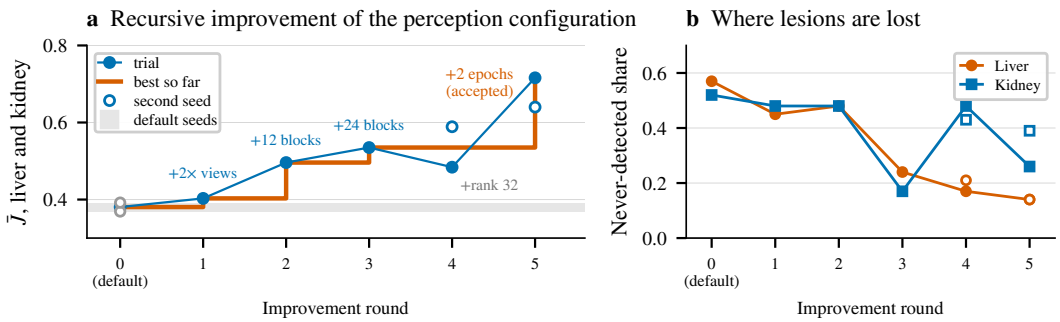


Figure 4: **Perception-level self-improvement. (a)** Recursive search over perception configurations, where $\bar{J}$ is the mean liver and kidney development objective. The orange line tracks the best configuration found so far and hollow markers show second-seed repeats. **(b)** The corresponding never-detected lesion share for liver and kidney. Adapting more vision blocks substantially improves $\bar{J}$ and reduces the never-detected share, while the two-epoch configuration is ultimately accepted.

Table 2: **Perception and inference revisions on liver and kidney.** Paired entries denote seeds s_0 / s_1; single entries report the available evaluation. Bottom rows compare a fixed seed program that merges candidates from successive accepted readers with the inference program selected on development ($P(\Delta J{>}0) = 0.96$).

		Development J (selection)			Test J (once)		
Perception configuration	Chosen by	Liver	Kidney	Mean	Liver	Kidney	Mean
Frozen encoder (language adapters only)	fixed	0.48	0.53	0.51	0.28	0.57	0.43
Adapters, hand (4 blocks, rank 16, 1×)	hand	0.50 / 0.47	0.67 / 0.76	0.58 / 0.61	0.39	0.68	0.54
Adapters, designer (24 blocks, rank 32, 2×)	designer	0.71 / 0.70	0.86 / 0.80	0.79 / 0.75	0.71 / 0.54	0.80 / 0.78	0.75 / 0.66
Inference level on the designer's configuration (program over successive accepted readers)							
+ successive readers merged (seed program)	fixed	0.68	0.80	0.74	0.78	0.80	0.79
+ designer's inference program (Claude Opus 5)	designer	**0.75**	**0.88**	**0.82**	**0.82**	**0.83**	**0.83**

adapting progressively more of the vision encoder raises the mean liver-kidney objective $\bar{J}$ from 0.381 to 0.496 with 12 blocks and 0.535 with all 24 blocks. Increasing adapter rank adds little, while extending the stronger configuration to two training epochs reaches $\bar{J} = 0.716$. An independent repeat reaches 0.640, and the revision is accepted.

Figure 4b shows that the same revisions also move the attribution term that motivated the search. Broader vision adaptation substantially reduces the never-detected share on both liver and kidney, whereas resolution changes and shallower adaptation have smaller effects. The accepted two-epoch configuration reduces this share to 0.14 on liver and 0.26 on kidney in its first trial. Thus, perception adaptation changes not only the operating point but also which lesions the reader is able to detect.

Held-out evaluation. We retrain the three perception configurations on the full training set and evaluate them once on held-out patients (Table 2). The development ordering persists on test: the designer-selected configuration achieves mean $J = 0.75$ and 0.66 across two seeds, compared with 0.54 for the human-selected adapters and 0.43 with the vision encoder frozen. For seed s_0, this corresponds to gains of $+0.21$ [$+0.14$, $+0.28$] over the human-selected configuration and $+0.33$ [$+0.26$, $+0.41$] over the frozen encoder.

The revised reader also provides complementary evidence for the subsequent inference level. Merging candidates from successive readers yields test $J = 0.78$ on liver and 0.80 on kidney. The inference program selected on development ($P(\Delta J > 0) = 0.96$) reaches 0.82 and 0.83 on test. Repeating inference-level search with four designers from three model families yields an accepted program in all twelve runs (Appendix Table A16), indicating that the improvement procedure is not tied to a single designer model.

5.4 Ablation Analysis

Appendix Table A17 isolates the effects of training signal, reading protocol, training recipe, and perception configuration on the same 31 held-out liver patients.

Training signal. A supervised warm start raises survey recall from 0.05 to 0.25, mainly for lesions above 20 mL. Reinforcement learning adds little under the same sparse survey (0.27 recall) and does not improve 1–5 mL lesions, while ten supervised epochs reach 0.29 recall at 1.71 false positives per patient. RL is more useful under dense reading, where development J rises from 0.38 to 0.46, but stronger training alone does not overcome the small-lesion bottleneck.

Reading protocol. Reading every slice raises overall recall from 0.27 to 0.45 and 1–5 mL recall from 0.08 to 0.29, while approximately halving false positives. Confirming detections with a second reader further raises overall recall to 0.56. Increased coverage therefore recovers additional findings, but most small lesions remain undetected.

Training recipe. Recipe changes mainly reshape the precision–recall trade-off rather than expand what the reader can detect. Overall recall remains within 0.41–0.49 and 1–5 mL recall within 0.23–0.31. Training on learnable prompts from fresh views reaches 0.49 recall versus 0.41 for the equal-compute control, while the learning frontier contracts from 25% to 7%, explaining why further recipe search saturates.

Perception configuration. Changing perception produces the largest gain, especially for small lesions. Under the same dense protocol, 1–5 mL recall rises from 0.06 with a frozen encoder and 0.12 with four hand-selected blocks to 0.69 with the designer-selected configuration. Overall recall rises from 0.19 and 0.25 to 0.79. Unlike recipe revisions, perception adaptation expands the set of lesions that the reader detects at all, consistent with the attribution analysis in Section 5.2.

6 Discussion

ReVision3D improves medical visual capability using feedback external to the reader's own judgment. The final reader reaches 0.79 liver-lesion recall and 0.83 kidney-tumor recall at under 0.4 false positives per patient, with the largest gains on 1–5 mL lesions. More importantly, the interventions affect different failure modes: recipe-level revisions refine existing detections, whereas perception adaptation reduces the never-detected share and raises the reachable recall ceiling. Similar inference-level outcomes across four tested designers suggest that fixed external evaluation reduces dependence on the model proposing revisions.

More broadly, recursive improvement benefits from separating **where the system fails** from **how it should be changed**. The tested interventions are not interchangeable: dense reading exposes more findings, recipe changes exploit stochastic detections, and perception adaptation raises the detection ceiling. Re-attributing failure after each accepted revision provides a mechanism for shifting search when the bottleneck moves.

Limitations. The routing rule uses training-patient learnability to estimate which intervention may reduce development loss and was formulated retrospectively, so prospective routing remains to be tested on a new organ or modality. The current loop also fixes rendering choices such as windowing and neighboring-slice context, extending the intervention space to these variables is a natural next step. Additional limitations are discussed in Appendix H.

7 Conclusion

We introduced ReVision3D, an attribution-guided recursive self-improvement framework for 3D medical perception. Annotated CT volumes serve as replay environments in which unvisited views can be rendered, localized findings verified against fixed geometry, and missed lesions attributed to where evidence is lost. This enables an evidence-gated improvement loop over acquisition, perception, training, and inference. On liver and kidney CT, perception emerges as the dominant remaining limitation and produces the largest gain when revised, particularly for small lesions, while unsupported revisions are rejected. These results highlight failure attribution and externally grounded verification as useful principles for recursive improvement of medical imaging systems.

AI USE STATEMENT

Large language models are part of the method: a frozen designer (Claude Opus 5, GPT-6 Astra, Gemini 3.1 Pro, and Gemini 3.8 Flash in Appendix G) proposes revisions of the reading system, and thirteen frontier models serve as baselines. Their roles and interfaces are described in Sections 3 and 4. Generative AI tools were also used to assist with language editing and with scripts used to produce tables and figures from logged results. All generated material was reviewed and verified by the authors against the underlying experiments. The authors take responsibility for the full content of the paper.

ETHICS STATEMENT

This work uses public, de-identified CT datasets (LiTS and KiTS23) under their licences, no new patient data were collected and no human-subject experiments were run. ReVision3D is a research system for lesion localization and is not a clinical device. Reference masks are used only for training, replay and evaluation, the system makes no claim about diagnosis, and clinical judgments that a scan cannot verify are deferred to clinicians (Section 3). Performance on these benchmarks may not transfer to other scanners, protocols or patient populations.

REPRODUCIBILITY STATEMENT

Section 3 defines the reward, objective, attribution identity, screening procedures, and acceptance rule. Appendix B gives data splits, rendering, adapter and optimization settings, the frontier-model protocol, and the self-improvement schedule. The appendix tables additionally report both accepted and rejected trials. Code, designer prompts, playbooks, and trial registries are prepared for release with the supplementary material.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. The liver tumor segmentation benchmark (lits). *Medical image analysis*, 84:102680, 2023.

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.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

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.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

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*, 2024.

Sijing Li, Zhongwei Qiu, Zhuoya Wang, Boxiang Yun, Zhenyu Yi, Jianwei Xu, Wenqiao Zhang, Yingda Xia, and Ling Zhang. E-mrl: Cross-view aligned evidence-driven multimodal reinforcement learning for reliable 3d tumor analysis. *arXiv preprint arXiv:2606.23888*, 2026a.

Yifan Li, Fenghe Tang, Yingtai Li, and Shaohua Kevin Zhou. Medreason-r1: Learning to reason for ct diagnosis with reinforcement learning and local zoom. In *2026 IEEE 23rd International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5. IEEE, 2026b.

Tianwei Lin, Zhongwei Qiu, Jie Cao, Jiang Liu, Wenjie Yan, Bo Zhang, Yu Zhong, Wenqiao Zhang, Yingda Xia, and Ling Zhang. Regulating anatomy-aware rewards via trajectory-integral feedback for volumetric computed tomography analysis. *arXiv preprint arXiv:2605.20277*, 2026.

Ruixun Liu, Bowen Fu, Jiayi Song, Kaiyu Li, Wanchen Li, Lanxuan Xue, Hui Qiao, Weizhan Zhang, Deyu Meng, and Xiangyong Cao. Zoomearth: Active perception for ultra-high-resolution geospatial vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34877–34888, 2026.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.

Yuren Mao, Wenyi Xu, Yuyang Qin, and Yunjun Gao. Ct-agent: A multimodal-llm agent for 3d ct radiology question answering. *Science China Information Sciences*, 69(5):150107, 2026.

Andriy Myronenko, Dong Yang, Yufan He, and Daguang Xu. Automated 3d segmentation of kidneys and tumors in miccai kits 2023 challenge. In *International Challenge on Kidney and Kidney Tumor Segmentation*, pp. 1–7. Springer, 2023.

Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco JR Ruiz, Abbas Mehrabian, et al. Alphaevolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M Pawan Kumar, Emilien Dupont, Francisco JR Ruiz, Jordan S Ellenberg, Pengming Wang, Omar Fawzi, et al. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475, 2024.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Weixiang Shen, Bailiang Jian, Jun Li, Che Liu, Johannes Moll, Xiaobin Hu, Daniel Rueckert, Hongwei Bran Li, and Jiazhen Pan. Evo-medagent: Beyond one-shot diagnosis with agents that remember, reflect, and improve. *arXiv preprint arXiv:2604.14475*, 2026.

Parshin Shojaee, Kazem Meidani, Shashank Gupta, Amir Barati Farimani, and Chandan Reddy. Llm-sr: Scientific equation discovery via programming with large language models. In *International Conference on Learning Representations*, volume 2025, pp. 16054–16085, 2025.

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

Tong Zheng, Xidong Wu, Zheng Zhang, Zhankui He, Chaoyi Zhang, Benjamin Coleman, Ruoqiao Wei, Di Bai, Haolin Liu, Rui Liu, et al. Dream-rsi: Recursive self-improvement through evolving worlds. *arXiv preprint arXiv:2609.14858*, 2026a.

Ziwei Zheng, Minghao Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, and Chao Shen. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 126775–126798, 2026b.

A TERMINOLOGY USED IN TABLES

For clarity, we collect the recurring terms used throughout the main manuscript and appendix tables below.

Reading protocols and inference.

8-slice survey. Eight evenly spaced axial slices presented together in a single prompt. This is the primary protocol used to compare frozen frontier models.

Dense read. Every axial slice is read at 5 mm spacing, one slice per prompt. Findings on neighbouring slices are merged into 3D candidates. Unless otherwise stated, our trained readers are evaluated with this protocol.

Lesion-guided slices. Slices selected to intersect every annotated lesion. This provides an oracle estimate of the benefit available from improved view selection.

Acquisition program. A 4-slice survey followed by up to four additional views selected by the learned view-selection program ψ.

Support. The number of rendered slices on which a 3D candidate is reported. A threshold "support $\geq k$" retains only candidates observed on at least k slices; increasing k typically reduces both false positives and recall.

Cascade. An inference rule in which one reader proposes candidate findings and a second reader must confirm each candidate within 12 mm in the shared 3D coordinate system.

Evaluation metrics and operating points.

Operating point. A recall–false-positive pair obtained by fixing an inference threshold. For example, "recall at ≤ 0.5 FP/patient" denotes the highest recall achievable while keeping false positives at or below 0.5 per patient.

R, FP, and J. R denotes lesion-level recall for lesions ≥ 1 mL; FP denotes false positives per patient; J is the system-level objective defined in Eq. (2).

Failure attribution.

Never detected. No evaluated read localizes the lesion.

Discarded. The lesion is detected by at least one read but removed during downstream aggregation or thresholding.

Kept. The lesion is detected and retained in the final prediction. Together, these terms correspond to the decomposition in Eq. (3).

Reader training and learning signal.

Warm start. Supervised initialization using mask-derived reports.

RL reader. The reader obtained after one round of reinforcement learning with the mask-grounded reward in Eq. (1).

Learning frontier / learnable prompt. A training prompt whose sampled rewards are non-identical and therefore provide non-zero group-relative learning signal. This use of "frontier" is unrelated to *frontier models*.

Cached views. Previously rendered training views reused when constructing a learning-frontier training pool.

Fresh views. Newly rendered views used to recompute the learning frontier under the current reader.

Learnable prompts, cached/fresh. Training recipes restricted to learnable prompts, using either cached or freshly rendered views, respectively.

Recursive improvement protocol.

Seed recipe. The incumbent training recipe at the beginning of a recipe-level search round.

Seed program. The incumbent inference program against which inference-level proposals are evaluated.

Trial scale. Short training or evaluation runs on a fixed development subset used to screen candidate revisions before full evaluation.

Noise floor. The range ν of J across repeated seeds of the incumbent. A training revision is promoted only if its mean gain exceeds ν and every matched seed improves.

B Experimental Details

Data and splits. LiTS provides 131 liver CT volumes and KiTS23 292 kidney CT volumes with voxel-level lesion masks. Lesions are the 26-connected components of the reference masks. The *liver study* (Section 5.2) uses 80/20/31 patients for training, development and held-out evaluation. The *multi-organ study* (Section 5.3) uses 262/60/101 patients for training, development and test (liver 80/20/31, kidney 182/40/70). Each study's splits are patient-disjoint. The 31 held-out liver patients form the liver test set of the multi-organ study. They were also used for the liver-study comparisons in Appendix C, which informed no multi-organ decision, and 40 of the 70 kidney test patients were used by no earlier experiment. A fixed 30-patient subset of the development cohort (10 liver, 20 kidney) scores short designer trials, while the full 60-patient development set governs promotion.

Readers and rendering. The reader is Qwen3-VL-4B-Instruct, a 24-block vision transformer feeding a 36-layer language model, with pretrained weights frozen. Volumes are reoriented to RAS and cropped to an organ box expanded by 15 mm; at deployment the box is supplied by an external organ segmenter, not the reference lesion mask. Axial slices are rendered at 384 pixels on the long side in a $[-160, 240]$ HU window and captioned with plane and position. Low-rank adapters (rank 16) are trained on the language model's attention and MLP projections. A perception configuration adds adapters to the visual mergers and a chosen number of vision blocks and may enlarge views before the vision tower. The warm start fits two epochs of mask-derived reports (learning rate 2×10^{-4}, effective batch 4), while reinforcement learning uses group-relative optimization with the Dr. GRPO loss (Liu et al., 2025) (group size 8–16, temperature 0.9, learning rate 10^{-5}, 100–400 steps).

Reading protocols. The *uniform survey* shows four or eight evenly spaced axial slices in one prompt. The *budgeted read* starts from a four-slice survey and adds at most four views requested by the acquisition program. The *dense read* shows every axial slice at 5 mm, one per prompt (about 43 per patient), and merges findings within 12 mm in 3D. A candidate's support is the number of slices that report it. Lesion-guided slices, which pass through the annotated lesions, serve as an oracle on where to look. Beyond the reasons given in Section 4 for the eight-slice survey, the four-slice survey exposes only 83% of liver lesions, and a slice within 5 mm of a small lesion may show only its edge, which is why we also report lesion-guided slices and the dense-protocol comparison.

Frontier readers. The thirteen frozen readers, in order of release, are Mistral Large 3, Qwen3.5-397B, Gemini 3.1 Pro, Gemma 4 31B, GPT-5.5, Claude Opus 5, Qwen3.8-Max, Qwen3.8-27B, GLM-5.3 Flash, Qwen3.8-Flash, Gemini 3.8 Flash, GPT-6 Astra and DeepSeek V4.1 Flash. Reasoning models may use up to 64k output tokens; if the initial response is empty or unparsable, it is retried once at low reasoning effort.

Metrics and statistics. Recall is pooled over lesions $\geq$1 mL and false positives are counted per patient; J (Eq. (2)) is computed over the dense read's operating points. Uncertainty is a patient-level bootstrap with 2,000 resamples; false-positive differences use a paired Wilcoxon signed-rank test. Loss attribution uses $G = 0$ on development and test patients and $G = 8$ on training patients.

Self-improvement protocol. The designer (Claude Opus 5; GPT-6 Astra, Gemini 3.1 Pro and Gemini 3.8 Flash in Appendix G) runs in a sandbox and receives only aggregate trial statistics, attribution summaries and trial histories. Each recipe search trains 4–10 proposals for 100 steps after 2–3 seeds of the seed recipe set the noise floor, and retrains the leading proposals for 400 steps before acceptance; for an accepted training revision, the median-J seed becomes the next incumbent. The perception round trains six single-change proposals at trial scale on a 120-patient subset and replicates the two best; inference programs are revised up to eight times on cached development reads. The acquisition study is reported as a separate component experiment, under the budgeted-read protocol with geometric screening only, because it was not submitted to the full promotion gate.

C FROZEN FRONTIER READERS

C.1 DOES BETTER ACQUISITION HELP FROZEN READERS?

Table A1: **Baseline check: better views do not help frontier readers.** Liver, 31 held-out patients (111 lesions $\geq$1 mL). In each column every reader receives the same images and prompt, columns change only which slices are shown. R: recall with patient-bootstrap 95% interval (8-slice column); FP: false positives per patient.

	8 uniform slices (protocol of Table 1)		Acquisition program (4 + up to 4 chosen views)		Lesion-guided (oracle) (slices through every lesion)	
Reader	R	FP	R	FP	R	FP
Frozen frontier models, sorted by 8-slice recall						
Gemini 3.8 Flash	**0.36** [0.26, 0.44]	0.19	0.40	0.19	0.38	0.10
Gemini 3.1 Pro	0.35 [0.24, 0.44]	1.13	0.38	0.90	0.35	1.10
GPT-6 Astra	0.32 [0.20, 0.43]	0.58	**0.42**	0.39	**0.43**	0.55
Qwen3.8-Max	0.32 [0.25, 0.38]	0.87	0.29	1.19	0.33	0.90
Gemma 4 31B	0.18 [0.09, 0.31]	0.45	0.26	0.77	0.23	0.90
GPT-5.5	0.15 [0.09, 0.21]	0.58	0.14	0.87	0.19	1.00
GLM-5.3 Flash	0.15 [0.08, 0.22]	1.26	0.23	2.03	0.24	1.97
Mistral Large 3	0.12 [0.06, 0.19]	5.19	0.10	4.97	0.13	4.16
Qwen3.8-Flash	0.11 [0.05, 0.18]	1.52	0.14	1.42	0.20	1.84
Claude Opus 5	0.10 [0.05, 0.15]	1.29	0.14	1.68	0.13	1.68
Qwen3.8-27B	0.10 [0.06, 0.14]	0.39	0.07	0.81	0.12	0.74
Qwen3.5-397B	0.07 [0.02, 0.13]	1.23	0.13	1.39	0.13	1.19
DeepSeek V4.1 Flash	0.05 [0.02, 0.10]	2.61	0.08	2.16	0.05	1.45
Our Qwen3-VL-4B reader with a frozen vision encoder, on the same views and prompt						
RL reader (supervised warm start + one RL round)	0.27 [0.17, 0.38]	0.97	0.33	0.61	0.37	0.35

To separate acquisition from perception, we evaluate thirteen frozen frontier readers and our frozen-encoder RL reader on the same 31 held-out liver patients while changing only the views they receive. We compare the standard 8-slice survey, views selected by the acquisition program, and oracle slices passing through every annotated lesion (Table A1).

On the survey, frontier-model recall ranges from 0.05 to 0.36, with Gemini 3.8 Flash reaching 0.36 at 0.19 false positives per patient. Better exposure produces only modest gains. The acquisition program improves the strongest reader by at most 0.10 (GPT-6 Astra, 0.32 to 0.42), while even oracle lesion-guided slices improve it by only 0.11. Several readers change almost not at all: Gemini 3.8 Flash moves from 0.36 to 0.38, Qwen3.8-Max from 0.32 to 0.33, and DeepSeek V4.1 Flash remains at 0.05. Our frozen-encoder RL reader exhibits the same pattern (0.27, 0.33, and 0.37).

Thus, simply exposing the reader to more informative views does not close the detection gap. Under this protocol, the dominant limitation of frozen readers lies in recognizing visible evidence rather than only in selecting where to look, consistent with the attribution analysis in Section 5.2.

C.2 Lesion Size Reveals the Perception Gap

Table A2: **Recall by lesion volume, 31 held-out liver patients.** Pooled over lesions; each cell is recall (hits/lesions). The same patients are the liver test set of the multi-organ study. Bold: best per column. The two largest bins hold 17 and 11 lesions, so one lesion moves recall by 0.06 and 0.09. Support: the number of slices on which a finding is reported (Appendix A).

Reader *lesions*	1–5 mL *n*=52	5–20 mL *n*=31	20–100 mL *n*=17	>100 mL *n*=11	All *n*=111	FP
Frozen frontier readers, 8 uniform slices						
Gemini 3.8 Flash	0.08 (4/52)	0.48 (15/31)	0.65 (11/17)	**0.91** (10/11)	0.36	0.19
Gemini 3.1 Pro	0.12 (6/52)	0.42 (13/31)	0.71 (12/17)	0.73 (8/11)	0.35	1.13
Qwen3.8-Max	0.13 (7/52)	0.35 (11/31)	0.47 (8/17)	**0.91** (10/11)	0.32	0.87
GPT-6 Astra	0.13 (7/52)	0.29 (9/31)	0.65 (11/17)	0.82 (9/11)	0.32	0.58
Gemma 4 31B	0.02 (1/52)	0.10 (3/31)	0.41 (7/17)	0.82 (9/11)	0.18	0.45
GPT-5.5	0.04 (2/52)	0.13 (4/31)	0.24 (4/17)	0.64 (7/11)	0.15	0.58
GLM-5.3 Flash	0.04 (2/52)	0.10 (3/31)	0.29 (5/17)	0.64 (7/11)	0.15	1.26
Mistral Large 3	0.06 (3/52)	0.00 (0/31)	0.06 (1/17)	0.82 (9/11)	0.12	5.19
Qwen3.8-Flash	0.02 (1/52)	0.10 (3/31)	0.18 (3/17)	0.45 (5/11)	0.11	1.52
Claude Opus 5	0.00 (0/52)	0.10 (3/31)	0.18 (3/17)	0.45 (5/11)	0.10	1.29
Qwen3.8-27B	0.00 (0/52)	0.03 (1/31)	0.24 (4/17)	0.55 (6/11)	0.10	0.39
Qwen3.5-397B	0.00 (0/52)	0.10 (3/31)	0.06 (1/17)	0.36 (4/11)	0.07	1.23
DeepSeek V4.1 Flash	0.00 (0/52)	0.00 (0/31)	0.00 (0/17)	0.55 (6/11)	0.05	2.61
Qwen3-VL-4B base	0.02 (1/52)	0.03 (1/31)	0.06 (1/17)	0.18 (2/11)	0.05	1.32
Ours, the same 8 uniform slices						
Warm start	0.08 (4/52)	0.23 (7/31)	0.47 (8/17)	0.82 (9/11)	0.25	0.42
RL reader	0.08 (4/52)	0.26 (8/31)	0.53 (9/17)	0.82 (9/11)	0.27	0.97
Fresh-view reader	0.08 (4/52)	0.26 (8/31)	0.71 (12/17)	**0.91** (10/11)	0.31	0.68
Second fresh-view reader	0.08 (4/52)	0.26 (8/31)	0.71 (12/17)	0.82 (9/11)	0.30	0.29
Ours, every slice with 3D reconciliation (support $\geq$2)						
RL reader	0.29 (15/52)	0.39 (12/31)	0.76 (13/17)	**0.91** (10/11)	0.45	0.52
Fresh-view reader	0.31 (16/52)	0.45 (14/31)	0.82 (14/17)	**0.91** (10/11)	0.49	0.42
Second fresh-view reader	0.25 (13/52)	0.48 (15/31)	0.76 (13/17)	0.82 (9/11)	0.45	0.32
Cascade: RL proposes, cached-view confirms	0.37 (19/52)	0.58 (18/31)	0.88 (15/17)	**0.91** (10/11)	0.56	0.61
Frozen reader under our protocol: dense read (every slice), $\leq$0.5 FP per patient						
Gemini 3.8 Flash, every slice	0.23 (12/52)	0.32 (10/31)	0.41 (7/17)	0.82 (9/11)	0.34	0.26
Ours, final multi-organ reader, every slice, $\leq$0.5 FP per patient (test evaluation)						
Final reader (designer's configuration, s_0)	**0.69** (36/52)	**0.84** (26/31)	**1.00** (17/17)	0.82 (9/11)	**0.79**	0.35

Table A2 stratifies lesion recall by volume for every reader on the 31 held-out liver patients. Lesion size strongly determines performance: below 5 mL, no frontier model detects more than 0.13 of lesions (7 of 52), whereas ten of thirteen detect at least 0.55 of lesions above 100 mL.

Training the language-side reader improves larger lesions without removing this small-lesion ceiling. On the survey, our readers increase 20–100 mL recall from 0.06 before training to 0.71 after the fresh-view recipe, while 1–5 mL recall remains between 0.02 and 0.08. Dense reading improves the smallest lesions to 0.29 for the RL reader, 0.31 for the fresh-view reader, and 0.37 for the cascade, but the largest change follows perception adaptation. The final reader detects 36 of 52 lesions of 1–5 mL (0.69), 26 of 31 lesions of 5–20 mL, and all 17 lesions of 20–100 mL.

The size-stratified results therefore localize the remaining difficulty: the largest gains from changing perception occur precisely in the small-lesion regime that frozen readers and recipe-level revisions fail to resolve. Because individual bins contain 11–52 lesions, one lesion changes a reported recall by 0.02–0.09.

C.3 Separating Reading Protocol from Perception

Table A3: **Reading protocol and perception configuration, by lesion volume** (test patients). Recall at $\leq$0.5 false positives per patient; FP: false positives per patient at that operating point. Reading every slice does not help the frozen frontier model at a comparable false-positive rate (0.36 on the survey, 0.34 dense on liver), whereas adapting the vision encoder raises our reader's recall in every size bin below 100 mL. Bold: best in column.

	Liver: 31 patients, 111 lesions						Kidney: 70 patients, 81 lesions					
Reader	1–5	5–20	20–100	>100	All	FP	1–5	5–20	20–100	>100	All	FP
Strongest frontier model: survey vs. our protocol												
Gemini 3.8 Flash, 8 uniform slices	0.08	0.48	0.65	**0.91**	0.36	0.19	0.21	0.46	0.68	**1.00**	0.60	0.13
Gemini 3.8 Flash, dense read (our protocol)	0.23	0.32	0.41	0.82	0.34	0.26	0.07	0.25	0.64	0.72	0.44	0.13
Ours: one 4B reader, dense read; perception configurations												
Frozen vision encoder	0.06	0.13	0.41	0.64	0.19	0.00	0.14	0.29	0.88	**1.00**	0.60	0.50
Adapters on 4 blocks (human-chosen)	0.12	0.19	0.53	0.64	0.25	0.06	0.29	0.50	0.92	**1.00**	0.70	0.46
ReVision3D (24 blocks, designer)	**0.69**	**0.84**	**1.00**	0.82	**0.79**	0.35	**0.57**	**0.71**	**0.96**	**1.00**	**0.83**	0.37

Table A3 controls for the reading protocol by evaluating the strongest frontier reader under both the survey and our dense read, then comparing our three perception configurations under the same dense protocol.

Dense reading alone does not explain the gain. For Gemini 3.8 Flash, it improves only the smallest liver-lesion bin (0.08 to 0.23) while reducing recall on larger lesions; overall liver recall remains essentially unchanged (0.36 to 0.34), and kidney recall falls from 0.60 to 0.44. Additional slices also introduce more false positives, forcing a stricter support threshold at the target operating point. In contrast, under the identical dense protocol, our liver reader moves from 0.19 with a frozen encoder to 0.25 with the hand-selected adapters and 0.79 with the designer-selected perception configuration. Kidney recall similarly rises from 0.60 to 0.70 and 0.83.

The advantage of ReVision3D therefore cannot be attributed to reading more slices alone. Holding the protocol fixed, the dominant improvement comes from changing the visual representation itself.

C.4 Kidney Provides a Higher-Contrast Test Case

Table A4: **Baseline check on kidney: frontier readers on the survey and on lesion-guided slices**

	8 uniform slices		Lesion-guided (oracle)	
Reader	R	FP	R	FP
Frozen frontier models, in order of release (bold: best per column)				
Mistral Large 3	0.15 [0.08, 0.23]	2.04	0.15	1.93
Qwen3.5-397B	0.41 [0.30, 0.54]	0.81	0.46	0.80
Gemini 3.1 Pro	0.52 [0.40, 0.66]	0.43	0.67	0.24
Gemma 4 31B	0.54 [0.44, 0.66]	0.26	0.59	0.27
GPT-5.5	0.42 [0.30, 0.56]	0.20	0.53	0.26
Claude Opus 5	0.26 [0.16, 0.37]	0.77	0.27	0.86
Qwen3.8-Max	0.46 [0.34, 0.59]	0.24	0.53	0.26
Qwen3.8-27B	0.31 [0.21, 0.42]	0.50	0.51	0.33
GLM-5.3 Flash	0.37 [0.28, 0.48]	0.83	0.47	0.83
Qwen3.8-Flash	0.42 [0.32, 0.54]	0.64	0.41	0.86
Gemini 3.8 Flash	0.60 [0.47, 0.75]	0.13	0.74	0.06
GPT-6 Astra	**0.70** [0.57, 0.85]	0.14	**0.79**	0.09
DeepSeek V4.1 Flash	0.25 [0.14, 0.36]	0.87	0.28	0.80
Ours, same 70 patients, dense read (every slice)				
Final reader, support $\geq$2	0.72 [0.59, 0.85] at 0.07 FP per patient			
Final reader, support $\geq$1	**0.83** [0.73, 0.93] at 0.37 FP per patient			
Final reader + inference program	0.79 / 0.81 / 0.85 within 0.25 / 0.5 / 1 FP			

Table A5: **Recall by tumour volume, 70 kidney test patients.** Pooled over tumours; each cell is recall (hits/tumours). Support $\geq$1 is the $\leq$0.5 FP operating point used in Table 1.

Reader *tumours*	1–5 mL *n*=14	5–20 mL *n*=24	20–100 mL *n*=25	>100 mL *n*=18	All *n*=81	FP
Frozen frontier models, 8 uniform slices, in order of release						
Mistral Large 3	0.00 (0/14)	0.04 (1/24)	0.00 (0/25)	0.61 (11/18)	0.15	2.04
Qwen3.5-397B	0.00 (0/14)	0.08 (2/24)	0.56 (14/25)	0.94 (17/18)	0.41	0.81
Gemini 3.1 Pro	0.21 (3/14)	0.17 (4/24)	0.68 (17/25)	**1.00** (18/18)	0.52	0.43
Gemma 4 31B	0.14 (2/14)	0.29 (7/24)	0.72 (18/25)	0.94 (17/18)	0.54	0.26
GPT-5.5	0.00 (0/14)	0.17 (4/24)	0.60 (15/25)	0.83 (15/18)	0.42	0.20
Claude Opus 5	0.00 (0/14)	0.00 (0/24)	0.28 (7/25)	0.78 (14/18)	0.26	0.77
Qwen3.8-Max	0.14 (2/14)	0.17 (4/24)	0.52 (13/25)	**1.00** (18/18)	0.46	0.24
Qwen3.8-27B	0.00 (0/14)	0.04 (1/24)	0.36 (9/25)	0.83 (15/18)	0.31	0.50
Qwen3.8-Flash	0.00 (0/14)	0.21 (5/24)	0.48 (12/25)	0.94 (17/18)	0.42	0.64
GLM-5.3 Flash	0.07 (1/14)	0.21 (5/24)	0.32 (8/25)	0.89 (16/18)	0.37	0.83
Gemini 3.8 Flash	0.21 (3/14)	0.46 (11/24)	0.68 (17/25)	**1.00** (18/18)	0.60	0.13
GPT-6 Astra	0.14 (2/14)	0.58 (14/24)	0.92 (23/25)	**1.00** (18/18)	0.70	0.14
DeepSeek V4.1 Flash	0.07 (1/14)	0.04 (1/24)	0.24 (6/25)	0.67 (12/18)	0.25	0.87
Frozen model under our protocol: dense read, $\leq$0.5 FP per patient						
Gemini 3.8 Flash, every slice	0.07 (1/14)	0.25 (6/24)	0.64 (16/25)	0.72 (13/18)	0.44	0.13
Ours, final reader, dense read						
Final reader, dense read, support $\geq$2	0.29 (4/14)	0.58 (14/24)	0.88 (22/25)	**1.00** (18/18)	0.72	0.07
Final reader, dense read, support $\geq$1	**0.57** (8/14)	**0.71** (17/24)	**0.96** (24/25)	**1.00** (18/18)	**0.83**	0.37

We repeat the frozen-reader comparison on 70 kidney test patients containing 81 tumors $\geq$ 1 mL (Tables A4 and A5). Kidney tumors are generally larger and higher contrast than liver lesions, and the frontier readers accordingly perform substantially better: survey recall ranges from 0.15 to 0.70, led by GPT-6 Astra at 0.70 recall and 0.14 false positives per patient.

Oracle exposure has a larger effect than on liver for some readers. GPT-6 Astra reaches 0.79 and Gemini 3.1 Pro rises from 0.52 to 0.67, but weak readers again change little. ReVision3D reaches 0.72 recall at 0.07 false positives per patient and 0.83 at 0.37; after inference-level revision it reaches 0.79, 0.81, and 0.85 recall within false-positive budgets of 0.25, 0.5, and 1 per patient.

The remaining difference again concentrates in small tumors. Below 5 mL, no frontier model detects more than 3 of 14 tumors (0.21), whereas our reader detects 8 (0.57). Above 20 mL, the strongest systems approach ceiling. Kidney therefore provides a complementary setting in which frontier readers are already competitive on large, conspicuous tumors, while the perception advantage remains visible in the smaller-lesion regime.

D THE FINAL READER: PERCEPTION LEVEL, LIVER AND KIDNEY

D.1 ATTRIBUTION ACROSS PERCEPTION CONFIGURATIONS

Table A6: **Where lesions are lost** (Eq. equation 3 with $G = 0$) at 0.5 false positives per patient, development patients. Each row splits the lesions $\geq$1 mL into those no slice read detects, those detected and then discarded by aggregation, and those kept (recall); rows sum to 1. Bold: lowest never-detected and highest kept share.

	Liver			Kidney		
Perception configuration	Never detected	Discarded	Kept	Never detected	Discarded	Kept
Frozen encoder	0.42	0.13	0.45	0.45	0.00	0.55
Adapters, hand	0.30	0.20	0.50	0.30	0.00	0.70
Adapters, designer	**0.25**	0.05	**0.70**	**0.14**	0.00	**0.86**

Table A6 applies the failure decomposition of Eq. (3) to the three perception configurations on development patients. Each lesion is classified as never detected, detected but discarded, or retained in the final prediction.

On liver, adapting perception progressively reduces the never-detected share from 0.42 with the frozen encoder to 0.30 with the hand-selected adapters and 0.25 with the designer-selected configuration. The corresponding kept share rises from 0.45 to 0.50 and 0.70. The hand-selected configuration illustrates why recall alone does not identify the responsible mechanism: it detects additional lesions but also discards more of them (0.13 to 0.20), yielding only a small net improvement. The designer-selected configuration reduces both sources of loss, with the discarded share falling to 0.05. On kidney, no detected tumor is discarded, so the entire gain comes from reducing the never-detected share from 0.45 to 0.30 and finally 0.14.

These decompositions show that perception adaptation changes the loss component it is expected to affect: previously unseen findings become detectable, rather than the gain arising only from a different aggregation threshold.

D.2 Perception Gains Persist Across Lesion Sizes

Table A7: Recall by lesion volume on the development patients (seed s_0), at the most permissive support threshold with at most 0.5 false positives per patient. The test-set counterpart is Table 1. Support: the number of slices on which a finding is reported (Appendix A).

Perception configuration	1–5 mL	5–20 mL	20–100 mL	>100 mL	FP	P
Liver (1–5: 26, 5–20: 15, 20–100: 16, >100: 3 lesions)						
Frozen encoder	0.19 (5/26)	0.60 (9/15)	0.62 (10/16)	**1.00** (3/3)	0.20	0.93
Adapters, hand	0.35 (9/26)	0.53 (8/15)	0.62 (10/16)	**1.00** (3/3)	0.30	0.92
Adapters, designer	**0.50** (13/26)	**0.73** (11/15)	**0.94** (15/16)	**1.00** (3/3)	0.05	0.99
Kidney (1–5: 10, 5–20: 10, 20–100: 13, >100: 11 lesions)						
Frozen encoder	0.10 (1/10)	0.40 (4/10)	0.62 (8/13)	**1.00** (11/11)	0.50	0.83
Adapters, hand	0.40 (4/10)	0.60 (6/10)	0.77 (10/13)	**1.00** (11/11)	0.38	0.88
Adapters, designer	**0.60** (6/10)	**0.80** (8/10)	**1.00** (13/13)	**1.00** (11/11)	0.25	0.92

Table A7 reports the three perception configurations on the development patients used for selection, stratified by lesion volume. The designer-selected configuration performs best in every sub-100 mL bin on both organs while also reducing false positives.

On liver, recall for 1–5 mL lesions increases from 0.19 to 0.35 and 0.50 across the frozen, hand-selected, and designer-selected configurations, while recall for 20–100 mL lesions rises from 0.62 to 0.94. On kidney, 1–5 mL recall increases from 0.10 to 0.40 and 0.60, while false positives fall from 0.50 to 0.25 per patient. The largest-volume bins contain few lesions and saturate across configurations.

The ordering observed on held-out test patients is therefore already present on the development cohort used for selection. Although all configurations exhibit some development-to-test drop in J, the relative advantage of the designer-selected perception configuration is consistent across organs and lesion sizes.

D.3 Searching the Perception Level

Table A8: **Perception-level round at trial scale** (120 training patients; 30 liver and kidney development patients). Each proposal changes one setting of the vision-encoder adaptation. $\bar{J}$ averages liver and kidney; Δ is the difference from the mean of the two default seeds (0.381), whose range, 0.023, is the noise floor. The selected configuration averages 0.678 over two seeds and was accepted.

Trial	Change	Configuration	$\bar{J}$	Δ	Liver J (never det.)	Kidney J (never det.)
base_s0	default	4 blocks, r16, 1.0×, 1 ep	0.369	–	0.293 (0.62)	0.446 (0.52)
base_s1	default (seed 2)	4 blocks, r16, 1.0×, 1 ep	0.392	–	0.328 (0.52)	0.457 (0.52)
p01	2× views	4 blocks, r16, 2.0×, 1 ep	0.403	+0.023	0.328 (0.45)	0.478 (0.48)
p02	12 blocks	12 blocks, r16, 2.0×, 1 ep	0.496	+0.115	0.491 (0.48)	0.500 (0.48)
p03	24 blocks	24 blocks, r16, 2.0×, 1 ep	0.535	+0.155	0.440 (0.24)	0.630 (0.17)
p04	rank 32	24 blocks, r32, 2.0×, 1 ep	0.484	+0.103	0.457 (0.17)	0.511 (0.48)
p05	2 epochs	24 blocks, r32, 2.0×, 2 ep	0.716	+0.335	0.716 (0.14)	0.717 (0.26)
p06	lr 3×10^{-4}	24 blocks, r32, 2.0×, 2 ep, lr 3×10^{-4}	0.000	diverged	0.000 (1.00)	0.000 (1.00)
p05_s1	p05, seed 2	24 blocks, r32, 2.0×, 2 ep	0.640	+0.260	0.716 (0.14)	0.565 (0.39)
p04_s1	p04, seed 2	24 blocks, r32, 2.0×, 1 ep	0.589	+0.208	0.612 (0.21)	0.565 (0.43)

Table A8 records the perception-level search. Two runs of the default configuration first establish a trial-scale noise floor of 0.023; subsequent proposals change one perception setting at a time, and the two strongest proposals are independently repeated.

Increasing image resolution alone produces a gain equal to the noise floor and is therefore insufficient. In contrast, adapting 12 and then all 24 vision blocks increases $\bar{J}$ by 0.115 and 0.155, accompanied by large reductions in the never-detected share. Increasing adapter rank alone provides little additional benefit, whereas a second training epoch on top of the stronger adaptation produces the largest trial-scale objective, $\bar{J} = 0.716$. Increasing the learning rate causes divergence. Replicates of the two leading configurations preserve their ordering, although their variation is larger than that of the default runs.

The search therefore identifies a qualitatively different intervention from additional views or recipe changes: exposing more of the vision encoder to adaptation, followed by longer training, directly reduces the never-detected component that motivated the perception-level search.

E Recipe level: trajectory, recipe searches and learnability

E.1 Accepted Revisions Change Different Parts of the Operating Curve

Table A9: **The liver system after each accepted revision** (20 development patients, 60 lesions $\geq$1 mL, dense read). The warm start is supervised on mask-derived reports; RL uses the mask-grounded reward. Max R is the recall reachable at any budget: it stays near 0.58 until the perception level is revised.

				R / P at FP per patient		Max
System after each stage	Change made by	J	ΔJ	$\leq$0.25	$\leq$0.5	R
Warm start	fixed pipeline	0.379	–	0.17 / 0.93	0.45 / 0.88	0.58
+ RL	fixed pipeline	0.463	+0.084	0.30 / 1.00	0.52 / 0.91	0.58
+ learnability recipe	designer, 3 seeds	0.529	+0.066	0.42 / 1.00	0.57 / 0.92	0.57
+ vision-encoder adapters	hand, on designer's diagnosis	0.575	+0.046	0.53 / 0.95	0.53 / 0.95	**0.70**
+ inference program (own readers)	designer (accepted)	**0.608**	+0.033	0.58 / 0.97	0.58 / 0.97	0.68

Table A9 reports the liver system after each accepted revision. Across the trajectory, J increases from 0.379 to 0.463, 0.529, 0.575, and finally 0.608. However, these gains arise through different mechanisms.

The early training stages primarily improve precision. Recall at 0.25 false positives per patient increases from 0.17 to 0.42, while the maximum recall attainable at any operating point remains approximately unchanged at 0.57–0.58. The perception revision is the first intervention that raises this ceiling, from 0.57 to 0.70. Precision remains at least 0.88 throughout.

Thus, an increasing scalar objective alone does not reveal what has improved. Recipe-level revisions make better use of detections already available to the reader, whereas the perception revision expands the set of lesions that the system can detect.

E.2 First Recipe Search: Short-Run Gains Do Not Survive Promotion

Table A10: **First recipe-search round from the RL reader.** Ten designer recipes are screened for 100 steps and the two strongest are retrained for 400 steps. Two seed repeats define the 0.050 noise floor. The final columns compare the measured and replay-predicted fractions of training steps with learning signal.

			Gain			R / FP at support		Steps with gradient	
Trial	Recipe by	J	vs. start	$P(\Delta J{>}0)$	vs. seed	≥ 1	≥ 2	measured	replay
seed s_0	seed recipe	0.467	+0.004	0.09	–	0.60 / 3.25	0.47 / 0.25	0.29	0.36
seed s_1	seed recipe	0.517	+0.054	0.75	–	0.58 / 0.95	0.45 / 0.05	0.34	0.36
p01	designer	0.533	+0.071	0.88	+0.041	0.58 / 0.35	0.38 / 0.00	0.55	0.76
p02	designer	0.517	+0.054	0.70	+0.025	0.57 / 1.00	0.47 / 0.10	0.53	0.74
p03	designer	0.500	+0.037	0.55	+0.008	0.58 / 0.90	0.42 / 0.10	0.40	0.83
p04	designer	0.525	+0.062	0.89	+0.033	0.57 / 0.35	0.40 / 0.00	0.37	0.83
p05	designer	0.500	+0.037	0.66	+0.008	0.58 / 0.80	0.42 / 0.05	0.66	0.84
p06	designer	0.517	+0.054	0.84	+0.025	0.58 / 0.70	0.45 / 0.05	0.66	0.83
p07	designer	0.525	+0.062	0.82	+0.033	0.58 / 0.85	0.47 / 0.05	0.50	0.82
p08	designer	0.508	+0.046	0.67	+0.016	0.58 / 0.95	0.43 / 0.10	0.65	0.79
p09	designer	0.508	+0.046	0.72	+0.016	0.58 / 0.85	0.43 / 0.10	0.50	0.77
p10	designer	0.483	+0.021	0.71	−0.009	0.57 / 0.60	0.40 / 0.05	0.40	0.80
p01, 400 steps	designer	0.508	+0.046	0.65	+0.016	0.58 / 1.30	0.48 / 0.20	0.34	0.76
p07, 400 steps	designer	0.479	+0.017	0.37	−0.013	0.62 / 1.80	0.43 / 0.10	0.31	0.82
hand*, 400 steps	hand-designed	0.529	+0.067	0.86	+0.037	0.57 / 0.40	0.42 / 0.00	–	–

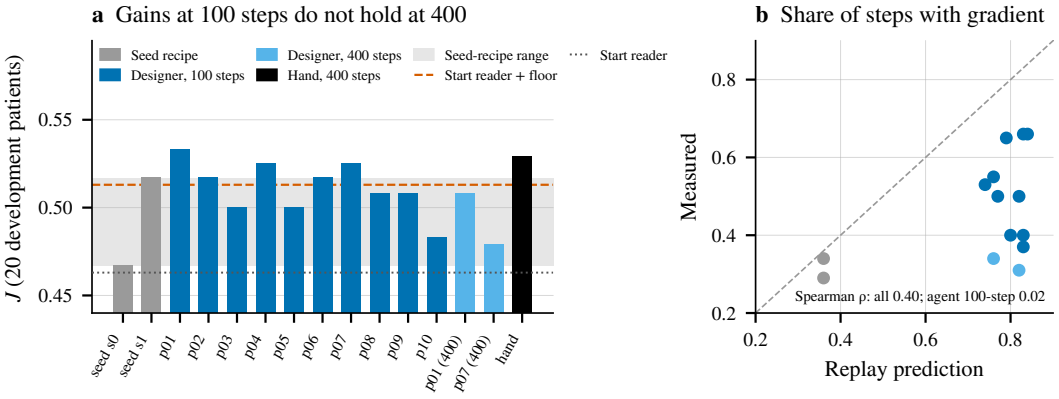


Figure A1: **Short-run recipe gains do not survive full training. (a)** Development objective J for seed, designer, and reference recipes; the dashed threshold is the incumbent plus its measured noise floor. **(b)** Replay-predicted versus measured fraction of training steps with learning signal.

The search starts from the RL reader at $J = 0.463$. Two seed-recipe repeats establish a noise floor of 0.050, after which the designer proposes ten recipes that are screened for 100 steps (Table A10; Figure A1). Three short trials appear to clear the acceptance threshold, but the two strongest candidates fail to retain these gains when extended to 400 steps: their improvements fall to $+0.046$ and $+0.017$, so neither is promoted.

The replay census is useful as a coarse screen but not as a predictor of final performance. It overestimates the realized fraction of training steps carrying learning signal, and its ranking of the designer's short-run proposals is nearly uninformative (Spearman $\rho = 0.02$). The hand-designed learnability recipe reaches $J = 0.529$ after full training and is included only as a reference.

This round illustrates the separation between *proposal* and *promotion* in ReVision3D: inexpensive screens and short trials can suggest promising revisions, but only improvement on the fixed objective beyond the measured noise floor is retained.

E.3 Recipe Search After Perception Adaptation: No Further Gain

Table A11: **Second recipe search, from the vision-adapter reader** ($J = 0.575$, never-detected share 0.30; 20 development patients, 100 steps per trial). Noise floor: range of the three seed-recipe runs, 0.042. No proposal improves on the start reader by more than the floor, and the search is rejected. Never detected: share of lesions no slice read detects (Eq. equation 3).

Trial	Recipe by	J	vs. start	Never detected
seed s_0	seed recipe	0.567	−0.008	0.27
seed s_1	seed recipe	0.546	−0.029	0.32
seed s_2	seed recipe	0.525	−0.050	0.23
p01	designer	0.583	+0.008	0.37
p02	designer	0.533	−0.042	0.38
p03	designer	0.554	−0.021	0.27
p04	designer	**0.592**	+0.017	0.32

After accepting the perception revision, ReVision3D revisits the recipe level from the stronger incumbent ($J = 0.575$, Table A11). Three seed-recipe repeats establish a noise floor of 0.042, and the designer proposes four new recipes using the updated attribution.

None passes the promotion gate. The strongest proposal reaches $J = 0.592$, only $+0.017$ over the incumbent and well within measured seed variability. Its never-detected share also remains within the range of the seed runs. The round is therefore rejected.

This negative result is part of the recursive process: after perception changes, attribution is recomputed and previously exhausted levels can be reconsidered. Here, revisiting the recipe level produces no measurable gain, providing evidence that the remaining loss is not addressed by further recipe changes.

E.4 Learnability Census and Exhaustion of Recipe-Level Signal

Table A12: **Learnability census on training patients.** Single-slice prompts are evaluated with eight samples at $T = 0.9$. *Learnable* denotes prompts whose sampled rewards differ and therefore provide non-zero group-relative learning signal. ‡ denotes the training pool used for the second fresh-view reader.

		Lesion-bearing prompts			Lesion-free prompts			
Reader, prompt pool	Prompts	Always hit	Learnable	Never hit	Always empty	Sometimes FP	Always FP	Learnable
RL reader, cached slices	1,593	24%	37%	39%	84%	13%	3%	**25%**
RL reader, fresh views	3,083	19%	31%	50%	84%	14%	1%	**23%**
Fresh-view reader, fresh views‡	2,797	27%	11%	62%	93%	4%	3%	**7%**

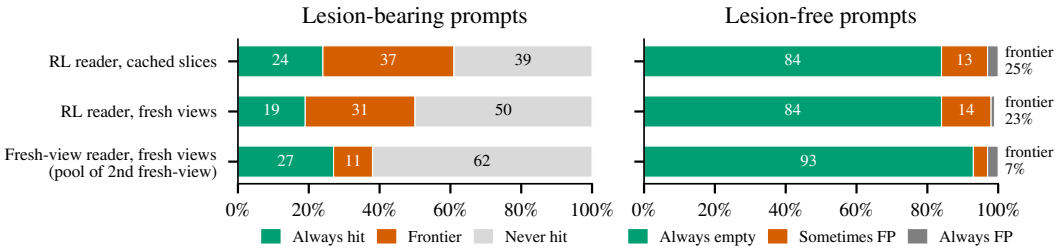


Figure A2: **Learnability census.** Training prompts are grouped by the outcomes of eight sampled reads. For lesion-bearing prompts, only the frontier group (i.e. where sampled rewards differ) provides group-relative learning signal. The overall frontier share is shown on the right.

Table A12 and Figure A2 partition prompts by the outcomes of eight sampled reads. Lesion-bearing prompts are classified as always hit, frontier (sampled rewards differ), or never hit; only frontier prompts provide non-zero group-relative learning signal.

For the RL reader, 25% of cached-view prompts and 23% of fresh-view prompts lie on this frontier, while 39% and 50% of lesion-bearing prompts are never hit. After training on learnable fresh

views, the frontier contracts to 7% overall: more prompts become consistently solved, but 62% of lesion-bearing prompts remain never hit. Lesion-free prompts also become more stable, with the always-empty share increasing from 84% to 93%.

The census explains why recipe-level optimization saturates. Training progressively consumes prompts that still exhibit reward variation, while never-hit prompts provide no group-relative signal. As the frontier contracts, additional recipe changes have diminishing leverage, motivating a move to the perception level.

E.5 The Training Procedure Transfers Better Than the Reader

Table A13: **Transfer of the liver recipe to kidney.** The same training procedure is applied unchanged to 30 held-out KiTS23 patients.

		8-slice survey		+ lesion-guided slices	
Reader	Training on kidney	R	FP	R	FP
Liver reader, zero-shot	none on kidney	0.16 [0.03, 0.30]	0.37	0.23	0.53
Kidney warm start	supervised	0.48 [0.31, 0.67]	0.40	0.58	0.30
Kidney RL reader	mask-grounded reward	0.58 [0.41, 0.77]	0.33	0.61	0.30

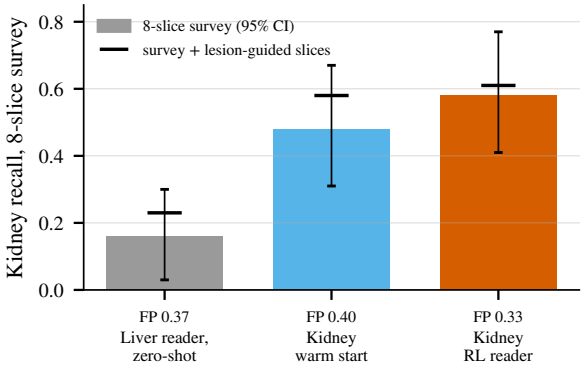


Figure A3: **Transfer from liver to kidney.** Bars show recall on the 8-slice survey (95% intervals); ticks show recall after adding lesion-guided slices. False positives per patient are reported below each bar.

Table A13 and Figure A3 distinguish transfer of the trained reader from transfer of the training procedure. Applying the liver reader directly to kidney generalizes poorly, reaching 0.16 recall at 0.37 false positives per patient. In contrast, reusing the same warm-start and mask-grounded RL procedure on kidney data reaches 0.48 recall after supervised initialization and 0.58 at 0.33 false positives per patient after RL.

Adding lesion-guided slices raises recall only from 0.58 to 0.61, suggesting that limited view exposure explains little of the remaining error. Thus, in this experiment, the training procedure transfers across organs better than the resulting organ-specific reader: the same optimization recipe produces useful readers for both liver and kidney, whereas the liver-trained reader itself does not transfer directly.

F Successive Readers and Cascades

F.1 Successive Recipe Revisions Reach a Similar Ceiling

Table A14: **Successive recipe-trained readers on 31 held-out liver patients.** All results use dense reading with support ≥ 2. R, FP, and P denote lesion recall (95% patient-bootstrap interval), false positives per patient, and precision. *Zero-variance steps* is the fraction of training steps whose sampled reward group is constant; *Learnable* is the pre-training fraction of candidate prompts whose eight sampled rewards differ. Support is defined in Appendix A.

Reader	Trained on	R	FP	P	Zero-variance steps	Learnable
RL reader	mined failures	0.45 [0.34, 0.56]	0.52	0.76	19%	–
Longer-training control	un-mined	0.41 [0.28, 0.54]	0.39	0.79	88%	–
Failure-mined reader	failure-mined cases	0.46 [0.35, 0.55]	0.74	0.69	55%	–
Cached-view reader	learnable, cached views	0.41 [0.31, 0.51]	**0.06**	**0.96**	75%	25%
Fresh-view reader	learnable, fresh views	**0.49** [0.38, 0.59]	0.42	0.81	62%	23%
Second fresh-view reader	learnable, fresh views	0.45 [0.33, 0.58]	0.32	0.83	53%	7%

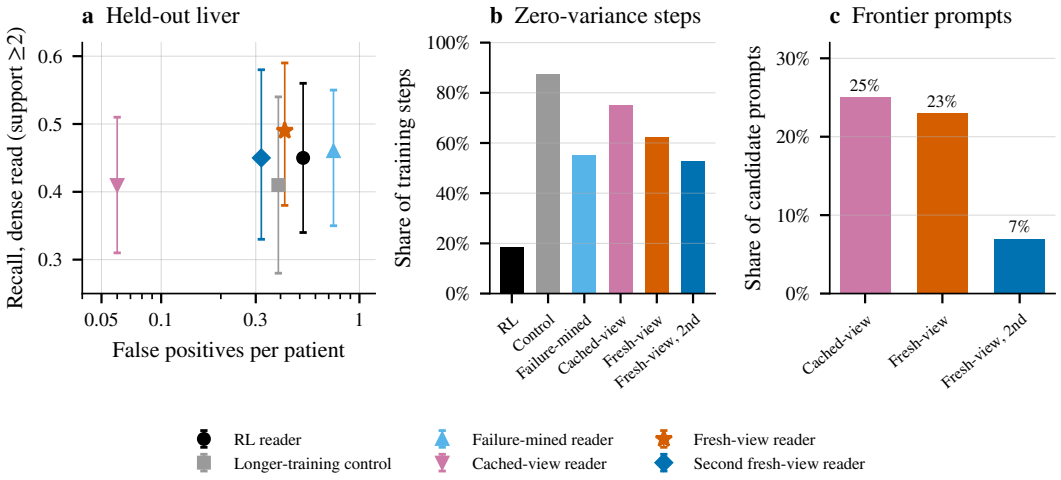


Figure A4: **Successive recipe-trained readers on held-out liver. (a)** Dense-read recall and false positives at support ≥ 2 (95% patient-bootstrap intervals). **(b)** Fraction of training steps with zero reward variance. **(c)** Fraction of candidate prompts on the learnable frontier before each reader is trained.

Table A14 and Figure A4 compare readers produced by successive recipe revisions from the same RL reader. Despite different curricula, dense-read recall remains within 0.41–0.49 with overlapping confidence intervals, indicating a largely unchanged single-reader detection ceiling.

The main differences are instead in precision. For example, the cached-view reader reaches precision 0.96 at 0.06 false positives per patient, whereas the failure-mined reader reaches 0.69 at 0.74. Meanwhile, the available learning signal contracts: 53–88% of training steps for post-RL readers have zero reward variance, compared with 19% for the RL reader, and the learnable frontier shrinks from 25% to 7% of candidate prompts.

Thus, successive recipe revisions mainly move the reader along the precision-recall trade-off while progressively exhausting the available group-relative learning signal. This supports treating the recipe level as exhausted.

F.2 Successive Readers Provide Complementary Evidence

Table A15: **Cascades and paired reader comparisons on 31 held-out liver patients.** Absolute columns report the first-named reader; ΔR uses 95% patient-bootstrap intervals and ΔFP Wilcoxon signed-rank p-values. Cascade proposals are retained when confirmed within 12 mm in 3D.

Comparison	R	FP	ΔR	95% CI	ΔFP	p
(a) Cascade: the RL reader proposes, a later reader confirms (dense read)						
RL reader alone	0.45 [0.34, 0.56]	0.52	reference		reference	
RL proposes, cached-view confirms	0.56 [0.44, 0.65]	0.61	+0.11	[+0.03, +0.18]	+0.10	0.821
RL proposes, fresh-view confirms	0.63 [0.53, 0.72]	1.77	+0.18	[+0.12, +0.24]	+1.26	<0.001
RL proposes, second fresh-view confirms	0.65 [0.56, 0.73]	1.03	+0.20	[+0.13, +0.26]	+0.52	0.003
(b) Paired reader comparisons, dense read at support ≥2						
Fresh-view vs RL	0.49 [0.39, 0.59]	0.42	+0.04	[−0.01, +0.08]	−0.10	0.382
Second fresh-view vs RL	0.45 [0.33, 0.57]	0.32	+0.00	[−0.05, +0.06]	−0.19	0.239
Second fresh-view vs fresh-view	0.45 [0.33, 0.57]	0.32	−0.04	[−0.08, +0.01]	−0.10	0.382
Cached-view vs longer-training control	0.41 [0.30, 0.50]	0.06	+0.00	[−0.06, +0.06]	−0.32	0.085
Fresh-view vs longer-training control	0.49 [0.38, 0.59]	0.42	+0.08	[+0.01, +0.14]	+0.03	0.746
Fresh-view vs cached-view	0.49 [0.39, 0.59]	0.42	+0.08	[+0.04, +0.12]	+0.35	0.137
(c) Paired reader comparisons, 8-slice survey						
Fresh-view vs RL	0.31 [0.19, 0.45]	0.68	+0.04	[−0.01, +0.09]	−0.29	0.380
Second fresh-view vs RL	0.30 [0.19, 0.41]	0.29	+0.03	[−0.02, +0.08]	−0.68	0.073

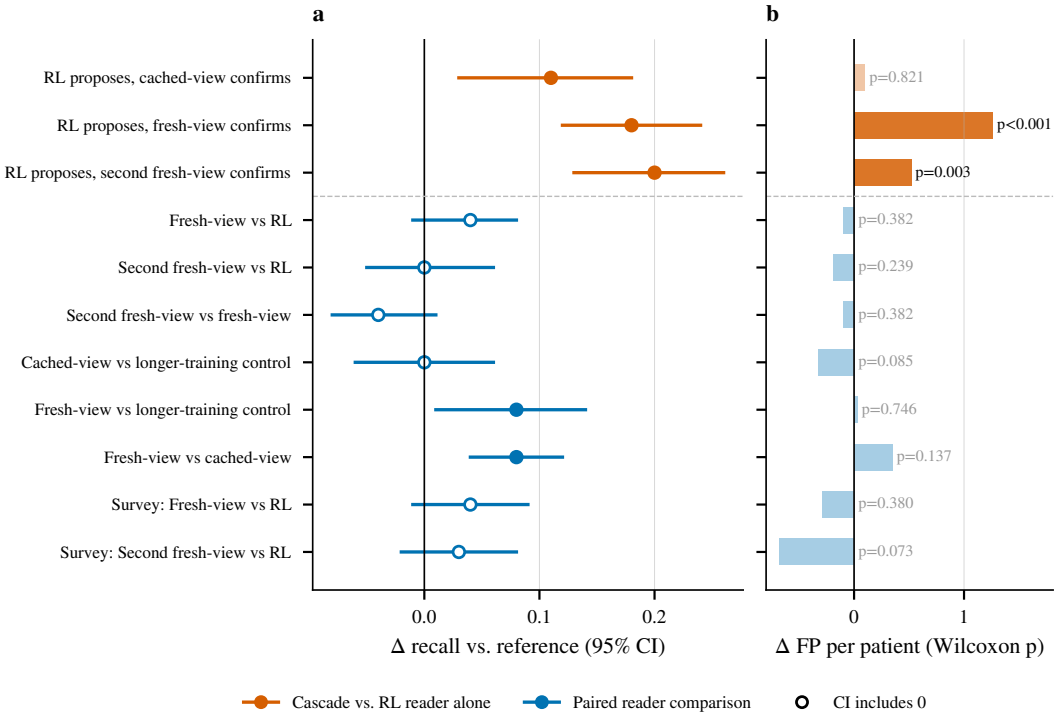


Figure A5: **Complementarity between successive readers on held-out liver. (a)** Paired change in recall with 95% patient-bootstrap intervals; hollow markers indicate intervals containing zero. **(b)** Paired change in false positives per patient; saturated bars denote Wilcoxon $p < 0.05$.

Although successive readers reach similar recall individually, their errors are not identical. Table A15 and Figure A5 measure this complementarity using proposal-confirmation cascades and paired reader comparisons on the same patients.

Using the cached-view reader to confirm the RL reader increases recall by 0.11 [0.03, 0.18] without a significant increase in false positives ($p = 0.82$). Fresh-view confirmation yields larger gains of 0.18–0.20, but at the cost of 0.52–1.26 additional false positives per patient. Among individual readers, only the fresh-view reader significantly improves over the equal-compute control (+0.08 [0.01, 0.14])

and the cached-view reader (+0.08 [0.04, 0.12]); the remaining pairwise recall differences include zero.

Successive readers are therefore more useful as *complementary evidence* than as replacements for one another. This motivates the inference-level revision in ReVision3D, which combines findings across readers rather than selecting a single final checkpoint.

G Designer Models and Ablations

G.1 The Improvement Loop Is Robust Across Tested Designers

Table A16: **Inference-level improvement is consistent across tested designers.** Four language-model designers are evaluated from the same three incumbents. J is the accepted development objective and ΔJ its gain over the incumbent. *Rev.* denotes the accepted revision index among up to eight proposals; acceptance requires $P(\Delta J > 0) \geq 0.9$.

	Liver, liver-study readers incumbent $J = 0.575$			Liver, multi-organ readers incumbent $J = 0.679$			Kidney, multi-organ readers incumbent $J = 0.801$		
Designer	J	ΔJ	Rev.	J	ΔJ	Rev.	J	ΔJ	Rev.
Claude Opus 5	0.608	+0.033	1	**0.750**	+0.071	2	0.881	+0.080	1
GPT-6 Astra	**0.613**	+0.038	1	0.746	+0.067	1	0.881	+0.080	1
Gemini 3.1 Pro	**0.613**	+0.038	1	0.746	+0.067	1	**0.892**	+0.091	1, 2
Gemini 3.8 Flash	**0.613**	+0.038	1	**0.750**	+0.071	1	0.886	+0.085	1
Spread across designers	0.005			0.004			0.011		

To test sensitivity to the language-model designer, Table A16 repeats the inference-level search with four designers from three model families and three different incumbents ($J = 0.575$, 0.679, and 0.801).

All twelve runs produce an accepted program, eleven on the first revision. Within each incumbent setting, accepted objectives differ across designers by at most 0.005, 0.004, and 0.011, respectively. Gemini 3.8 Flash also performs comparably as a designer despite being substantially weaker as a medical image reader under the same evaluation protocol.

Across the tested designers, inference-level improvement is therefore stable to the model proposing revisions. The designer generates candidate programs, while the fixed external objective determines which candidates are retained.

G.2 Ablations Localize the Largest Gain to Perception

Table A17: **Component-wise ablations on the same 31 held-out liver patients.** Each block varies one intervention while holding the others fixed: (a) training signal, (b) reading protocol, (c) training recipe, and (d) perception configuration. Recall is stratified by lesion volume; FP denotes false positives per patient. Block (a) uses the 8-slice survey, while blocks (b)–(d) use the dense protocol at ≤ 0.5 FP per patient or support ≥ 2 as indicated. Bold denotes the best result within each block.

	Lesion volume (mL)					
Setting (liver, 31 test patients)	1–5	5–20	20–100	>100	All	FP
Number of lesions	52	31	17	11	111	
(a) Training signal: our 4B reader after each training stage, 8-slice survey						
Untrained Qwen3-VL-4B (zero-shot)	0.02	0.03	0.06	0.18	0.05	1.32
+ supervised warm start on mask-derived reports	0.08	0.23	0.47	**0.82**	0.25	0.42
+ RL with the mask-grounded reward (the RL reader)	0.08	**0.26**	0.53	**0.82**	0.27	0.97
control: 10 supervised epochs instead of RL	**0.12**	0.19	**0.65**	**0.82**	**0.29**	1.71
(b) Reading protocol, RL reader						
8 uniform slices	0.08	0.26	0.53	0.82	0.27	0.97
dense read (every slice)	0.29	0.39	0.76	**0.91**	0.45	0.52
+ a second reader confirms	**0.37**	**0.58**	**0.88**	**0.91**	**0.56**	0.61
(c) Training recipe: readers trained from the RL reader, dense read						
RL reader (reference)	0.29	0.39	0.76	**0.91**	0.45	0.52
400 more steps, no selection	0.27	0.29	0.71	**0.91**	0.41	0.39
trained on its failures	**0.31**	0.39	0.76	**0.91**	0.46	0.74
learnable prompts, cached views	0.23	0.45	0.59	0.82	0.41	0.06
learnable prompts, fresh views	**0.31**	0.45	**0.82**	**0.91**	**0.49**	0.42
learnable prompts, fresh views, repeated	0.25	**0.48**	0.76	0.82	0.45	0.32
(d) Perception configuration: multi-organ reader, dense read						
Frozen vision encoder	0.06	0.13	0.41	0.64	0.19	0.00
Adapters, 4 blocks (human-chosen)	0.12	0.19	0.53	0.64	0.25	0.06
Adapters, 24 blocks (designer): *ReVision3D*	**0.69**	**0.84**	**1.00**	**0.82**	**0.79**	0.35

Table A17 changes one component at a time on the 31 held-out liver patients and reports recall by lesion volume. The pattern separates the effects of training signal, reading protocol, training recipe, and perception.

A supervised warm start raises overall survey recall from 0.05 to 0.25, mainly for lesions above 20 mL, while RL and a 10-epoch supervised control leave 1–5 mL recall at only 0.08–0.12. Dense reading raises small-lesion recall from 0.08 to 0.29, and confirmation by a second reader reaches 0.37. Recipe variants remain within 0.23–0.31 on the same bin. In contrast, changing the perception configuration raises 1–5 mL recall from 0.06 with the frozen encoder and 0.12 with hand-selected adapters to 0.69, while overall recall rises from 0.19 and 0.25 to 0.79.

The ablations are consistent with the attribution analysis: acquisition and inference recover or reconcile evidence already within reach, recipe changes primarily alter how that evidence is used, and perception adaptation produces the largest increase in the set of lesions detected by the reader.

H Extended Limitations and Future Directions

Dependence on an external verifier. ReVision3D can optimize only outputs that can be checked independently of the reader. In this work, reference lesion masks provide such a verifier for spatial localization, but not for characterization, malignancy, etiology, or other clinical judgments. The framework therefore improves a verifiable component of medical perception rather than clinical reasoning as a whole. A natural extension is to broaden the verifier beyond spatial masks by incorporating additional externally grounded signals, such as longitudinal follow-up, pathology, structured measurements, or cross-modality agreement, so that more dimensions of clinical interpretation can enter the improvement loop.

Bounded intervention space. The designer searches only four predefined levels: acquisition, perception configuration, training recipe, and inference. This makes revisions auditable and comparable under a fixed objective, but also limits what the system can discover. The current loop cannot introduce a new backbone, imaging modality, rendering operator, or downstream reasoning strategy unless that capability is explicitly exposed as an action. Future systems could progressively expand their own intervention space, for example by proposing new visual modules, rendering strategies, or tool interfaces and admitting them only after the same external promotion test.

Replay is exact for geometry, not perception. Volumetric replay can render unvisited views exactly and determine their geometric lesion exposure from reference masks, but it cannot know how the reader will respond to an unseen view without running the model. Acquisition and inference can therefore be screened cheaply, whereas perception revisions still require training and evaluation. A promising direction is to learn lightweight surrogates that predict which perception changes are likely to reduce the never-detected share, using previous trials as experience while retaining the fixed evaluator for final promotion.

Attribution guides intervention but is not causal. The missed-lesion decomposition is exact, but the split between training-level and perception-level loss relies on finite sampled reads and the learnability census. It is therefore a routing heuristic rather than a causal guarantee. Future work could make routing itself adaptive, for example by explicitly estimating the expected improvement of each intervention from prior revisions or by allocating a small exploration budget across multiple levels before committing expensive training compute.

A fixed objective defines the boundary of improvement. Keeping the verifier and system objective fixed is important for preventing the designer from redefining success, but it also constrains what counts as progress. Improvements in lesion localization may not always translate to better downstream clinical decisions. An important next step is to study hierarchical objectives in which spatially verifiable perception remains the grounded inner loop, while higher-level clinical utility is evaluated separately under independent evidence and human oversight.