

MediSketch: Test-Time Scaling for Multimodal Patient Education

Zhipeng Tang¹, Benlu Wang^{1,2}, Zhenyu Zhang³, Wuao Liu¹, Pengda Xiang⁴,
Juncheng Huang⁵, Chin Siang Ong⁶, Hong Yu^{1,7,8}, Zonghai Yao^{1,7,8}

¹UMass Amherst, ²Yale University, ³Stanford University,

⁴University of Southern California, ⁵National University of Singapore,

⁶Yale School of Medicine, ⁷UMass Lowell, ⁸VA Bedford Health Care

{zhipengtang,zonghaiyao}@umass.edu

Abstract

Patient education is a major NLP application, but current evaluation largely focuses on text-only assistance. In radiology, this overlooks an important part of clinical communication: patients often need both a plain-language explanation and a visual aid showing where to look and what it refers to. We introduce MediSketch, a multimodal patient-education task in which an agent must produce an image–text bundle: an educational image and a grounded plain-language explanation. We instantiate this setting with a clinician-curated radiology testbed, a structured multimodal rubric, and a rubric-following judge that evaluates bundle quality rather than text-only correctness. Under a matched image-generation budget, we compare three test-time scaling strategies: parallel sampling, sequential revision, and hybrid search. Across five backbone models, the no-scaling baseline remains weak, while naive parallel search often provides little gain despite using the full budget. Hybrid performs best on 4 of 5 backbones, suggesting that test-time compute works best when exploration and revision are combined. Further analysis identifies weak multimodal self-critique as a major bottleneck: many models stop revising too early and overestimate the quality of their own image–text outputs. MediSketch exposes both an evaluation gap in multimodal patient education and a concrete failure mode in test-time scaling.

1 Introduction

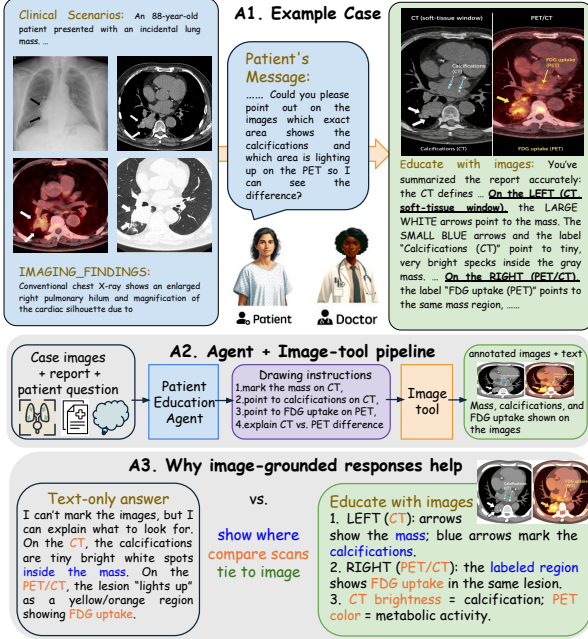
Patient education is one of the clearest ways NLP can create direct social value (OpenAI, 2026; Anthropic, 2026; Delbanco et al., 2012; Walker et al., 2019). Systems for simplification, lay summarization, and question answering aim to help patients understand medical information that would otherwise remain inaccessible (Mehan Jr et al., 2021; Ondov et al., 2022; Welivita and Pu, 2023). Yet

the dominant evaluation target remains narrow (Aydin et al., 2024; Yao and Yu, 2026). Most work still asks whether a model can produce a fluent text explanation, simplify a report, or answer a medical question in pure words. These are useful capabilities, but they do not fully capture whether a system helps patients gain an intuitive understanding of what matters, where to look, and how the explanation is grounded in the underlying medical evidence.

Radiology makes this gap especially visible. Patients increasingly see both radiology reports and the underlying images through portals and follow-up discussions, but access does not guarantee understanding (Mehan Jr et al., 2021; Pollock et al., 2024; Norris et al., 2022). Reports are usually written for clinicians, not patients, and the images are harder still. A patient may read that there is a lesion, calcification, or interval change, yet still not know where it is, how to compare it across images, or why one region matters more than another (Martin-Carreras et al., 2019; Vincoff et al., 2022; Alarifi et al., 2021; Van der Mee et al., 2025). In real interactions, clinicians often bridge this gap visually, for example by circling a finding, pointing to a region, contrasting modalities, or sketching a simple explanation. More broadly, work on visual aids and pictorial health communication suggests that visual support can improve understanding, recall, and health literacy, especially for readers with lower literacy (Schubbe et al., 2020; Galmarini et al., 2024; Card et al., 2022).

However, current evaluation reflects this weakly. Existing medical benchmarks often focus on recognition, short-form visual question answering, report generation, or text-only patient-facing rewriting (Lau et al., 2018; Liu et al., 2021; Boecking et al., 2022; Lyu et al., 2023). These settings are valuable, but they usually stop short of requiring the model to turn medical evidence into a visually grounded explanation for a patient. As Figure 1

A. Task protocol for multimodal patient education



B. Test-time scaling under equal generation budget

Each candidate is one image+text answer bundle.

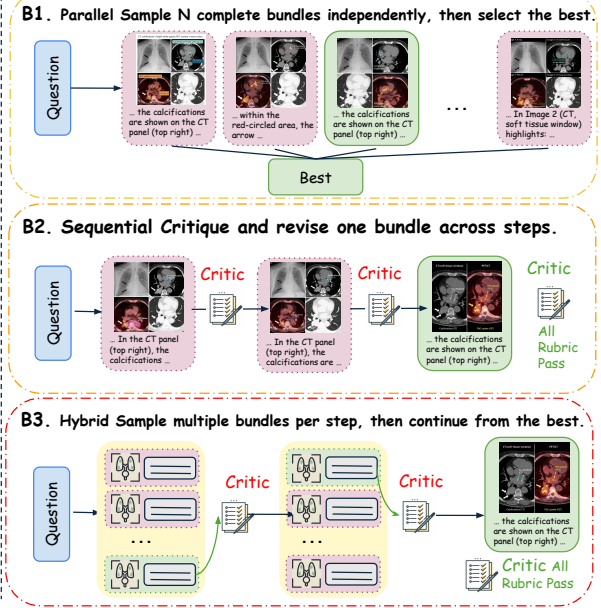


Figure 1: **Overview of MediSketch and equal-budget test-time scaling.** **Left: Multimodal patient education as a bundled output task.** Given radiology images, report findings, and a patient question, the agent must produce a complete image–text response bundle: an annotated educational image and a grounded plain-language explanation. A1 shows an example case in which the patient asks where the calcifications are on CT and which region is “lighting up” on PET. A2 shows the agent–tool pipeline: the agent reads the case materials, generates drawing instructions, calls an image tool to produce the educational image, and then generates the patient-facing explanation conditioned on that image. A3 contrasts a text-only answer with an image-grounded answer, illustrating that success depends not only on verbal explanation but on showing the patient what the explanation refers to. **Right: Test-time scaling under a shared image-generation budget cap.** Each search candidate is a complete image–text bundle rather than a text-only response. We compare **parallel sampling**, which generates multiple bundles independently and selects the best; **sequential revision**, which critiques and improves one bundle across steps; and **hybrid search**, which combines exploration with revision by sampling multiple bundles at each step and continuing from the strongest one. By treating each image–text bundle as the search unit, this setup lets us compare how different compute-allocation strategies improve multimodal patient education quality under the same budget cap.

shows, the visually grounded explanation changes the evaluation target itself: the system must decide what to show, produce an educational visual aid, and generate text that correctly interprets the generated educational image. A model that looks strong at medical interpretation or lay rewriting may still fail at the more applied task of showing patients what matters and explaining it in a way tied to visible evidence.

To address this gap, we introduce **MediSketch**, a multimodal patient-education framework in which an agent must answer a patient question with a complete **image–text** response bundle¹. Given radiology images, report findings, and a patient question, the agent generates drawing instructions, calls an image tool to obtain an annotated educational image, and then produces grounded plain-language

text conditioned on that image. We instantiate this setting on MedThinkVQA (Yao et al., 2026b), built from peer-reviewed Eurorad cases spanning diverse radiology subspecialties and imaging scenarios. The key difference from text-only patient education is the output contract itself: success requires not only a fluent explanation, but also a visual aid that localizes or contrasts the clinically relevant evidence.

This bundled output setting also requires a different evaluation lens. Instead of evaluating systems only by text correctness or a single end-task score, we evaluate whether the complete image–text bundle is medically grounded, patient-useful, visually coherent, and aligned across modalities. To do so, we use a clinician-structured multimodal rubric and a rubric-following judge that scores bundle quality dimension by dimension. This turns multimodal patient education into a more inspectable evalua-

¹MediSketch resources and code: <https://github.com/zptang1210/MediSketch>.

tion problem: system failures can be attributed to specific breakdowns, such as a misleading image, a mismatch between the text and image, an explanation that is not useful to the patient, or a medically unsafe bundle despite fluent language.

With this evaluation lens in place, we study one controlled question: whether test-time scaling improves performance, and if so, how test-time compute should be allocated when each candidate is a coupled image–text educational bundle rather than a text-only answer. In text-only settings, extra inference-time compute is often allocated through best-of- N sampling or iterative self-revision (Wang et al., 2023; Madaan et al., 2023). Here, revision is harder because both the visual aid and the text must improve together, and each additional attempt incurs a real image-generation cost. Under a matched maximum image-generation budget, we compare three test-time scaling strategies: parallel sampling, which generates multiple independent bundles before selection; sequential revision, which repeatedly revises one bundle using critic feedback; and hybrid search, which combines parallel exploration with revision (Wang et al., 2023; Madaan et al., 2023; Yao et al., 2023). Separately, we report direct answering as a non-scaling baseline that generates a single image–text bundle in one pass². Our results first confirm that test-time scaling can improve bundle quality, but the gains depend strongly on how the additional compute is allocated. More specifically, across five backbone models, direct answering remains weak under rubric-based evaluation. Parallel often yields little or no gain over Direct, despite always using the full budget. Hybrid achieves the best model-wise result on 4 of 5 backbone models, suggesting that extra test-time compute helps most when exploration and revision are combined rather than spent on independent samples alone.

Further analysis suggests that a major bottleneck is weak multimodal self-critique rather than budget alone. Many models stop revising after only one or two editing attempts on average, even under a much larger budget cap. This suggests that simply increasing the token budget or image-generation cap does not reliably improve bundle quality. A key issue is therefore not simply whether more com-

putation is available, but whether the model can evaluate and improve an image–text output. This is consistent with broader concerns about self-bias in self-refinement and the difficulty of controlling multimodal hallucination and self-feedback (Setlur et al., 2025; Xu et al., 2024; Lee et al., 2024; Bai et al., 2025).

Our contributions are threefold. First, we introduce MediSketch, a new framework for multimodal patient education in which an agent must generate both an educational image and a grounded patient-facing explanation. Second, we provide a clinician-curated evaluation setting and structured rubric for measuring complete bundle quality rather than text-only correctness. Third, under matched maximum generation budgets, we compare parallel, sequential, and hybrid test-time scaling strategies, showing that hybrid is strongest overall, while weak multimodal self-critique limits how effectively current models use extra test-time compute.

2 Related Work

Medical multimodal modeling and evaluation. Multimodal modeling aims to connect language and vision across a broad range of tasks, such as language-guided segmentation (Lai et al., 2024; Yang et al., 2024) and interleaved image–text generation (Zhou et al., 2025). In medicine, prior work has mainly focused on medical visual question answering, report generation, image–text retrieval, and visual grounding. Representative benchmarks include VQA-RAD, SLAKE, and PathVQA for medical VQA (Lau et al., 2018; Liu et al., 2021; He et al., 2021), MIMIC-CXR for image-report modeling (Johnson et al., 2019), and MS-CXR and RadGraph for grounding and report-structured supervision (Boecking et al., 2022; Jain et al., 2021). More recent datasets have expanded scale and instruction style, including PMC-VQA, LLaVA-Med, HuatuoGPT-Vision, and MedThinkVQA (Zhang et al., 2024; Li et al., 2023; Chen et al., 2024; Yao et al., 2026b). These resources substantially improve medical interpretation and image–text alignment, but they rarely require the model to generate a visual teaching aid as part of a patient-facing response. MediSketch extends this line of work from clinician-facing interpretation to visually grounded patient education.

Patient education and lay-language generation. Patient-facing medical NLP has largely focused on text-only support, including simplification, lay

²Throughout the paper, we use *parallel sampling*, *sequential revision*, and *hybrid search* as canonical names, abbreviated as *Parallel*, *Sequential*, and *Hybrid*, respectively. We occasionally use *best-of- N* , *self-reflection/self-revision*, and *hybrid reflection* interchangeably with these three strategies, respectively. *Direct* denotes direct answering.

summarization, jargon explanation, and patient question answering (Ondov et al., 2022; Welivita and Pu, 2023; Chen et al., 2018; Yao et al., 2024; Tran et al., 2025; Cai et al., 2023; Kwon et al., 2022; Jang et al., 2025; Yao et al., 2025b). In radiology, prior work has translated reports into plain language, simplified terminology with consumer health vocabularies, and studied patient-centered report summarization (Qenam et al., 2017; Lyu et al., 2023; Park et al., 2024; Sunshine et al., 2025). Shared evaluations of biomedical lay summarization have also expanded this direction (Xiao et al., 2025). This literature has improved the accessibility of medical information, but the dominant response format remains text-only. Inspired by recent work on patient education (Yao et al., 2026a), our work studies a complementary setting in which the response itself must be multimodal, with a focus on how test-time scaling can improve its generation.

Test-time scaling. Recent work shows that additional inference-time compute can improve model performance through best-of- N sampling, self-refinement, critique-and-revision loops, and hybrid search (Snell et al., 2024; Wang et al., 2023; Madaan et al., 2023; Yao et al., 2023, 2025a; Maram et al., 2025). At the same time, recent analysis suggests that test-time scaling is limited without strong verification signals (Setlur et al., 2025), and self-refinement can be weakened by self-bias (Xu et al., 2024). In multimodal settings, the problem is further complicated by hallucination and weak cross-modal self-correction (Lee et al., 2024; Bai et al., 2025). In MediSketch, each candidate is an image–text educational bundle, and revision must improve both visual grounding and text-image consistency. This makes multimodal critique central to test-time scaling and motivates our comparison under a fixed generation budget cap.

3 Task and Evaluation Setup

Task definition. MediSketch studies a radiology-focused multimodal form of patient education in which the system must do more than produce a fluent answer. Each example contains radiology images, report findings, and a patient question, and the agent must return a complete image–text response bundle: an educational image together with a grounded plain-language explanation. The image is part of the answer itself rather than a decorative add-on. It should help the patient locate, compare, or distinguish the clinically relevant finding refer-

enced in the case. This differs from standard medical visual question-answering, where the model typically interprets the input and provides answers in text only. In MediSketch, success depends on whether the model can turn medical evidence into a patient-usable visual explanation. Appendix A gives the full task specification.

Protocol and data. We instantiate MediSketch with the tool-augmented protocol in Figure 1. Given the case context, the agent generates drawing instructions, calls a shared image tool to produce an annotated educational image, and then writes a patient-facing explanation conditioned on that image. This shared interface keeps the visual-generation protocol fixed across backbones and makes the intended visual explanation explicit for scoring and revision. Our 100 cases are built from the MedThinkVQA test split (Yao et al., 2026b; Kim et al., 2025; Wang et al., 2024), which is derived from peer-reviewed Eurorad teaching files and spans diverse radiology subspecialties, modalities, and anatomical categories. With clinician input, we formulate patient questions for cases where visual grounding provides substantive rather than merely cosmetic value, such as helping patients understand where reported findings appear, how visually distinct findings relate to the report, or what has changed over time. Additional details are provided in Appendix B.

Rubric and judge. We evaluate each image–text response bundle using a clinician-grounded checklist rubric and a rubric-following multimodal judge. A strong multimodal educational response must satisfy several constraints at once: the image should address the patient’s question and remain medically faithful, the text should be understandable and safe, and the two modalities should be consistent with each other. We therefore use a clinician-grounded checklist rubric that evaluates medical fidelity, image–text consistency, patient-facing usefulness, and visual quality. Final scoring is performed by GPT-5.4 as the primary rubric-following multimodal judge (Arora et al., 2025; Gu et al., 2026), using structured checklist execution rather than a free-form holistic preference score. Appendix C provides the full rubric, judge design, and radiologist validation for GPT-5.4 and an additional open-weight judge, Qwen3.5-397B-A17B.

4 Test-Time Scaling Strategies

We next study test-time scaling to improve MediSketch beyond a single pass. Some visual and cross-modal errors appear only after the image is rendered and paired with text, so additional image-tool calls can explore alternatives or repair the response without retraining. Because these calls are costly, we study finite-budget allocation across exploration, revision, and stopping, complementing fine-tuning, prompting, and grounding. Using Direct answering as the non-scaling baseline, we compare three scaling strategies: **1. Parallel sampling.** In the parallel strategy, the model independently generates multiple complete image-text bundles and then selects the best one using rubric-based scoring. Its main strength is exploration across different explanatory framings. Its limitation is that it cannot incrementally improve a promising but imperfect candidate. **2. Sequential revision.** In the sequential strategy, the model generates one bundle, critiques it, and then revises it over subsequent steps. Each revision regenerates the educational image and produces a revised final explanation for it, so improvement must occur jointly in the visual aid and the text. This strategy can strengthen partially correct solutions, but it depends heavily on the quality of the critique. It can fail when the model incorrectly judges an imperfect bundle as already sufficient. **3. Hybrid search.** Hybrid search combines exploration and revision. At each step, the model generates multiple bundles, selects the strongest candidate, and then continues sequential revision from that candidate into the next step. This strategy is well matched to MediSketch because the task benefits from both trying alternative visual framings and refining image-text consistency within a promising candidate. The three strategies are illustrated on the right side of Figure 1. Additional strategy definitions and the bundle-level search setup are provided in Appendix D.2, and the critic and judge prompts are provided in Appendices C.11 and C.12, respectively.

5 Experimental Setup

Models and unified interface. We evaluate a compact but diverse set of multimodal-capable backbone models spanning proprietary and open-weight systems, including GPT-5-nano, GPT-5-mini, GPT-5.2, Qwen3.5-9B, and Qwen3.5-27B. All models are tested under the same agent protocol (see Section 3). This shared process makes the compari-

son fairer by keeping the image tool fixed across models, so the main differences come from how well models critique image-text bundles, propose effective revisions, and generate grounded patient-facing explanations from the resulting image. For the image generation/edit tool, we used OpenAI GPT Image 1.5 in all experiments. Model details are provided in Appendix D.1, the shared agent and image-tool protocol is described in Appendix B.2, and the drawing and response prompts are provided in Appendix C.8 and Appendix C.9.

Budget cap, realized usage, and strategy instantiation. To compare different test-time scaling strategies, we first need to define a shared budget. In MediSketch, the primary resource is image generation, because each additional attempt requires a new drawing-instruction call and a new educational image, which dominates the cost. We therefore compare methods under a matched maximum image-generation budget, defined as the maximum number of image-producing attempts allowed per example. Token usage is reported as a secondary efficiency measure, but the image-generation budget is the primary fairness constraint because it more directly captures the practical cost of multimodal patient assistance.

In our main experiments, we set the maximum image-generation budget cap to 12, a practical cap large enough to test nontrivial Parallel, Sequential, and Hybrid strategies while avoiding excessive image-generation cost. Appendix D.3 provides further analysis and motivation for this choice.

Under this cap, different strategies use the budget differently. Direct answering, the non-scaling baseline, generates a single image-text bundle in one pass. Parallel sampling uses the full budget by generating 12 independent bundles and selecting the best one. Sequential revision starts from one bundle and repeatedly critiques and revises it, with early stopping when the critic returns a *pass* judgement, indicating that the current image-text bundle satisfies the criteria and does not require further image or response edits. Hybrid search combines exploration and revision by generating multiple candidates at each stage and then continuing from the strongest one; like sequential revision, it can stop early when the critic returns *pass*. We write a hybrid setting as $a \times b$, where a is the number of candidates generated per stage and b is the maximum number of stages. For example, 2×6 generates 2 candidates at each stage for up to 6 stages, so the total number of image-generation

TTS	Model	IFC	TMC	SAFE	BASE	USE	ALIGN	COMM	IQA	SIMPLE	Avg (/29)	#Img
Direct	qwen3.5-9b	0.68 ± 0.03	1.10 ± 0.10	2.04 ± 0.18	0.44 ± 0.05	3.74 ± 0.09	1.89 ± 0.15	1.72 ± 0.05	2.82 ± 0.05	1.96 ± 0.05	16.40 ± 0.28	1.00 ± 0.00
	qwen3.5-27b	0.84 ± 0.14	1.25 ± 0.36	2.32 ± 0.27	0.31 ± 0.07	1.00 ± 0.23	2.23 ± 0.35	2.00 ± 0.00	2.87 ± 0.05	1.97 ± 0.01	18.44 ± 1.27	1.00 ± 0.00
	gpt5-nano	0.49 ± 0.28	1.85 ± 0.36	2.57 ± 0.36	0.58 ± 0.04	2.56 ± 0.22	2.29 ± 0.35	2.00 ± 0.00	2.76 ± 0.08	1.98 ± 0.04	17.08 ± 1.70	1.00 ± 0.00
	gpt5-mini	1.52 ± 0.08	2.06 ± 0.04	3.23 ± 0.09	0.48 ± 0.06	5.30 ± 0.09	2.69 ± 0.05	2.00 ± 0.00	2.83 ± 0.07	1.84 ± 0.03	21.96 ± 0.29	1.00 ± 0.00
	gpt5.2	1.36 ± 0.14	2.41 ± 0.07	3.49 ± 0.14	0.37 ± 0.01	5.23 ± 0.17	2.97 ± 0.14	1.99 ± 0.01	2.51 ± 0.09	1.32 ± 0.03	21.64 ± 0.70	1.00 ± 0.00
Parallel	qwen3.5-9b	0.33 ± 0.31	1.00 ± 0.35	1.53 ± 0.23	0.53 ± 0.12	3.67 ± 0.31	1.67 ± 0.46	1.80 ± 0.00	2.80 ± 0.00	1.93 ± 0.12	15.27 ± 0.12	12.00 ± 0.00
	qwen3.5-27b	0.80 ± 0.20	1.20 ± 0.72	2.27 ± 0.23	0.33 ± 0.12	5.07 ± 0.31	2.40 ± 0.20	2.00 ± 0.00	3.00 ± 0.00	2.00 ± 0.00	19.07 ± 0.99	12.00 ± 0.00
	gpt5-nano	0.40 ± 0.53	1.73 ± 0.58	2.27 ± 0.81	0.60 ± 0.00	2.73 ± 0.42	2.40 ± 0.53	2.00 ± 0.00	2.73 ± 0.12	2.00 ± 0.00	16.87 ± 2.20	12.00 ± 0.00
	gpt5-mini	1.33 ± 0.42	1.93 ± 0.23	2.87 ± 0.70	0.20 ± 0.20	5.73 ± 0.12	2.27 ± 0.70	2.00 ± 0.00	2.87 ± 0.12	1.93 ± 0.12	21.13 ± 1.97	12.00 ± 0.00
	gpt5.2	1.33 ± 0.50	2.47 ± 0.31	3.47 ± 0.58	0.53 ± 0.12	5.80 ± 0.20	3.07 ± 0.23	2.00 ± 0.00	2.80 ± 0.00	1.67 ± 0.31	23.13 ± 1.86	12.00 ± 0.00
Sequential	qwen3.5-9b	1.33 ± 0.81	1.60 ± 0.20	2.53 ± 0.99	0.47 ± 0.50	4.27 ± 0.90	2.20 ± 0.53	1.60 ± 0.00	3.00 ± 0.00	2.00 ± 0.00	19.00 ± 2.80	1.47 ± 0.12
	qwen3.5-27b	1.27 ± 0.61	1.80 ± 0.20	2.60 ± 0.60	0.07 ± 0.12	4.60 ± 0.35	2.53 ± 0.42	2.00 ± 0.00	2.80 ± 0.20	2.00 ± 0.00	19.67 ± 1.94	3.00 ± 0.35
	gpt5-nano	0.53 ± 0.76	1.93 ± 0.46	2.93 ± 0.81	0.33 ± 0.12	3.47 ± 0.81	2.87 ± 0.23	2.00 ± 0.00	2.73 ± 0.31	1.67 ± 0.31	18.47 ± 2.80	1.33 ± 0.31
	gpt5-mini	0.87 ± 0.42	1.60 ± 0.40	2.27 ± 0.61	0.27 ± 0.12	5.67 ± 0.31	2.20 ± 0.87	2.00 ± 0.00	2.87 ± 0.12	1.80 ± 0.20	19.53 ± 2.64	1.27 ± 0.23
	gpt5.2	2.80 ± 0.35	3.00 ± 0.00	4.87 ± 0.23	0.53 ± 0.12	5.93 ± 0.12	4.00 ± 0.00	2.00 ± 0.00	2.93 ± 0.12	1.33 ± 0.31	27.40 ± 0.92	5.20 ± 0.53
Hybrid	qwen3.5-9b	0.60 ± 0.00	0.93 ± 0.23	2.20 ± 0.00	0.47 ± 0.12	4.40 ± 0.53	1.60 ± 0.40	1.87 ± 0.12	2.87 ± 0.23	2.00 ± 0.00	16.93 ± 0.31	3.33 ± 0.61
	qwen3.5-27b	1.07 ± 0.64	1.47 ± 0.12	2.67 ± 0.50	0.20 ± 0.20	5.13 ± 0.23	2.53 ± 0.12	2.00 ± 0.00	3.00 ± 0.00	2.00 ± 0.00	20.07 ± 1.70	4.13 ± 1.22
	gpt5-nano	1.07 ± 0.23	2.47 ± 0.12	3.20 ± 0.20	0.40 ± 0.00	3.13 ± 1.14	2.47 ± 0.50	2.00 ± 0.00	2.73 ± 0.23	1.73 ± 0.46	19.20 ± 1.04	2.40 ± 0.69
	gpt5-mini	1.27 ± 0.70	2.13 ± 0.12	3.27 ± 0.31	0.33 ± 0.12	5.67 ± 0.42	3.07 ± 0.31	2.00 ± 0.00	2.80 ± 0.20	1.80 ± 0.20	22.33 ± 1.50	2.13 ± 0.23
	gpt5.2	2.80 ± 0.35	2.87 ± 0.23	4.87 ± 0.23	0.80 ± 0.20	6.00 ± 0.00	3.93 ± 0.12	2.00 ± 0.00	3.00 ± 0.00	1.80 ± 0.20	28.07 ± 0.92	6.27 ± 0.83

Table 1: **Main results under a shared image-generation budget cap.** Direct, Parallel, Sequential, and Hybrid denote direct answering, parallel sampling, sequential revision, and hybrid search, respectively. For a standardized main comparison, we report one representative Hybrid setting, Hybrid 2×6 ; additional Hybrid allocations are analyzed in the broader budget sweeps in Figures 2 and 3. Values are mean $\pm$ standard deviation; best and second-best values in each column are bolded and underlined, respectively. The nine rubric dimensions cover medical grounding, patient-facing usefulness, and visual execution; see Appendix C.1. Avg is the total score across all nine dimensions (maximum 29), and #Img is the realized average number of edited images per case, which can fall below the budget cap when critique-based methods stop early.

attempts is capped at 12. Parallel always uses the full cap, whereas sequential and hybrid methods may stop early when the critic accepts the current bundle. We therefore report both final quality and the realized average number of edited images #Img in Table 1, so readers can distinguish the maximum allowed budget from the actual image-generation usage. See search-strategy and budget-instantiation details in Appendices D.2 and D.3, respectively.

Evaluation metrics. We evaluate each image-text bundle using nine rubric dimensions organized into three broader categories. The first category, *medical grounding, safety, and source integrity*, covers Image Factual Correctness (IFC), Text Medical Correctness (TMC), Safety-First Visualization & Content (SAFE), and Base Image Usage Integrity (BASE). These metrics ask whether the visual aid and final response stay faithful to the report, avoid unsupported inference, handle uncertainty conservatively, and preserve the integrity of any original medical image used as the visual substrate.

The second category, *patient-facing usefulness and image-response alignment*, covers Patient-Facing Usefulness & Coverage (USE), Clarity & Correct Image-Response Alignment (ALIGN), and Communication Quality (COMM). These metrics focus on whether the bundle actually helps the patient understand the question they asked, whether the response correctly refers to what appears in the image, and whether the language remains clear,

coherent, and professional for a lay audience.

The third category, *visual execution, simplicity, and technical quality*, covers Image QA (IQA) and Simplicity & Non-Escalation (SIMPLE). These metrics focus on the internal consistency and readability of the generated image, including labels and artifacts, while also ensuring the visual aid remains minimal and avoids unnecessary visual clutter or long explanatory text within the image. Our primary summary metric is the total score across all nine rubrics, which lets us compare overall multimodal educational quality while still preserving interpretable sub-dimensions. Appendix D.5 reports a human-expert SAFE error audit; additional rubric, judge I/O, and scoring details are provided in Appendices C.10, C.2, and D.4.

6 Experimental Results

Main Results Table 1 shows three main findings. First, direct answering is relatively weak across models, with the strongest result reaching 21.96 ± 0.29 . Second, additional budget does not help by default: Parallel always uses the full image budget, but often improves little over Direct and sometimes performs worse. Third, budget allocation matters more than budget size alone: strategies that include critique and revision can outperform simple full-budget sampling, but their gains depend on how effectively the model uses the budget.

A strategy-level reading of the table makes

the pattern clearer. Parallel results sharpen the second finding: simply spending the full budget through best-of- N sampling is often inefficient in this setting. For example, GPT-5-mini drops from 21.96 ± 0.29 under Direct to 21.13 ± 1.97 under Parallel, despite using 12 images instead of 1. GPT-5.2 improves from 21.64 ± 0.70 to 23.13 ± 1.86 under Parallel, but this gain is small relative to the budget increase and remains well below the critique-based strategies, which reach 27.40 ± 0.92 with Sequential and 28.07 ± 0.92 with Hybrid.

Compared with Parallel, Sequential and Hybrid show that better budget allocation can matter more than the nominal budget size. Sequential improves over Direct for 4 of 5 backbones, with the clearest gain on GPT-5.2, where the score rises from 21.64 ± 0.70 to 27.40 ± 0.92 . It also gives the best result for Qwen3.5-9B, reaching 19.00 ± 2.80 compared with Hybrid at 16.93 ± 0.31 . Hybrid is strongest overall, achieving the best model-wise mean score for 4 of 5 backbones. Its clearest gain appears on GPT-5.2, where it improves over Direct from 21.64 ± 0.70 to 28.07 ± 0.92 . The main exception is Qwen3.5-9B, where Sequential reaches 19.00 ± 2.80 and outperforms Hybrid at 16.93 ± 0.31 . Taken together, these results suggest that test-time scaling helps most when the budget is used for critique-guided revision or a combination of exploration and revision, rather than for independent sampling alone; Appendix D.5 provides a complementary SAFE audit.

Budget and Critique We now look beyond the fixed comparison in Table 1 and examine the finer-grained budget sweeps in Figures 2 and 3. Table 1 reports results under one shared budget cap and uses a single representative Hybrid configuration, Hybrid 2×6 , to keep the main comparison uniform across backbones (see Appendix D.3). By contrast, Figures 2 and 3 provide a more detailed view of how different sequential and hybrid configurations use the available budget, and how their average scores change with realized budget use.

Specifically, across models, neither more edited images nor more tokens guarantees better performance. A key reason is premature stopping: critique-based methods often stop far below the budget cap, suggesting that many models treat an imperfect bundle as already sufficient. This pattern is mild for GPT-5.2, whose higher-budget hybrid variants continue improving, but much stronger for weaker critics such as Qwen3.5-27B, whose gains

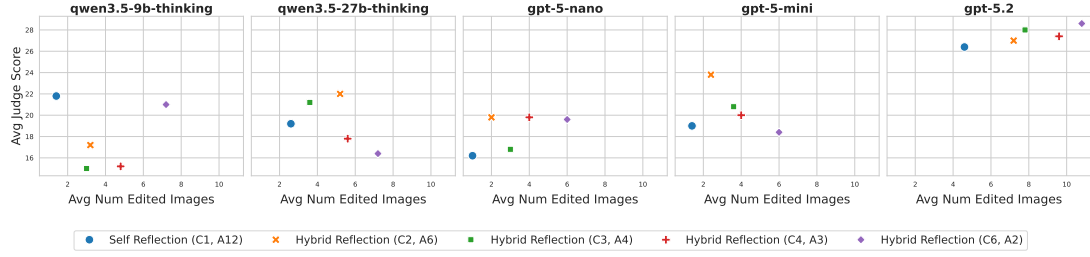
are smaller and less reliable. Together, these results suggest that the central challenge is not only producing more candidates, but also recognizing when a multimodal bundle still needs revision; qualitative examples of premature stopping and hybrid recovery are provided in E. We next unpack this summary by examining the cost–quality frontier in Figure 2 and the early-stopping behavior behind it.

Figure 2 shows that cost and quality are only loosely coupled. Different hybrid allocations with similar realized image counts or token costs can yield noticeably different scores, and some higher-cost settings underperform lower-cost ones for weaker models. This is especially visible for Qwen3.5-9B, Qwen3.5-27B, and GPT-5-mini, where spending more budget does not reliably move the model to a better part of the quality frontier. The main issue, therefore, is not the size of the budget in isolation, but how effectively it is used.

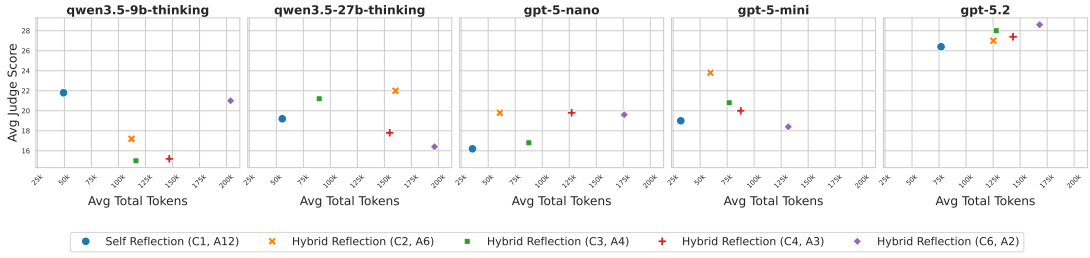
The realized #Img values in Table 1 show where this inefficiency comes from. Many critique-based runs stop after only one or two image edits on average, even though the cap is much larger. This pattern is consistent with weak multimodal self-critique: current models often fail to recognize that an imperfect bundle still needs revision. In this setting, critique must assess not only wording, but also whether the image highlights the right evidence, whether the text correctly interprets that image, and whether the full bundle actually helps answer the patient’s question. The distinction between the final judge and the in-loop critique signal is described in Appendix C.5 and C.6.

Figure 3 makes the model-level contrast concrete. For GPT-5.2, the self-reflection curve rises to 26.4 in the main comparison, while broader hybrid budget-sweep variants continue higher in the figure, reaching 27.0, 27.4, 28.0, and 28.6 at larger realized budgets. This indicates that GPT-5.2 can still convert additional exploration and revision into higher-quality bundles. Qwen3.5-27B behaves differently: its self-reflection result reaches 19.2 in the main comparison and then largely plateaus, while hybrid variants range from weaker settings below self-reflection to stronger settings that reach 22.0. This comparison shows that the ability to exploit additional budget varies across backbones.

These patterns help explain why Hybrid works well in MediSketch. Pure Sequential depends on the model noticing when its current path is still inadequate. Pure Parallel avoids early commitment, but it cannot refine a promising partial solution. Hy-



(a) Average judge score versus the number of edited images.



(b) Average judge score versus total token usage.

Figure 2: **Quality is only weakly coupled to generation cost.** The top panel plots average rubric-based bundle score against the number of edited images, and the bottom panel plots the same score against total token usage. Self-reflection denotes the Sequential strategy. Across strategies, using more image edits or more tokens does not reliably move models to a better quality frontier. A major bottleneck is therefore not budget availability itself, but allocating it effectively through multimodal critique and revision.

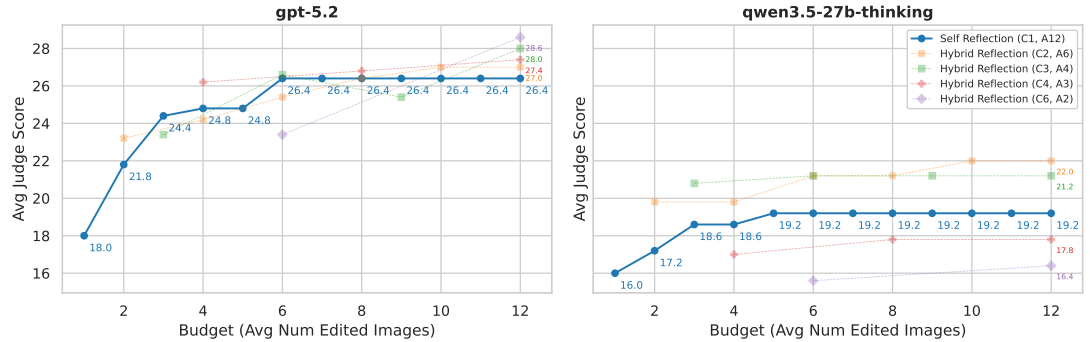


Figure 3: **Test-time scaling depends on how well a model can use extra multimodal budget.** We plot average rubric-based bundle score against the average number of edited images for two representative models, *gpt-5.2* and *qwen3.5-27b-thinking*, under different hybrid configurations. *gpt-5.2*, especially with hybrid search, shows a stronger and more stable scaling trend: additional budget is more consistently converted into better image-text bundles. By contrast, *qwen3.5-27b-thinking* fluctuates more across nearby settings and shows smaller gains from additional budget. This pattern suggests that the limiting factor is not budget alone, but the strength of multimodal self-evaluation needed to turn extra exploration and revision into higher-quality patient-facing outputs.

brid relaxes both assumptions: it maintains some exploration while still allowing the selected candidate to improve over time. In tasks that require both alternative visual framings and careful image-text consistency, this allocation is often the most effective use of additional compute. Representative qualitative examples are provided in Appendix E.

7 Conclusion

MediSketch exposes an evaluation gap in patient-facing NLP. Prior work has advanced simplification, explanation, and lay-language generation (On-

dov et al., 2022; Welivita and Pu, 2023; Lyu et al., 2023; Yao et al., 2024; Tran et al., 2025), but mostly under text-only output targets. MediSketch instead evaluates complete image-text bundles, where success depends on medical grounding, image-response alignment, patient usefulness, and visual execution.

Within this setting, we evaluate test-time scaling on our clinician-curated 100-case testbed. Direct answering remains weak, and naive best-of- N search often spends the full budget inefficiently. Hybrid search performs strongest overall because it

combines exploration with revision. Overall, budget size alone does not determine performance; gains depend on how it is allocated and whether the model recognizes when revision is needed.

These findings suggest that multimodal self-evaluation is a major, but not exclusive, bottleneck. Models must recognize when an image–text bundle is incomplete or misaligned and continue improving it. Future progress will require better multimodal critics, stronger verification signals, and inference procedures that reduce premature stopping and support grounded revision (Setlur et al., 2025; Lee et al., 2024; Bai et al., 2025).

Limitations

Our study has several methodological limitations. Final evaluation still relies primarily on a rubric-following LLM judge, although we complement it with radiologist preference validation and a non-GPT judge check in Appendix C.3, as well as a six-participant lay preference pilot in Appendix C.4. Rubric-based scoring is structured and scalable but remains an approximation of real patient interpretation and clinical appropriateness; sharing the rubric between critique and evaluation may also overstate the reported gains from Sequential and Hybrid search.

The six-participant lay pilot provides complementary evidence of perceived patient-facing preference, but it does not assess comprehension, recall, or calibrated confidence or substitute for evaluation with actual patients. Its overall A/B judgments also do not directly validate individual rubric dimensions or the aggregate score.

Evaluator-family alignment presents an additional limitation. Because the GPT-5.2 backbone and GPT-5.4 judge are from the same model family, evaluator preference may partly contribute to GPT-5.2’s reported cross-model advantage (Panickssery et al., 2024). This concern is less likely to fully explain within-backbone strategy differences, where the backbone and judge are fixed across strategies.

The current evaluation uses a focused set of 100 clinician-curated cases covering multiple radiology subspecialties, imaging modalities, and anatomical categories. Our conclusions are therefore scoped to the current MediSketch testbed, and broader validation across additional institutions, imaging modalities, and patient populations is an important direction for future work.

The source images and reports were created for

professional education rather than lay communication, and our task construction uses clinician input to focus on questions where visual support is helpful. This curation limits the scope of our conclusions because the benchmark may not reflect the broader distribution of real patient-portal inquiries, which may be ambiguous, emotional, poorly structured, or unsuitable for visual explanation.

The current task also focuses on radiology, where visual explanation is especially natural, so transfer to other patient-education settings such as laboratory interpretation, medication counseling, or longitudinal chronic-care support remains open.

In addition, our setup uses the proprietary GPT Image 1.5 as a shared image-generation tool rather than native multimodal output models. Sharing the backend keeps the comparison controlled, but makes the results conditional on it and, together with the use of proprietary backbone models, limits exact end-to-end reproducibility; the setup may also not reflect future systems with tighter image–text integration or different generation interfaces.

Ethical Considerations

These limitations connect directly to patient-facing safety risks. In this setting, errors are not merely cosmetic. A visual explanation can aid understanding, but an incorrect annotation may appear to localize a finding with clinical certainty, making errors more convincing than text-only mistakes. This risk is especially relevant to patients with low health or visual literacy or limited ability to verify radiology images. MediSketch should therefore not be interpreted as deployment-ready, and any patient-facing use would require clinician oversight and patient-comprehension testing.

The benchmark is also English-only and does not evaluate non-English or culturally different explanation needs or accessibility for visually impaired or color-blind users.

We therefore view MediSketch as an evaluation resource rather than a clinical recommendation engine. Future work should include larger controlled studies with patients comparing image–text bundles with text-only responses on comprehension, recall, and calibrated confidence, alongside direct expert validation of dimension-level and aggregate rubric scores, uncertainty-aware response strategies, and safeguards that explicitly limit overconfident visual or textual explanations in high-stakes settings.

Acknowledgments

We used ChatGPT (OpenAI) throughout the manuscript for language polishing. All AI-generated text was reviewed and edited by the authors, who verified its accuracy and take full responsibility for the content, claims, and analyses. All of the experiments were aided by OpenAI Codex, where the codes and results were manually reviewed by the authors.

References

- Mohammad Alarifi, Timothy Patrick, Abdulrahman Jabour, Min Wu, and Jake Luo. 2021. Understanding patient needs and gaps in radiology reports through online discussion forum analysis. *Insights into imaging*, 12(1):50.
- Anthropic. 2026. Advancing claude in healthcare and the life sciences. <https://www.anthropic.com/news/healthcare-life-sciences>.
- Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñero-Candela, Foivos Tsimpouras, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. 2025. [Healthbench: Evaluating large language models towards improved human health](#). Preprint, arXiv:2505.08775.
- Serhat Aydin, Mert Karabacak, Victoria Vlachos, and Konstantinos Margetis. 2024. Large language models in patient education: a scoping review of applications in medicine. *Frontiers in medicine*, 11:1477898.
- Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2025. [Hallucination of multimodal large language models: A survey](#). Preprint, arXiv:2404.18930.
- David W. Baker, Mark V. Williams, Ruth M. Parker, Julie A. Gazmararian, and Joanne Nurss. 1999. [Development of a brief test to measure functional health literacy](#). *Patient Education and Counseling*, 38(1):33–42.
- Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. 2022. Making the most of text semantics to improve biomedical vision–language processing. In *Computer Vision – ECCV 2022*, pages 1–21, Cham. Springer Nature Switzerland.
- Pengshan Cai, Zonghai Yao, Fei Liu, Dakuo Wang, Meghan Reilly, Huixue Zhou, Lingxi Li, Yi Cao, Alok Kapoor, Adarsha Bajracharya, Dan Berlowitz, and Hong Yu. 2023. [PaniniQA: Enhancing patient education through interactive question answering](#). *Transactions of the Association for Computational Linguistics*, 11:1518–1536.
- Elizabeth B Card, Julian Lejbman, Eo Trueblood, Brittany C Bennett, Brian P Dunham, Amanda J Swain, and Horace M Delisser. 2022. Drawing for visual communication in medicine: a novel pilot course for senior medical students. *Journal of surgical education*, 79(2):389–396.
- Jinying Chen, Emily Druhl, Balaji Polepalli Ramesh, Thomas K Houston, Cynthia A Brandt, Donna M Zulman, Varsha G Vimalananda, Samir Malkani, and Hong Yu. 2018. A natural language processing system that links medical terms in electronic health record notes to lay definitions: system development using physician reviews. *Journal of medical Internet research*, 20(1):e26.
- Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, and Benyou Wang. 2024. [Towards injecting medical visual knowledge into multimodal LLMs at scale](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7346–7370, Miami, Florida, USA. Association for Computational Linguistics.
- Tom Delbanco, Jan Walker, Sigall K Bell, Jonathan D Darer, Joann G Elmore, Nadine Farag, Henry J Feldman, Roanne Mejilla, Long Ngo, James D Ralston, and 1 others. 2012. Inviting patients to read their doctors’ notes: a quasi-experimental study and a look ahead. *Annals of internal medicine*, 157(7):461–470.
- European Society of Radiology. 2026. [Eurorad](#). Accessed: May 25, 2026.
- Elisa Galmarini, Laura Marciano, and Peter Johannes Schulz. 2024. The effectiveness of visual-based interventions on health literacy in health care: a systematic review and meta-analysis. *BMC Health Services Research*, 24(1):718.
- Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Zhouchi Lin, Bowen Zhang, Lionel Ni, Wen Gao, Yuanzhuo Wang, and Jian Guo. 2026. [A survey on llm-as-a-judge](#). *The Innovation*, 7(6):101253.
- Xuehai He, Zhuo Cai, Wenlan Wei, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2021. Towards visual question answering on pathology images. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 708–718.
- Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis P. Langlotz, and Pranav Rajpurkar. 2021. [Radgraph: Extracting clinical entities and relations from radiology reports](#). Preprint, arXiv:2106.14463.

- Won Seok Jang, Hieu Tran, Manav Shaileshkumar Mistry, Sai Kiran Gandluri, Yifan Zhang, Sharmin Sultana, Sunjae Kwon, Yuan Zhang, Zonghai Yao, and Hong Yu. 2025. [Chatbot to help patients understand their health](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6598–6627, Suzhou, China. Association for Computational Linguistics.
- Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.
- Su Hwan Kim, Severin Schramm, Lisa C Adams, Rickmer Braren, Keno K Bressen, Matthias Keicher, Paul-Sören Platzek, Karolin Johanna Paprottka, Claus Zimmer, Dennis M Hedderich, and 1 others. 2025. Benchmarking the diagnostic performance of open source llms in 1933 eurorad case reports. *NPJ digital medicine*, 8(1):97.
- Sunjae Kwon, Zonghai Yao, Harmon Jordan, David Levy, Brian Corner, and Hong Yu. 2022. Medjex: A medical jargon extraction model with wiki’s hyper-link span and contextualized masked language model score. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11733–11751.
- Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. 2024. Lisa: Reasoning segmentation via large language model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9579–9589. IEEE.
- Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):180251.
- Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Minjoon Seo. 2024. Volcano: mitigating multimodal hallucination through self-feedback guided revision. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 391–404.
- Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in neural information processing systems*, 36:28541–28564.
- Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1650–1654. IEEE.
- Qing Lyu, Josh Tan, Michael E Zapadka, Janardhana Ponnatapura, Chuang Niu, Kyle J Myers, Ge Wang, and Christopher T Whitlow. 2023. Translating radiology reports into plain language using chatgpt and gpt-4 with prompt learning: results, limitations, and potential. *Visual computing for industry, biomedicine, and art*, 6(1):9.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: iterative refinement with self-feedback. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS ’23*, Red Hook, NY, USA. Curran Associates Inc.
- Durga Prasad Maram, Dhruvin Gandhi, Zonghai Yao, Gayathri Akkinapalli, Franck Dernoncourt, Yu Wang, Ryan A. Rossi, and Nesreen K. Ahmed. 2025. [Iterative critique-refine framework for enhancing llm personalization](#). Preprint, arXiv:2510.24469.
- Teresa Martin-Carreras, Tessa S Cook, and Charles E Kahn Jr. 2019. Readability of radiology reports: implications for patient-centered care. *Clinical imaging*, 54:116–120.
- William A Mehan Jr, James A Brink, and Joshua A Hirsch. 2021. 21st century cures act: patient-facing implications of information blocking. *Journal of the American College of Radiology*, 18(7):1012–1016.
- Evan C Norris, Ciarra Halaska, Peter B Sachs, Chen-Tan Lin, Kate Sanfilippo, and Justin M Honce. 2022. Understanding patient experiences, opinions, and actions taken after viewing their own radiology images online: web-based survey. *JMIR Formative Research*, 6(4):e29496.
- Brian Ondov, Kush Attal, and Dina Demner-Fushman. 2022. A survey of automated methods for biomedical text simplification. *Journal of the American Medical Informatics Association*, 29(11):1976–1988.
- OpenAI. 2026. [Introducing chatgpt health.](https://openai.com/index/introducing-chatgpt-health/) <https://openai.com/index/introducing-chatgpt-health/>.
- Arjun Panickssery, Samuel R Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems*, 37:68772–68802.
- Jiwoo Park, Kangrok Oh, Kyunghwa Han, and Young Han Lee. 2024. Patient-centered radiology reports with generative artificial intelligence: adding value to radiology reporting. *Scientific Reports*, 14(1):13218.
- Jordan R Pollock, Skye A Buckner Petty, John J Schmitz, Jacob Varner, Allie M Metcalfe, and Nelly Tan. 2024. Patient access of their radiology reports before and after implementation of 21st century

- cures act information-blocking provisions at a large multicampus health system. *American Journal of Roentgenology*, 222(6):e2330343.
- Basel Qenam, Tae Youn Kim, Mark J Carroll, and Michael Hogarth. 2017. Text simplification using consumer health vocabulary to generate patient-centered radiology reporting: translation and evaluation. *Journal of medical Internet research*, 19(12):e417.
- Danielle Schubbe, Peter Scalia, Renata W Yen, Catherine H Saunders, Sarah Cohen, Glyn Elwyn, Maria van den Muijsenbergh, and Marie-Anne Durand. 2020. Using pictures to convey health information: A systematic review and meta-analysis of the effects on patient and consumer health behaviors and outcomes. *Patient education and counseling*, 103(10):1935–1960.
- Amrith Setlur, Nived Rajaraman, Sergey Levine, and Aviral Kumar. 2025. [Scaling test-time compute without verification or rl is suboptimal](#). *Preprint*, arXiv:2502.12118.
- Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. [Scaling llm test-time compute optimally can be more effective than scaling model parameters](#). *Preprint*, arXiv:2408.03314.
- Alexis Sunshine, Grace H Honce, Andrew L Callen, David A Zander, Jody L Tanabe, Samantha L Pisani Petrucci, Chen-Tan Lin, and Justin M Honce. 2025. Evaluating the quality and understandability of radiology report summaries generated by chatgpt: survey study. *JMIR formative research*, 9(1):e76097.
- Hieu Tran, Zonghai Yao, Won Seok Jang, Sharmin Sultana, Allen Chang, Yuan Zhang, and Hong Yu. 2025. Medreadctrl: Personalizing medical text generation with readability-controlled instruction learning. *medRxiv*.
- FAM Van der Mee, RPG Ottenheijm, EGS Gentry, JM Nobel, FM Zijta, JWL Cals, and J Jansen. 2025. The impact of different radiology report formats on patient information processing: a systematic review. *European Radiology*, 35(5):2644–2657.
- Nina S Vincoff, Matthew A Barish, and Gregory Grimaldi. 2022. The patient-friendly radiology report: history, evolution, challenges and opportunities. *Clinical Imaging*, 89:128–135.
- Jan Walker, Suzanne Leveille, Sigall Bell, Hannah Chmowitz, Zhiyong Dong, Joann G Elmore, Leonor Fernandez, Alan Fossa, Macda Gerard, Patricia Fitzgerald, Kendall Harcourt, Sara Jackson, Thomas H Payne, Jocelyn Perez, Hannah Shucard, Rebecca Stamez, Catherine DesRoches, and Tom Delbanco. 2019. [Opennotes after 7 years: Patient experiences with ongoing access to their clinicians’ outpatient visit notes](#). *J Med Internet Res*, 21(5):e13876.
- Junda Wang, Yujan Ting, Eric Z. Chen, Hieu Tran, Hong Yu, Weijing Huang, and Terrence Chen. 2024. [Semi-hvision: Enhancing medical multimodal models with a semi-human annotated dataset and fine-tuned instruction generation](#). *Preprint*, arXiv:2410.14948.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. [Self-consistency improves chain of thought reasoning in language models](#). *Preprint*, arXiv:2203.11171.
- Anuradha Welivita and Pearl Pu. 2023. A survey of consumer health question answering systems. *Ai Magazine*, 44(4):482–507.
- Chenghao Xiao, Kun Zhao, Xiao Wang, Siwei Wu, Sixing Yan, Tomas Goldsack, Sophia Ananiadou, Noura Al Moubayed, Liang Zhan, William K. Cheung, and Chenghua Lin. 2025. [Overview of the BioLaySumm 2025 shared task on lay summarization of biomedical research articles and radiology reports](#). In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 365–377, Vienna, Austria. Association for Computational Linguistics.
- Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Wang. 2024. Pride and prejudice: Llm amplifies self-bias in self-refinement. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15474–15492.
- Senqiao Yang, Tianyuan Qu, Xin Lai, Zhuotao Tian, Bohao Peng, Shu Liu, and Jiaya Jia. 2024. [Lisa++: An improved baseline for reasoning segmentation with large language model](#). *Preprint*, arXiv:2312.17240.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.
- Zonghai Yao, Nandyala Siddharth Kantu, Guanghao Wei, Hieu Tran, Zhangqi Duan, Sunjae Kwon, Zhichao Yang, and Hong Yu. 2024. Readme: Bridging medical jargon and lay understanding for patient education through data-centric nlp. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12609–12629.
- Zonghai Yao, Aditya Parashar, Huixue Zhou, Won Seok Jang, Feiyun Ouyang, Zhichao Yang, and Hong Yu. 2025a. Mcqg-srefine: Multiple choice question generation and evaluation with iterative self-critique, correction, and comparison feedback. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10728–10777.
- Zonghai Yao, Michael Sun, Won Seok Jang, Sunjae Kwon, Soie Kwon, and Hong Yu. 2025b. [DischargeSim: A simulation benchmark for educational](#)

- [doctor–patient communication at discharge](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10772–10798, Suzhou, China. Association for Computational Linguistics.
- Zonghai Yao, Zhipeng Tang, Chengtao Lin, Xiong Luo, Benlu Wang, Juncheng Huang, Chin Siang Ong, and Hong Yu. 2026a. [Rethinking patient education as multi-turn multi-modal interaction](#). *Preprint*, arXiv:2604.14656.
- Zonghai Yao, Benlu Wang, Yifan Zhang, Junda Wang, Iris Xia, Zhipeng Tang, Shuo Han, Feiyun Ouyang, Zhichao Yang, Arman Cohan, and Hong Yu. 2026b. [Medical thinking with multiple images](#). *Preprint*, arXiv:2604.16506.
- Zonghai Yao and Hong Yu. 2026. Llm-based multi-agent systems for clinical workflows: A survey of ai hospitals. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 45772–45793.
- Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. [Pmc-vqa: Visual instruction tuning for medical visual question answering](#). *Preprint*, arXiv:2305.10415.
- Pengfei Zhou, Xiaopeng Peng, Jiajun Song, Chuanhao Li, Zhaopan Xu, Yue Yang, Ziyao Guo, Hao Zhang, Yuqi Lin, Yefei He, Lirui Zhao, Shuo Liu, Tianhua Li, Yuxuan Xie, Xiaojun Chang, Yu Qiao, Wenqi Shao, and Kaipeng Zhang. 2025. Opening: A comprehensive benchmark for judging open-ended interleaved image-text generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 56–66.

A Task Details

A.1 Formal Task Definition

MediSketch defines a radiology-focused multi-modal patient-education task in which an agent must answer a patient question with a complete image–text response bundle. We focus on radiology because it is a visually intensive medical specialty in which patient understanding often depends on connecting report language to visible image evidence. A patient may need to know where a finding is, how it compares across views or modalities, or why a highlighted region matters in relation to the report. Radiology therefore provides a concrete and clinically motivated setting for studying image-grounded patient education.

In MediSketch, each example contains three inputs: (1) one or more radiology images, (2) report findings or a report summary, and (3) a patient question written to reflect a realistic patient-understanding need. The required output has two parts: (1) an educational image and (2) a grounded plain-language explanation text.

The educational image is not a generic illustration. It is expected to help the patient locate, compare, or distinguish the clinically relevant finding referenced in the report and question. The text must then explain the image in simple and medically faithful language. Relative to standard medical visual question answering, MediSketch requires the model not only to interpret the medical input, but also to externalize that interpretation in a patient-usable visual form.

Additional task examples and case studies are provided in Appendix E.

B Data and Protocol Details

B.1 Source Data and Case Construction

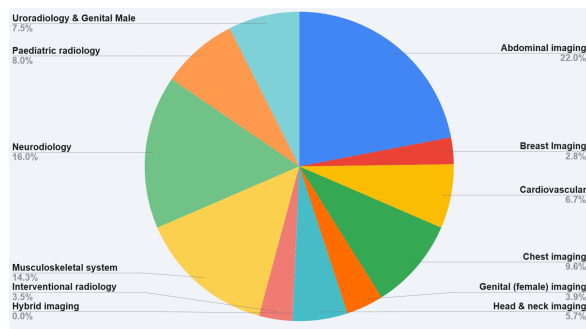


Figure 4: Distribution of radiology imaging categories.

MediSketch is built on the MedThinkVQA

benchmark (Yao et al., 2026b; Kim et al., 2025; Wang et al., 2024), which is derived from peer-reviewed Eurorad (European Society of Radiology, 2026) teaching files and covers diverse radiology subspecialties, anatomical categories, and imaging settings. We use its test split (720 datapoints) as the base source for constructing MediSketch examples.

From this source set, we construct patient-facing questions for cases in which visual grounding is likely to provide substantive added value beyond text-only paraphrase. Representative question types include identifying where a lesion is located, distinguishing calcification from uptake, comparing findings across modalities, and understanding interval change across views or studies. Clinician input is used during question construction to ensure that the resulting patient questions are medically meaningful and that the intended educational image serves a real explanatory purpose rather than a cosmetic one. Following the aforementioned processing, we ultimately retained 100 data points to serve as the experimental data for this paper.

We release the 100-case MediSketch testbed materials and evaluation code through the project repository. Additional examples are provided in Appendix E.

B.2 Agent and Image-Tool Protocol

All backbones share the same tool-augmented data generation pipeline. Given the case images, report findings, and patient question, the model first generates drawing instructions for the educational image. These instructions specify the intended educational transformation and are passed to a shared image tool, which returns the edited or annotated educational image. The model then generates the final patient-facing explanation conditioned on both the original case context and the returned image, yielding one complete image–text response bundle for evaluation.

The shared image tool is intended for lightweight educational edits rather than unrestricted image synthesis. Typical operations include arrows, circles, highlights, labels, region emphasis, and simple cross-view or cross-modality visual cues. The purpose is to mark or contrast medically relevant content while preserving the underlying clinical image. In all experiments, we use the state-of-the-art OpenAI GPT Image 1.5 as the image generation/editing backend. Keeping this backend fixed across backbones ensures that the visual-generation interface is shared, so model comparisons focus on

how well each backbone proposes, critiques, and revises complete image–text bundles rather than on differences in the image tool itself.

C Rubric, Judge, and Prompt Details

C.1 Rubric Dimensions

We evaluate each image–text output using nine rubric dimensions, grouped into three broader categories that capture medical safety, patient-facing utility, and visual quality. The first category, *medical grounding, safety, and source integrity*, assesses whether the output remains faithful to the medical report and avoids unsupported clinical inference. It includes **IFC** (Image Factual Correctness), which checks that the illustration does not introduce unsupported findings, anatomy, or spatial relationships; **TMC** (Text Medical Correctness), which verifies that all medical statements in the response are supported by the report or accepted medical knowledge; **SAFE** (Safety-First Visualization & Content), which penalizes speculative diagnoses, unsupported pathology, or over-specific claims under uncertainty; and **BASE** (Base Image Usage Integrity), which ensures that any reused medical image is preserved except for non-destructive overlays such as arrows, labels, or highlights.

The second category, *patient-facing usefulness and image–response alignment*, measures whether the multimodal output helps the patient understand the issue they asked about. It includes **USE** (Patient-Facing Usefulness & Coverage), which evaluates whether the response and image address the patient’s main concern and clarify a relevant concept; **ALIGN** (Clarity & Correct Image-Response Alignment), which checks that the written explanation accurately refers to the visual content without contradictions; and **COMM** (Communication Quality), which evaluates whether the response is clear, coherent, professional, and understandable to a lay audience.

The third category, *visual execution, simplicity, and technical quality*, focuses on whether the generated illustration is technically reliable and appropriately concise. It includes **IQA** (Image QA), which checks label readability, internal consistency, and the absence of visible artifacts; and **SIMPLE** (Simplicity & Non-Escalation), which ensures that the visual aid remains focused, avoids excessive annotations or panels, and does not embed long explanatory text in the image.

The full rubric is provided in Appendix C.10.

Our primary summary metric is the total score across all nine rubric dimensions. This aggregate score enables comparison of overall multimodal educational quality, while the individual dimensions preserve interpretability by separating factual grounding, patient-facing usefulness, and visual execution.

C.2 Judge Inputs and Output Format

Final scoring is primarily performed by GPT-5.4 in a rubric-following mode (with Qwen3.5-397B-A17B included as an additional non-GPT judge for validation; see Appendix C.3). For each candidate bundle, the judge receives the original radiology images, report findings or summary, the patient question, the generated educational image, and the final patient-facing explanation. The judge outputs structured dimension-level scores and an aggregate total score according to the rubric in Appendix C.1.

The evaluation is single-output scoring rather than pairwise preference ranking. We report the sum of the dimension-level rubric scores as the final metric in the main experiments. The full judge prompt and schema are provided in Appendix C.12.

C.3 Radiologist Preference Validation

To validate whether the rubric-following automated judge agrees with clinician preference, we conducted a preference study on 100 randomly sampled head-to-head image–text response pairs. For each pair, we shuffled the A/B order and asked two radiologists to independently select the preferred response. The nine-dimensional rubric was provided as evaluation guidance, but the radiologists were asked to make one final preference choice and provide a short rationale, rather than assign separate numeric scores to each dimension.

We resolved disagreements through discussion and used the adjudicated radiologist preference as a common human reference for evaluating judge behavior. In addition to GPT-5.4, which serves as our primary judge, we include open-weight Qwen3.5-397B-A17B as a non-GPT judge to test whether the rubric-following evaluation signal transfers across judge families.

Appendix Table 2 reports agreement between each automated judge and the adjudicated radiologist reference. GPT-5.4 reaches 90.0% agreement and Cohen’s $\kappa = 0.805$, while Qwen3.5-397B-A17B reaches 82.0% agreement and Cohen’s $\kappa = 0.633$. Although Qwen3.5-397B-A17B is lower than GPT-5.4, its agreement with the radiolo-

gist reference provides a complementary non-GPT evaluation signal.

We therefore use GPT-5.4 as the primary judge because it has stronger agreement with the adjudicated human reference, while reporting Qwen3.5-397B-A17B as an additional non-GPT judge to reduce dependence on a single GPT-family evaluator.

C.4 Lay-Participant Preference Pilot

To complement the clinician validation, we conducted a blinded preference pilot with six lay participants, all of whom demonstrated adequate functional health literacy on the Short Test of Functional Health Literacy in Adults (S-TOFHLA) (Baker et al., 1999) (mean: 33.3/36). Each participant evaluated the same 100 A/B response pairs used in the radiologist validation, reviewing the medical report, patient question, and two image-text bundles before selecting the response they preferred from the patient’s perspective.

Of the 600 judgments, 553 (92.2%) matched the adjudicated radiologist preference, with participant-level agreement ranging from 82% to 99%. Inter-participant agreement was substantial (Fleiss’ $\kappa = 0.721$). Three cases resulted in a 3–3 tie; for all remaining 97 cases, the lay majority matched the radiologist reference.

These results show that radiologist-identified relative preferences are also reflected in non-clinician judgments. However, the pilot measures perceived preference rather than comprehension, recall, or calibrated confidence; it neither compares image-text bundles with text-only responses nor directly validates the nine rubric dimensions and their aggregate score.

C.5 Final Judge versus In-Loop Critique

We use the same rubric logic in two roles that serve different functions. The final judge is GPT-5.4 and is applied only after search to score the selected image-text bundle. By contrast, during Sequential and Hybrid search, the evaluated backbone itself serves as the in-loop critic. The critic evaluates intermediate bundles and guides revision, candidate selection, and early stopping rather than producing the reported score. Thus, although the critic and judge share rubric logic, they use distinct prompt invocations, objectives, and output formats.

The exact prompt forms for both are provided in Appendix C.12 and Appendix C.11.

C.6 Cross-Family External-Critic Ablation

Because our scaling analysis points to weak self-critique and premature stopping as a source of limited scaling, we test whether a cross-family external critic can improve search while holding the generator fixed. We conduct this ablation under the main Hybrid 2×6 setting. For each backbone, we keep the generator model, image backend, prompts, search allocation, and final GPT-5.4 judge unchanged, replacing only the backbone’s in-loop self-critic with together-kimi-k2.6 (hereafter Kimi-K2.6). We retain GPT-5.4 as the final judge because it showed the strongest agreement with the adjudicated radiologist preference in Appendix C.3.

The external critic improves three of the five generators, with the largest gain on GPT-5-nano (+3.80), and increases the realized number of image edits for four of the five generators. This supports the hypothesis that critic quality and premature stopping are important components of the scaling bottleneck. Qwen3.5-27B remains effectively unchanged, with a difference of only -0.07 points. GPT-5.2 instead performs better with its own critic, which is consistent with either stronger self-evaluation capabilities or better compatibility between its generator and self-critic. Overall, critic quality is an important factor but not the only one: generator ability and critic-generator compatibility also affect final bundle quality.

C.7 Prompt Templates

For reproducibility, we provide the prompt templates used throughout the MediSketch pipeline. These include prompts for (i) generating drawing instructions for the shared image tool, (ii) generating or updating the patient-facing response conditioned on the educational image, (iii) producing in-loop critique and revision guidance during Sequential and Hybrid search, and (iv) performing final rubric-based judge scoring. The drawing-instruction, response-generation/updating, critique, and judge prompts are reported in Appendix C.8, Appendix C.9, Appendix C.11, and Appendix C.12, respectively. We also report the rubric used by both the final judge and the in-loop critic in Appendix C.10.

C.8 Drawing-Instruction Prompts

Figure 5 shows the prompt used to generate drawing instructions for the shared image tool. Given

Validation target	Agreement	Cohen’s κ
GPT-5.4 judge vs. adjudicated human reference	0.9000	0.8050
Qwen3.5-397B-A17B judge vs. adjudicated human reference	0.8200	0.6330

Table 2: Agreement between each automated judge and the adjudicated radiologist reference. The reference is obtained through discussion-based resolution of the two radiologists’ disagreements.

Generator	Self-critic		Kimi-K2.6 critic		
	Avg (/29)	#Img	Avg (/29)	#Img	Δ Avg
Qwen3.5-9B	16.93	3.33	18.80	4.00	+1.87
Qwen3.5-27B	20.07	4.13	20.00	5.20	−0.07
GPT-5-nano	19.20	2.40	23.00	5.60	+3.80
GPT-5-mini	22.33	2.13	24.20	3.20	+1.87
GPT-5.2	28.07	6.27	26.60	5.20	−1.47

Table 3: Cross-family external-critic ablation under the main Hybrid 2×6 setting. Kimi-K2.6 denotes together-kimi-k2.6. Only the in-loop critic is changed; the generator, image backend, prompts, search algorithm, and final judge remain fixed. Δ Avg is Kimi-K2.6 performance minus self-critic performance.

the case images, report findings, and patient question, this prompt asks the model to specify the educational edits needed for the visual aid before the image tool produces the annotated image.

C.9 Response Generation Prompts

Figure 6 shows the prompt used for patient-facing response generation in the initial data generation protocol. The response-generation prompt is applied after the image tool returns the educational image, so the model can produce an explanation conditioned on both the original case context and the generated visual aid. Figure 7 shows the prompt used for updating the patient-facing response after the visual aid has been edited.

C.10 Rubric

Figure 8 lists the rubric used by both the final judge and the in-loop critic. The rubric defines the checklist dimensions for evaluating the complete image–text bundle, including medical fidelity, image–text consistency, patient-facing usefulness, and visual quality.

C.11 Critique and Revision Prompts

Figure 9 shows the prompt used for in-loop critique during Sequential and Hybrid search. The **[RUBRIC]** in this prompt refers to the rubric in Appendix C.10. The critic evaluates intermediate image–text bundles and returns feedback used for candidate selection, image editing, response updating, or early stopping.

C.12 Judge Prompts

Figure 10 shows the prompt used for final judge scoring. The **[RUBRIC]** in this prompt refers to the rubric in Appendix C.10. Unlike the in-loop critic, the judge is applied only for the evaluation of the selected image–text bundle.

D Implementation and Evaluation Details

D.1 Models

We evaluate five multimodal-capable models: GPT-5-nano, GPT-5-mini, GPT-5.2, Qwen3.5-9B, and Qwen3.5-27B.

We report results separately for each model rather than averaging within model families, because MediSketch performance depends not only on overall model capability but also on each model’s search behavior and multimodal critique quality.

D.2 Search Strategies

All strategies operate on complete image–text bundles rather than on text-only answers. In one bundle-generation attempt, the backbone model first produces drawing instructions, the shared image tool returns an educational image, and the backbone then generates a patient-facing explanation conditioned on that image. Thus, each search candidate is a full multimodal response bundle consisting of an image and its associated explanation.

Direct Answering. Direct is the non-scaling baseline. It generates one image–text bundle in

a single pass and does not perform candidate selection, critique, or revision.

Parallel. Parallel samples multiple complete bundles independently under the budget cap and selects the best-scoring bundle using the rubric-based critique signal. This strategy explores different candidate bundles but does not revise any individual candidate over time.

Sequential. Sequential starts from one generated bundle and then repeatedly applies critique and revision to the current bundle. At each step, the critic evaluates the current bundle and either returns revision suggestions or accepts the bundle for early stopping. The process continues until the image-generation budget is exhausted or the critic determines that the bundle is sufficient.

Hybrid. Hybrid combines parallel exploration with sequential revision. We denote a Hybrid configuration as $a \times b$, where a is the number of candidate bundles sampled at each step and b is the maximum number of revision steps. At each step, the method samples a candidate bundles, selects the strongest one using the critique signal, and continues revision from that selected candidate into the next step. The total image-generation budget is therefore capped at $a \cdot b$. Sequential can be viewed as a special case of Hybrid with $a = 1$, where the method keeps revision but removes parallel exploration.

D.3 Budget Instantiation

Our main comparison uses a shared maximum image-generation budget cap of 12 per example. We choose this cap based on preliminary budget-sweep experiments, which suggested that, for most models, additional image-generation attempts beyond this range did not reliably produce higher-quality image-text bundles. The same pattern is also reflected in our broader budget-sweep analysis: many models plateau or fluctuate as realized budget use increases, and the strongest backbone, GPT-5.2, is already close to the upper range of observed rubric scores under the 12-attempt cap. Thus, a cap of 12 serves as a practical maximum that is large enough to test nontrivial Parallel, Sequential, and Hybrid strategies, while avoiding excessive image-generation cost in regions where additional attempts show limited marginal gain.

This budget is equal only as a cap, not necessarily as realized consumption. Direct answering is

a non-scaling baseline and generates one image-text bundle, so the multi-step budget cap is not directly applicable. Parallel uses the full cap in the main comparison by generating 12 independent image-text bundles before selecting the best one. Sequential and Hybrid use the same maximum cap, but their realized consumption can be smaller because the search may stop before the cap is reached (i.e. early stopping mechanism). We therefore report the realized average number of edited images per case, denoted as #Img in Table 1, so readers can distinguish allowed budget from actual image-generation usage.

The early-stopping behavior is implemented through the in-loop critic’s discrete overall judgement. At each critique step, the critic must choose one of three outcomes: pass, edit_image, or edit_response. A pass judgement is returned only when the critic judges that all rubric criteria are satisfied and that the current image-text bundle needs no further image or response edits; the Sequential or Hybrid process then stops even if the maximum cap has not been exhausted. If the critic returns edit_image, the system applies the image-editing suggestions and then updates the patient-facing response to reflect the edited image. If the critic returns edit_response, the system applies only the response-editing suggestions while keeping the image unchanged. If no pass judgement is returned, the process continues until the cap is reached.

The main table reports one selected Hybrid setting, namely Hybrid 2×6 . In this configuration, the method samples 2 candidate bundles per step for up to 6 steps, yielding the same maximum cap of 12 image-generation attempts. By contrast, the scaling figures in the main paper include additional Hybrid allocations beyond this single setting. This is why some plotted points in Figure 3 exceed the best Hybrid value shown in Table 1: the table reports the chosen main comparison setting, while the figure reports a broader budget sweep over multiple Hybrid configurations.

D.4 Metrics and Scoring

Our primary metric is the total rubric score, computed as the sum of the nine dimension-level scores in Appendix C.1. We also report realized image usage (#Img) and token cost as secondary efficiency measures.

To make the reported totals interpretable, Appendix C.1 includes the score range for each dimension.

sion. This is especially important for low-range dimensions such as BASE, whose contribution is intentionally bounded.

D.5 SAFE Error Taxonomy

To make the SAFE dimension more directly interpretable, we performed an error-count audit on a 15-case subset. We use a subset because the audit requires human experts to review the judge’s error-type annotations. Each generated image–text bundle can contain more than one error type, so the values in Table 5 are counts of error instances rather than counts of cases. Lower values indicate safer outputs.

Table 4 defines the five error types used in the audit. The taxonomy separates unsupported pathology, unsupported visual details, unsupported text claims, unsupported clinical decision-making, and overconfident localization. This separation is useful because patient-facing safety failures in MediSketch are often caused by visually specific but unsupported details, not only by explicit treatment recommendations.

The audit shows that the main SAFE failures are unsupported visual details, unsupported text claims, and overconfident localization, rather than overt treatment recommendations. It also helps explain why SAFE varies across models and TTS settings. Hybrid and Sequential generally control unsupported visual details and overconfident localization better than Direct, while Parallel does not reliably reduce these errors. This pattern is consistent with the main finding that simply sampling more independent candidates is not enough for safe patient-facing image–text generation.

Error type	Definition
Unsupported visual pathology	The image depicts a disease, lesion, abnormality, or pathological finding that is not explicitly supported by the report. A simplified schematic of a reported condition is allowed if it represents the same reported concept.
Unsupported visual detail	The image adds unsupported non-pathological visual details, such as a specific location, structure, spatial relation, pathway, label, arrow, highlighted region, or measurement cue that is not supported by the report or source image.
Unsupported text claim	The response states a descriptive medical fact that is not supported by the report or by widely accepted medical knowledge.
Unsupported clinical decision	The response gives or implies unsupported diagnosis, staging, exact severity, prognosis, treatment plan, triage advice, or other medical decision-making content, instead of omitting it or explicitly deferring to clinicians.
Overconfident localization	When the report or image evidence does not allow high-confidence identification, the response or image still uses precise region-level wording, exact anatomical labels, or apparent measurements.

Table 4: SAFE error taxonomy used in the 15-case audit.

TTS	Model	Unsupported visual pathology	Unsupported visual detail	Unsupported text claim	Unsupported clinical decision	Overconfident localization	Total
Direct	qwen3.5-9b	6	13	13	1	10	43
Direct	qwen3.5-27b	7	13	12	1	11	44
Direct	gpt-5-nano	9	15	8	1	9	42
Direct	gpt-5-mini	2	9	8	0	8	27
Direct	gpt-5.2	4	10	6	0	4	24
Parallel	qwen3.5-9b	9	14	14	2	13	52
Parallel	qwen3.5-27b	6	13	12	1	9	41
Parallel	gpt-5-nano	8	14	9	2	8	41
Parallel	gpt-5-mini	4	10	9	0	9	32
Parallel	gpt-5.2	4	9	6	0	4	23
Sequential	qwen3.5-9b	5	9	14	0	9	37
Sequential	qwen3.5-27b	8	10	10	0	8	36
Sequential	gpt-5-nano	6	13	8	0	4	31
Sequential	gpt-5-mini	5	13	12	1	10	41
Sequential	gpt-5.2	1	1	0	0	0	2
Hybrid	qwen3.5-9b	5	12	13	1	11	42
Hybrid	qwen3.5-27b	3	11	11	1	9	35
Hybrid	gpt-5-nano	8	10	5	0	4	27
Hybrid	gpt-5-mini	2	10	7	1	6	26
Hybrid	gpt-5.2	0	1	1	0	0	2

Table 5: SAFE error-count audit on a 15-case subset. Each bundle may contain multiple error types, so totals are counts of error instances. Lower is better.

Prompt for Drawing Instruction Generation

You will be given:

1. A medical report
2. A single turn of conversation between a doctor and a patient in which the doctor explains the report.

Your task is to generate a **detailed drawing instruction** for an AI image generation model (e.g., GPT Image 1.5) to create a **medical illustration** that visually supports the doctor's explanation.

The instruction must be precise, concrete, and directly actionable by an image-generation model.

OUTPUT FORMAT (STRICT)

- Output only step-by-step drawing instructions.
- Do NOT include explanations, disclaimers, justifications, or reasoning.
- Do NOT summarize the report or the conversation.
- Do NOT mention safety policies or uncertainty explicitly.

ABSOLUTE PRIORITY: FACTUAL CORRECTNESS (NON-NEGOTIABLE)

The generated image must be 100% medically factual.

You must NOT introduce any visual element unless it is explicitly supported by:

1. The provided medical report
2. The doctor's explanation
3. Universally accepted basic human anatomy

When uncertainty exists, always choose the safest and most conservative visualization:

- Prefer simplified, schematic, or cartoon-style anatomy.
- Show only well-established anatomical structures and basic spatial relationships.
- Do NOT add any lesion shape, size, color, location, laterality, or severity unless explicitly stated.
- Do NOT infer, extrapolate, or fill in missing information.
- If a detail is uncertain, omit it and use a neutral schematic alternative (e.g., highlight a general region rather than a specific structure).

STRICT GROUNDING RULE (FATAL IF VIOLATED)

You must NOT add any new:

- Findings
- Diagnoses
- Implications
- Measurements
- Locations
- Interpretations

If something is not explicitly stated in the medical report or doctor's explanation, it must NOT appear in the image.

BASIC DRAWING REQUIREMENTS

- Specify every visual element needed to draw the illustration.
- Clearly and simply visualize only the key points the doctor is explaining.
- Follow the patient's requests expressed in their **last message**:
 - If the patient asks for annotations, arrows, labels, highlights, or overlays, include them explicitly in the instruction.
- Avoid unnecessary anatomical detail.

BASE IMAGE USAGE

You may be given base images (original medical images or previously generated illustrations).

Default behavior:

- Use simplified, non-patient-specific cartoon diagrams whenever possible.
- Use the given base images only if the patient explicitly requests to draw on them.

When using a base image:

- Explicitly state which base image(s) to use.
- Specify exactly where to place annotations, arrows, labels, highlights, or overlays.

If using an original medical image:

- Do NOT change the imaging modality, anatomy, or findings.
- Do NOT enhance, sharpen, clean up, distort, or reinterpret image content.
- You MAY add transparent overlays, arrows, labels, or highlights.
- The underlying medical content must remain completely unaltered.

SAFETY-FIRST VISUALIZATION POLICY

- Do NOT depict disease appearance, progression, or severity unless explicitly stated.
- Do NOT invent diagnoses, staging, or anatomical locations.
- Do NOT add implied structures or findings.
- If a detail is uncertain, omit it and default to a neutral anatomical schematic.

ULTRA-MINIMAL SKETCH MODE (HIGH PRIORITY)

Treat the illustration as a fast clinical doodle used only to index the doctor's explanation. It is **not** an educational poster or teaching graphic.

Overall Intent

- Aim for the simplest possible visual that still clearly supports the doctor's explanation.
- Minimize visual complexity.
- Allow extra elements only when strictly necessary for accuracy, safety, or clarity.

Layout

- Prefer 1 panel.
- Use 2 panels only if:
 - the doctor is explicitly comparing two concepts, OR
 - a second panel is required to avoid confusion.
- Use more panels only if unavoidable for correctness.

Markers & Visual Elements

- Prefer no more than 3 total markers (arrows, circles, checkmarks, short highlight strokes).
- If more markers are needed:
 - Add only the minimum number required.
- Every marker must map 1-to-1 to a statement in the doctor's explanation.
- Do NOT add:
 - teaching extras
 - example abnormality grids
 - decorative elements
 - supplementary concepts

Text Rules

- Prefer zero text in the image.
- If indexing is required, use only one system:
 - colored arrows / circles / checkmarks
 - 1. 2. 3.
- If the patient explicitly asks for labels:
 - Use only if absolutely necessary.
- Keep each label as short as possible (ideally 2–3 words, up to 7 words max if unavoidable).

Color Rules

- Prefer no more than 3–4 colors total.
- Anatomy: light gray only.
- Color: markers and highlights only.
- Use additional colors only if required for clarity.
- Do NOT use color decoratively.

Semantic Grounding

- Every visual element must directly support a statement from the doctor's explanation.
- No extra markers.
- No unstated concepts.
- No inferred findings.

Clarity & Minimalism

- Keep the image visually uncluttered.
- Prefer the fewest possible panels, markers, colors, and text elements.
- Add complexity only when omission would reduce correctness or safety.
- Every visual element must support the doctor's explanation and nothing else.

Figure 5: Prompt for Drawing Instruction Generation

Prompt for Initial Doctor Response Generation

You are a doctor who is helping a patient understand a medical report.
You will be provided with the patient's medical report and the patient's most recent response.

TASK:

Your task is to make a chatting response based on the patient's recent response to continue the discussion with the patient so that you can make sure the patient fully understands the medical report.
More specifically:

1. If the patient answered a question in the last response, evaluate the answer:
 - If the answer is correct, acknowledge the patient's understanding and briefly summarize the key points in one sentence.
 - If the answer is incorrect or incomplete, gently correct the misunderstanding and provide the correct information strictly based on established medical knowledge, the medical report, and the patient info.
2. If the patient asks questions, address the patient's questions as follows:

STEP 1: Answer the question clearly and concisely using your medical knowledge.

- Your explanation must be directly supported by clearly locatable evidence in the given report text and/or images, and/or by established medical knowledge. The evidence or knowledge must explicitly support the explanation itself. Support based on inference, assumption, indirect relevance, or "it can be reasonably deduced" is not allowed.
- Use established medical knowledge, clear and simple language, and appropriate analogies to make the explanation easy for the patient to understand.
- Do NOT include any medical decision-making content (such as follow-up plans, treatment options, cause analysis, medications, further testing, or disease staging) unless it is explicitly covered in the provided materials.
- When necessary, use information from relevant images in the report to support your explanation. For example, "As you can see from the image, ...", "The first image shows ..., which means ...".
- If the patient's question involves requests drawing on or annotating the original medical images, state that you cannot modify them and instead provide a clear, step-by-step description of what to look for on the images. Remember that you can review images but you cannot draw on them.

NOTICE (HARD STOP RULE):

If the patient's question is outside the scope of the medical report or your medical expertise, you **MUST** politely refuse to answer and **STOP** immediately.

- Do NOT provide an explanation.
- Do NOT continue to STEP 2.
- Do NOT ask any follow-up or clarification questions.

All responses must be based only on the medical report, the patient information, and established medical knowledge. Do not rely on or include any inference, speculation, or extrapolation from the given information.

STEP 2: Use the teach-back method to confirm the patient's understanding.

- You **MUST** ask the patient to explain the information back in their own words to confirm understanding. DO NOT skip this step.
 - The follow-up question must prompt the patient to restate or summarize what was just explained (e.g., "Can you explain it back to me in your own words?").
 - Use varied and natural expressions for the teach-back question. Avoid repeating the same wording across turns.
 - You must not ask any questions or make offers to point out, draw on, mark, highlight, annotate, or otherwise visually indicate anything on an image (e.g., "Do you want me to point out specific landmarks?", "Would you like me to draw or mark this area?", "Should I highlight or annotate the image for you?", "Can I show you where this structure is located?").
3. If the patient deviates from the topic or talks about unrelated issues, gently steer the conversation back to the topic.
 - You need to politely remind the patient about the purpose of the consultation and guide them back to discussing the medical report by repeating the explanation you provided and the question you asked.

REQUIREMENTS FOR YOUR RESPONSE:

1. Your response should be short (4–5 sentences). *Only discuss one aspect at a time to avoid overwhelming the patient.*
2. Your response should be related to the context of the conversation.
3. Always ensure that your responses and questions are empathetic, informative, easy to understand, and aiming at the patient's concern.
4. Do not assume the patient has medical expertise. By default, avoid complex medical terminology, abbreviation, and jargon unless the conversation history clearly shows the patient has high health literacy.
 - When explaining any medical finding, test result, or concept, you **MUST** actively include at least one simple analogy or everyday example to help the patient understand.
 - If a medical term is necessary for clarity, provide a simple explanation right after the term, for example "tuberculosis (a serious infectious lung disease)".
5. **(CRITICAL)** Your response **MUST** be based only on the given medical report and established medical knowledge.
 - Do NOT include any information that is not explicitly supported by these sources.
 - Do NOT rely on inference, speculation, or assumptions beyond what is directly supported.
 - Do NOT include any medical decision-making content (such as follow-up plans, treatment options, cause analysis, medications, further testing, or disease staging) unless it is explicitly covered in the provided materials.
6. Incorporate information from images in the report to support your explanations. For example, "As the image shows, ...", "The ... showed in the image indicates ...".

Figure 6: Prompt for Initial Doctor Response Generation

Prompt for Editing Doctor Response to Refer the Illustrative Image

Doctor Response Editing Prompt

You will be given:

1. An original doctor's response to a patient, and
2. An image generated to visually support the doctor's explanation.

Your task is to rewrite the doctor's response so that it explicitly references and uses the image to help the patient visually understand the explanation. The revised response must remain faithful to the original meaning and tone.

Core Instructions

- Integrate references to the image naturally into the explanation.
- Use the image to reinforce and clarify the doctor's points; do not simply repeat the original wording.
- Preserve all key medical concepts, claims, and explanations from the original response.
- Do NOT introduce any new findings, diagnoses, interpretations, or advice beyond what appears in the original response and the image.
- Write as if the doctor is speaking directly to the patient while referring to the image.
- Keep the same tone, style, and level of formality as the original response.
- Ensure the explanation aligns exactly with any labels, highlights, or visual elements shown in the image.
- Tailor the explanation to the patient's most recent concern or question:
→ Patient last message: {patient_last_message}
- The image-based explanatory portion should be concise: target **2–5 sentences**.
- No bullet points. No step-by-step instructions. The image is an aid, not a walkthrough.

Content Preservation Rules (CRITICAL)

The original response may contain parts that are not directly related to medical explanation, such as:

“Great, you are mostly correct”,

“I understand your concern”,

“Let's move on to the next part”,

“Another important point to note is ...”,

“Do you have any other questions?”,

“Can you explain that back to me in your own words?”,

“I cannot give advice on treatment plans”, etc.

Rules:

- **(CRITICAL)** You MUST NOT discard any of these parts. Your major task is to rewrite the medical explanation-related parts using the generated image as an aid for better patient understanding.
- You may lightly rephrase them to fit the new context, but their meaning must remain unchanged.
- **(EXCEPTION)** You may delete statements that say the doctor cannot draw or generate an image, since the image has already been generated and your role is to help the patient understand it using the image.

Output Requirements

- Output **only** the edited doctor's response.
- Do NOT include any reasoning, commentary, formatting, or extra text.

Figure 7: Prompt for Editing Doctor Response to Refer the Illustrative Image

Rubric

Overall Evaluation Guidelines for Criteria 1.x–9.x

- This benchmark evaluates whether the response and illustration help the patient understand the doctor's explanation while remaining medically safe.
- Do NOT penalize reasonable explanatory schematics merely because they are simplified or illustrative.
- Do NOT reward images that are overly minimal if they fail to visually represent the concept being explained in the response.
- Prefer failing when:
 - the image introduces unsupported medical claims, OR
 - the image fails to address the patient's main concern.
- Prefer passing when:
 - the illustration clearly supports the doctor's explanation,
 - * Illustrations may be either:
 - schematic diagrams, OR
 - overlays on the original base medical images from the report.
 - * Neither type should be preferred automatically.
 - * The appropriate type depends on the patient's question.
 - * For example: Location questions may be best supported by overlays on the original images, while conceptual questions may be best supported by simplified schematics.
 - no new medical claims are introduced,
 - and the image meaningfully supports the doctor's explanation.

Evaluation Criteria

Medical Grounding, Safety, and Source Integrity (Absolute priority + Non-Negotiable)

1. Image Factual Correctness

- 1.1 The generated illustrative image must NOT introduce any visual element unless it is explicitly supported by:

- a) The provided medical report, OR
- b) Universally accepted basic human anatomy (generic structures only).

Note: Generic schematic elements used purely for explanation (e.g., simplified normal anatomy used for comparison) are allowed as long as they are clearly not presented as patient-specific findings. For example:

- a schematic “normal alignment” panel used to compare with a reported abnormal alignment
- a generic anatomical diagram illustrating the structure mentioned in the report

- 1.2 Any inferred, extrapolated, or invented details that are not explicitly stated in the medical report or universally accepted basic human anatomy are strictly prohibited.

Note: Generalized schematic representations or highlighted regions corresponding to anatomical structures explicitly mentioned in the doctor's explanation are allowed if they are clearly illustrative rather than claiming pixel-precise localization. Highlighting a general anatomical region (e.g., gastric wall, tibial plateau, frontal bone inner surface) is acceptable even when the report does not specify exact pixel boundaries.

- 1.3 The image must not depict any anatomical structure, lesion, finding, or spatial relationship not explicitly supported by the report unless it is universally accepted basic anatomy.

Note: Generic anatomical relationships used purely for explanation are allowed if clearly presented as illustrative anatomy rather than patient imaging interpretation.

2. Response Text Medical Correctness

- 2.1 Every factual statement in the doctor's response must be directly supported by:

- a) The provided medical report (text and images), OR
- b) Universally accepted medical knowledge.

- 2.2 Any inferred, extrapolated, or invented details in the doctor's response beyond report/medical knowledge are strictly prohibited.

- 2.3 The doctor's response must not include diagnosis, staging, exact severity grading, exact anatomical sub-location, or a treatment plan unless supported by the report or universally accepted medical knowledge.

3. Safety-First Visualization & Content Policy

- 3.1 The image must not depict any disease/lesion/pathological finding unless explicitly stated in the report.

Note: simplified schematic depiction of a reported condition is allowed if it represents the same concept described in the report rather than introducing a new pathology.

- 3.2 The image must not infer, invent, or imply diagnoses, staging, anatomical locations, structures, findings, or other visual elements not supported by the report.

Note: Illustrative highlighting of anatomical regions mentioned in the response is allowed when used to help the patient understand the doctor's explanation.

- 3.3 The doctor's response must be supported by the report + universally accepted medical knowledge; no unsupported inference.

- 3.4 Any medical decision-making content (treatment plans, diagnoses, staging, exact severity, etc.) must be supported; otherwise omit or explicitly defer.

- 3.5 Uncertainty handling: If the image evidence does not allow high-confidence identification, the image/response must use region-level wording.

Fail if a specific anatomical name is used when identification is uncertain (e.g., labeling “femoral condyle” when the structure is not clearly identifiable).

RULE for Evaluating 1.x, 2.x, and 3.x (Image + Response Together): The image is a visual aid; the response is the primary source of medical claims. Evaluate them together, but note: text clarification in the response cannot “fix” a truly

incorrect or invented visual finding.

Process:

1. First check the image alone against 1.x and 3.x:
 - If the image depicts a specific unsupported structure/finding/pathology/spatial relationship → FAIL (cannot be rescued by text).
2. If the image is merely simplified or potentially ambiguous (not asserting a specific unsupported finding):
 - PASS only if the response explicitly limits interpretation to what the report states (e.g., “illustrative/schematic” or “report does not assess X / this view cannot assess X”).
3. Independently check the response text against 2.x and 3.x:
 - Any unsupported medical claim in the response → FAIL.

4. Base Image Usage Integrity

- 4.1 Determine whether any provided base images (original medical images) are used in the generated illustrative image:
 - If base images are used, identify which base image(s) were used, and compare the generated image to the base image(s) to verify ALL of the following:
 - a) Modality, anatomy, and findings remain unchanged.
 - b) No enhancement, sharpening, cleanup, distortion, or reinterpretation of the base image.
 - c) Only overlays (arrows, labels, highlights) are allowed, and they do not alter the underlying pixels.
 - Focus on the integrity of the base image usage rather than the inferred medical content for this criterion.
 - The explanation and supporting evidence for this criterion must include:
 - The base image(s) used (if any), and
 - Which of the above rules (a/b/c) are violated, if applicable.
 - If no base images are used, still enforce Criteria 1.x and all applicable safety rules.

Patient-Facing Usefulness and Image-Response Alignment

5. Patient-Facing Usefulness & Coverage (Response + Image Together)

These criteria evaluate whether the response and image help the patient, without forcing speculation.

- 5.1 On-topic: The doctor’s response should address the patient’s concerns and remain on-topic.
EXCEPTION A: Connecting phrases are acceptable, such as “Great, you are mostly correct”, “I understand your concern”, “Let’s move on to the next part”, “Another important point to note is ...”, “Do you have any other questions?”, “Can you explain that back to me in your own words?”, “I cannot give advice on treatment plans”.
Only evaluate the parts of the doctor’s response that are intended to address the patient’s concerns or questions.
- 5.2 Visual support: The illustrative image should visually support the doctor’s explanation and help the patient understand at least one key concept that directly relates to the patient’s question.
If the image omits the main explanatory concept that the patient asked about (while only showing unrelated or secondary elements), this criterion should fail even if the image remains safe and minimal.
- 5.3 Image referencing specificity: The response should reference the image in a way that guides the patient to specific visual elements (e.g., “see the arrow pointing to...”), not vague “look at the image.”
Examples of acceptable referencing:
 - “Look at the arrow: it points to the area where the two bones are not aligned as expected.”
 - “The label shows the front of the joint; the arrow highlights the shift described in the report.”
- 5.4 Coverage rule: The response+image should cover key concerns raised by the patient. More specifically, the response should address the key concerns. The image is a visual aid to support the explanation and it is never designed to cover every single concern on its own. Therefore, this criterion passes if:
 - The text response covers the patient’s key concerns within the constraints of Sections 1.x–4.x, referencing the image as needed.
 - The image visually supports the explanation for at least one key concern without violating Sections 1.x–4.x.
 - The image does not omit the primary concept that the doctor is trying to explain.
EXCEPTION B (to prevent forced speculation): If, to cover a patient concern, the doctor would have to speculate about content that violates 1.x, 2.x, or 4.x rules (e.g. diagnosis, staging, exact severity, exact anatomical sub-location, treatment plan, or unsupported spatial relationships), then the doctor must explicitly defer, omit, or refuse to answer that portion, and the image must omit that content without penalty. This coverage rule is still satisfied.
- 5.5 Relevance of visual elements: Any visual element in the image that does not help address the patient’s concerns—AND is not referenced in the response—is unnecessary. If it distracts or confuses, it may be treated as a failure.
- 5.6 If the patient refers to specific locations or features on the report images (e.g., “this spot”, “which slice”, “where exactly”), the illustrative image should preferably use the original medical image with overlays rather than replacing it with a purely schematic diagram. Using only a schematic illustration in such cases may fail if it prevents the patient from identifying the feature on their own medical images.

6. Clarity & Correct Image-Response Alignment

- 6.1 Clarity: The doctor’s response should be clear, coherent, and easy to understand for a layperson unless the patient indicates high medical literacy.
- 6.2 Accurate description of what’s present: Check whether the response describes spatial relationships that are clearly inconsistent with the image. More specifically,
 - Check if the response references a visual element that is not present in the image.
 - Check if the response describes a visual element that is present but inaccurately (e.g. wrong shape/color/size/location/orientation/laterality).
 - Check if the response describes spatial relationships that are inaccurate (e.g. left/right/front/back, anterior/posterior, superior/inferior, medial/lateral).

Any violation of these rules constitutes a failure of this criterion.

Note: Minor approximations are acceptable when the image itself is schematic. Only fail when the description contradicts what the image actually depicts.

6.3 No misinterpretation: The response must not misinterpret the illustrative image or make inaccurate statements about what it depicts.

6.4 No contradictions: The response must not contradict the medical report or the generated image.

7. Communication Quality (Non-medical)

7.1 Natural language: The response should sound natural and human-like; avoid awkward phrasing, repetitive patterns, or AI-like templating.

7.2 Professional tone: No insulting, dismissive, impatient, or inappropriate language.

Visual Execution, Simplicity, and Technical Quality

8. Image QA (Internal Consistency, Text Quality, Artifacts)

8.1 Internal consistency: No contradictory visual messaging within the image.

For example: Fail if:

- Two labels refer to the same structure but point to different locations.
- A label text conflicts with what the arrow/highlight is actually pointing to.
- The same structure is labeled with different names.
- Left/right, medial/lateral, or anterior/posterior indicators contradict each other.
- The image contains visual elements that imply mutually incompatible spatial relationships.

8.2 Text quality: All text in the illustrative image (labels, annotations, legends) must be:

- a) Free of typos, corrupted characters, or unreadable text.
- b) Grammatically correct and medically appropriate.
- c) Clearly legible (sufficient contrast, resolution, and spacing).
- d) Consistent in font style and size unless variation is intentionally used for hierarchy.

Minor stylistic differences are acceptable if they do not affect clarity or readability.

8.3 Generation quality: No artifacts (distortion, mosaic, weird shapes) from image generation.

9. Simplicity & Non-Escalation

9.1 Minimal-necessary annotations: Fail if the image uses more than needed to support the explanation, including any of the following:

- More than 2 panels.
- More than 5 annotation colors (for arrows/circles/highlights/text).
- More than 10 total annotation elements (arrows, circles, checkmarks, text labels, highlights).
- Decorative or unused elements not referenced in the response (see 5.5).

9.2 Keep explanations in the response: Fail if the image contains long-form text (sentences/paragraphs), disclaimers, or explanation boxes. The image is a schematic aid. The clarification and explanation should be in the response text, not embedded in the image.

Allowed: short labels only. Use long labels only when necessary.

Figure 8: Rubric

Prompt for Criticism and Suggestion Making

BACKGROUND

A doctor is trying to respond to a patient's message expressing their concerns or questions about a medical report. The doctor answered the patient's message with a response that includes an illustrative image to help the patient better understand what the doctor is explaining.

TASK

You are a critical judge evaluating the quality of the doctor's response that includes an illustrative image.

Your task is two-fold:

1. to evaluate the quality of both the generated image and the doctor's response that references the image based on a series of evaluation criteria and principles.
2. to provide specific, actionable suggestions for improving the image and the response if they do not meet the evaluation criteria.
 - The suggestions should be detailed and programmatically actionable, which can be directly used as prompts to guide edits to the image or the response using AI.

INPUT

- The patient's message expressing their concerns or questions about the medical report.
- The doctor's response that uses the generated illustrative image to answer the patient's message.
- The image generated to illustrate the doctor's response to the patient's message in order to help the doctor better address the patient's concerns.
- The medical report text and any relevant images that were used as references for generating the illustrative image.

EVALUATION CRITERIA

You will evaluate each criterion and determine if it passes or fails based on the evidence from the generated image, the doctor's response, and the medical report. The evaluation criteria are as follows:

[RUBRIC]

EVALUATION PROCESS

You should evaluate the generated image and the doctor's response based on the above criteria in a systematic and comprehensive manner. More specifically:

For each evaluation criterion, you should first determine whether it is violated or not. You must provide a concise explanation for your judgement and the corresponding evidence from the generated image, the doctor's response, and the medical report to support your evaluation.

MAKING SUGGESTIONS

After the evaluation process, you should then provide a list of specific, actionable editing suggestions for improving the image or the response if they do not meet the evaluation criteria.

The suggestions should be detailed and programmatically actionable, which can be directly used as prompts to guide edits to the image or the response using AI.

More specifically:

STEP 1: Summarize which essential principles are violated.

STEP 2: Choose ONE overall judgement:

- "pass": all criteria met.
- "edit_image": image violates factual correctness, base image usage policy, or safety-first visualization sections.
- "edit_response": image is usable/safe but response needs edits.

STEP 3: Provide specific, programmatically actionable suggestions based on judgement:

- If judgement="edit_response": ONLY suggest response edits; do not suggest any image changes.
- If judgement="edit_image": ONLY suggest image edits; do not suggest any response changes.
- If judgement="pass": do not provide any suggestions.

REQUIREMENTS:

1. When judgment=edit_response, DO NOT suggest any edit to the image. You should focus on providing suggestions for editing the doctor's response to make it pass all the evaluation criteria while keeping the generated image unchanged. When judgment=edit_image, DO NOT suggest any edit to the doctor's response. You should focus on providing a series of commands for editing the image to make it pass all the evaluation criteria. Your suggestions should be like instructions or commands that can be directly executed by an AI image editing model.
2. The suggestions will be directly input to AI models for guiding the edits to the response and the image, so they should be detailed, definitive, and programmatically actionable.

You should not provide general suggestions like "Make the response more relevant to the patient's message" or "Change the image to make it more accurate" or "Replace the caption with a more neutral label such as ...".

Instead, you should provide specific instructions that can be directly used as prompts for editing, such as "Change the annotation text in the image to say 'This is the liver' instead of 'This is the kidney' to align with the medical report and the doctor's wording" or "Edit the response to remove the incorrect statement about the diagnosis and replace it with a statement that accurately reflects what is stated in the medical report".
3. Every suggestion you provide should not violate any of the evaluation criteria. If a suggestion violates any of the evaluation criteria, then it is not a valid suggestion and should not be included in the suggestions list.

OUTPUT

The output of your evaluation and suggestions should be in a JSON format with the following structure:

```
{
  "criteria_evaluation": [
    {
      "criteria": "1.1",
      "judgement": "pass" or "fail",
      "explanation": "explanation to the judgement.",
      "evidence": "evidence for the judgement."
    },
    ...
    {
      "criteria": "2",
      "judgement": "pass" or "fail",
      "explanation": "explanation to the judgement.",
      "evidence": "evidence for the judgement."
    },
    ...
  ],
  "judgement": "edit_response"/ "edit_image"/ "pass",
  "suggestions": ["The first suggestion", "The second suggestion", ...]
}
```

If the judgement is “edit_response”, then these suggestions should be specific, actionable instructions for editing the doctor’s response. If the suggestion_type is “edit_image”, then these suggestions should be specific, actionable instructions for editing the generated image.

Do not include any formatting, reasoning, or explanation outside of the specified JSON response format. The response must be a valid JSON object that can be parsed without errors.

VERIFICATION

Double check the suggestions do not violate any of the evaluation criteria and the overall judgement is consistent with the criteria evaluations.

Double check the output JSON to make sure:

For the “criteria_evaluation” part of the output JSON:

- All criteria are present as a list of dictionaries in the output JSON.
Criteria list: ['1.1', '1.2', '1.3', '2.1', ...]
- Each criterion has a valid “judgement” value (“pass” or “fail”).
- Each criterion has a valid “explanation” value (a non-empty string that provides a concise explanation of the reasoning behind the judgement for this criterion).
- Each criterion has a valid “evidence” value (a non-empty string that provides specific evidence from the generated image, the doctor’s response, and the medical report that supports the judgement for this criterion).

For the overall “judgement” field in the output JSON:

- The “judgement” field in the output JSON must be one of the following values: “edit_response”, “edit_image”, or “pass”.

For the “suggestions” part of the output JSON:

- The “suggestions” field is a list of strings that provide specific, actionable editing suggestions for improving the doctor’s response and/or the generated image based on the evaluation of the criteria.
- If the “judgement” field is “edit_response”, then the “suggestions” field must contain specific, actionable instructions for editing the doctor’s response, and should not contain any suggestions for editing the image.
- If the “judgement” field is “edit_image”, then the “suggestions” field must contain specific, actionable instructions for editing the generated image, and should not contain any suggestions for editing the doctor’s response.
- If the “judgement” field is “pass”, then the “suggestions” field should be empty (i.e., an empty list) since no edits are needed or possible.

Figure 9: Prompt for Criticism and Suggestion Making

Prompt for Judge

BACKGROUND

A doctor is responding to a patient's message that raises concerns or questions about a medical report. In the response, the doctor includes an illustrative image to help the patient better understand the explanation being provided.

TASK

You are a critical judge evaluating the quality of the doctor's response that includes an illustrative image. Your task is to evaluate the quality of both the generated image and the doctor's response that references the image based on a series of evaluation criteria. You need to evaluate each criterion carefully and determine whether the generated image and the doctor's response meet the criterion or not (pass/fail), and provide a concise explanation and specific evidence to support your judgement for each criterion.

INPUT

- The image generated to illustrate the doctor's response to the patient's message in order to help the doctor better address the patient's concerns.
- The patient's message expressing their concerns or questions about the medical report.
- The doctor's response that uses the generated illustrative image to answer the patient's message.
- The medical report text and any relevant images that were used as references for generating the illustrative image.

EVALUATION CRITERIA

You will evaluate each criterion and determine if it passes or fails based on the evidence from the generated image, the doctor's response, and the medical report. The evaluation criteria are as follows:

[RUBRIC]

EVALUATION PROCESS

You should evaluate the generated image and the doctor's response based on the above criteria in a systematic and comprehensive manner. More specifically:

For each evaluation criterion, you should first determine whether it is violated or not. You must provide a concise explanation for your judgement and the corresponding evidence from the generated image, the doctor's response, and the medical report to support your evaluation.

If the criterion is met, then you should give a "pass" judgement. If the criterion is violated, then you should give a "fail" judgement.

OUTPUT

The output of your evaluation and suggestions should be in a JSON format with the following structure:

```
{
  "criteria_evaluation": [
    {
      "criteria": "1.1",
      "judgement": "pass" or "fail",
      "explanation": "explanation to the judgement.",
      "evidence": "evidence for the judgement."
    },
    ...
    {
      "criteria": "2.",
      "judgement": "pass" or "fail",
      "explanation": "explanation to the judgement.",
      "evidence": "evidence for the judgement."
    },
    ...
  ]
}
```

Do not include any formatting, reasoning, or explanation outside of the specified JSON response format. The response must be a valid JSON object that can be parsed without errors.

VERIFICATION

Double check the "criteria_evaluation" part of the output JSON to make sure:

1. All criteria are present in the output.
Criteria list: ['1.1', '1.2', '1.3', '2.1', ...]
2. Each criterion has a valid "judgement" value ("pass" or "fail").
3. Each criterion has a non-empty "explanation" and "evidence" field that provides a concise explanation and specific evidence to support the judgement for that criterion.

Figure 10: Prompt for Judge

E Case Study

In this section, we present a specific medical case, including its medical report (Section E.1) and associated images (Figure 11), and show how test-time compute can critique and improve an image–text bundle through sequential revision.

E.1 Medical Report

CLINICAL HISTORY A 65-year-old male patient presented to the emergency department with diarrhoea of 8 days duration, accompanied by tremors and poor general status. Leukocytosis was detected and he was admitted to further investigation and treatment. During hospitalisation, he presented with subtle desaturation and fecaluria.

IMAGING FINDINGS CT cystography (Image 1) revealed well-distended bladder with thick walls, multiple diverticula, an air-fluid level and leakage of contrast material out of the bladder and into the colon. Associated active diverticulitis was found, with sigmoid wall thickening. The affected section of the sigmoid was contiguous to the bladder dome wall, with obliteration of the fascial plane between them. The patient underwent surgery with identification of the fistulised colon and resection of the diseased segment, and a protective colostomy was performed. The bladder fistula was not identified intraoperatively and the bladder was relieved by a permanent urinary catheter. Seven months later, the patient complained of urine leakage by anus, despite permanent urinary catheter. No additional imaging was done during this period of time. A cystography (Image 2) and a water-soluble contrast enema were then performed. The water-soluble contrast enema showed persistence of the colovesical fistula, with moderate flow (Image 3).

E.2 Original Conversation

Patient’s Question

Yes — I get it. The later studies showed the abnormal tunnel was still there: the CT/enema images demonstrate contrast passing between the sigmoid and the bladder (the cystogram alone didn’t clearly show the tract, but the water-soluble enema filled the bladder, proving the fistula persisted).

Could you please point out the exact fistulous tract on the images and show how large or wide it appears?

Doctor’s Response Without Image Illustration

You’ve understood that correctly: the enema images show contrast traveling from the colon into the bladder, which is like pouring dyed water into one pipe and watching it unexpectedly appear in a neighboring pipe through a hidden “leak” (the fistula). I can’t draw on or directly point to the tract on the images, but in **Image 1 (CT cystography)** the report describes the **red arrow** as the fistulous tract, with **blue-arrow contrast leaking** from the bladder into the colon. In the later testing, **Images 4–5 (water-soluble enema)** don’t outline a clear “tunnel” shape; instead, the key sign is that the **bladder area becomes filled with contrast (the “*” in Image 5)**, which demonstrates that a connection still exists. These

pictures (and the report text) **don’t provide a measurable width/diameter of the tract**, so we can’t reliably say how “wide” it is from what’s shown here.

To make sure we’re on the same page, can you describe—in your own words—what specific visual clue in **Image 1** and what specific clue in **Image 5** tells you there is a fistula?

E.3 Case Study: Initial Generation

Figure 12 shows the initial image–text bundle that is later critiqued and improved through sequential revision. The corresponding critic feedback and judge results are shown in Table 6. The critic’s revision suggestions are listed below:

- S1. Use only two panels: the CT cystography sagittal image (report Image 1) and the water-soluble contrast enema image showing bladder filling (report Image 5). Remove all other panels.
- S2. On the CT panel, keep overlays minimal: add one semi-transparent red circle around the report-indicated fistulous tract region and a short external label, ‘Fistulous tract region,’ with a thin leader line. Do not add new arrows beyond those already present.
- S3. On the enema panel, add one thin white outline around the opacified bladder and a short label, ‘Bladder fills with contrast,’ with a thin leader line. Do not add extra asterisks or duplicate markers.
- S4. Remove embedded explanatory sentences from the image, especially ‘Tract width/diameter not measurable from these images.’ Keep only short labels.
- S5. Fix panel titles to match the source images exactly: use ‘CT cystography (sagittal)’ and ‘Water-soluble contrast enema.’ Avoid duplicated or mismatched numbering.
- S6. Preserve the base images: do not smooth, sharpen, crop away relevant anatomy, or otherwise alter radiology pixels. Add only vector overlays.

E.4 Case Study: Image Editing

Figure 13 shows the image–text bundle after image edition based on critic feedback from Section E.3. The text response is then updated conditioned on the edited image. The corresponding critic feedback and judge results are shown in Table 7. The critic’s revision suggestions are listed below:

- S1. Edit the sentence ‘the small red arrow points to the tract itself’ to a lower-certainty, image-consistent phrasing: replace with ‘the small red arrow is pointing to the area where the report describes the fistulous tract (the exact channel can be hard to see as a distinct tube on a single reproduced screenshot).’ Keep all other meaning unchanged.
- S2. Add one clarifying sentence that explicitly distinguishes what can be localized versus what cannot: after describing the left panel, insert ‘On this CT image we can localize the connection only to this general contact point between the bladder dome and adjacent sigmoid colon, not measure its diameter reliably.’
- S3. Add brief guidance to reduce confusion from other arrows present in the CT panel: insert a sentence such as ‘In this composite image there are other colored arrows from the original figure; for the fistula, focus on the small red arrow and the nearby red-circled region.’ Do not introduce any new interpretation of the yellow/white arrows.

- S4. Revise the right-panel description to be maximally precise about what the image shows: replace ‘oval-highlighted region where the bladder fills with contrast’ with ‘the white oval outlines the bladder; seeing contrast inside this outlined bladder during the enema is the key sign of a persistent connection.’

E.5 Case Study: Response Editing

Figure 14 shows the image–text bundle after response edition based on critic feedback from Section E.4. The critic judged the edited image from the previous round to be sufficient, so the visual aid is unchanged and only the response text is revised. The corresponding critic feedback and judge results are shown in Table 8. The critic’s revision suggestions are listed below:

- S1. Revise the sentence about the blue arrow to avoid overclaiming what is visually obvious on the reproduced screenshot. Replace: “the blue arrow indicates contrast outside its expected compartment, supporting leakage between the bladder and adjacent bowel” with: “the blue arrow (from the original published figure caption) marks the area where contrast is reported to be leaking from the bladder into the adjacent sigmoid/colon; on a single reproduced screenshot it may not look like a distinct stream, so we rely on the radiology interpretation/caption for that detail.”
- S2. Add one clarifying sentence explicitly distinguishing the non-fistula arrows so the patient is not distracted when trying to find the tract. Insert after mentioning ‘other colored arrows’: “For orientation: the yellow arrow marks bladder diverticula and the white arrow marks the air-fluid level in the bladder—those are related findings, but they are not the fistulous tract itself.”
- S3. Tighten the ‘exact tract’ request with a clear limitation statement tied to the imaging type. Add: “Even on CT cystography, the tract may be seen only as a tiny soft-tissue/air/contrast-containing connection rather than a measurable tube; defining an ‘exact’ pixel-by-pixel path or width usually requires the full DICOM series and radiologist measurements, not this single screenshot.”
- S4. When asking the patient to ‘teach-back’, make the question more specific to the overlays to improve usability. Replace the final question with: “Looking at the left panel, what does the small red arrow within the red circle suggest? Looking at the right panel, what does the presence of contrast inside the white-oval-outlined bladder imply?”



Image 1: CT cystography – Reconstructed sagittal CT image Bladder with diverticula (yellow arrow), an air-fluid level (white arrow) and a leakage of contrast material out of the bladder and into the colon (blue arrow). Red arrow shows the fistulous tract.

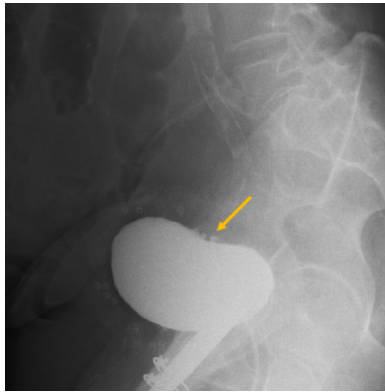


Image 2: Cystography Lateral view - Cystography performed with water-soluble contrast immediately before the enema, showed bladder diverticula (yellow arrow), but could not demonstrate the fistulous tract.

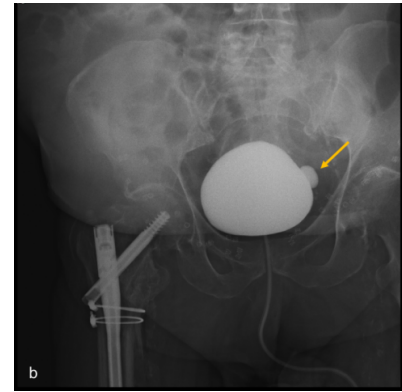


Image 3: Cystography Anteroposterior view - Bladder diverticula (yellow arrow).

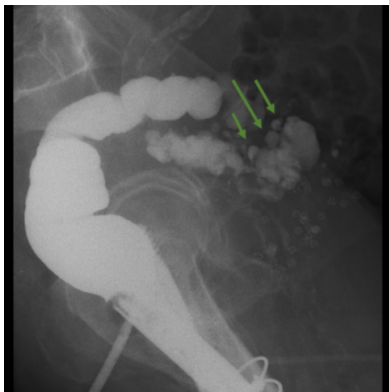


Image 4: Water-soluble contrast enema At first, the enema showed multiple colonic diverticulae (green arrows).

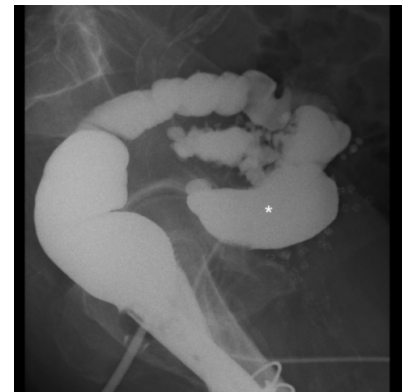


Image 5: Water-soluble contrast enema During the examination, there was filling of the bladder (*), demonstrating the maintenance of the fistulous tract.

Figure 11: Images from the Medical Report

Generated Doctor Response

You've understood that correctly: the enema images show contrast traveling from the colon into the bladder, which is like pouring dyed water into one pipe and watching it unexpectedly appear in a neighboring pipe through a hidden "leak" (the fistula). In **Image 1 (CT cystography, sagittal)**, the "exact" tract is what's circled and labeled "**Fistulous tract**"—the **red circle** encloses the short connection between bowel and bladder, and the **blue arrow** highlights the contrast that's escaping across that connection (the leakage the report describes). In the later testing, **Images 4–5 (water-soluble enema)** don't cleanly draw out a distinct tunnel; instead, the key visual proof is that the **bladder becomes opacified/filled with contrast**, marked by the "*" and the **white outline** around the bladder in **Image 5**, showing contrast has crossed over again. These pictures (and the report text) **don't provide a measurable width/diameter of the tract**, so we can't reliably say how "wide" it is from what's shown here.

To make sure we're on the same page, can you describe—in your own words—what specific visual clue in **Image 1** and what specific clue in **Image 5** tells you there is a fistula?

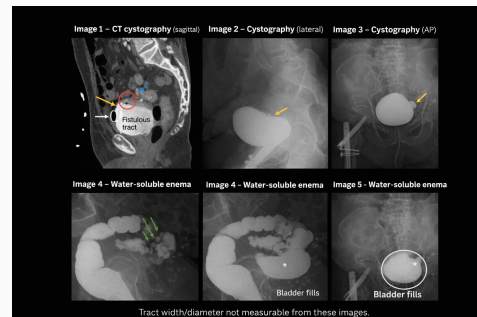


Figure 12: Initial generated image-text bundle for Section E.3

Edited Doctor Response

You've understood that correctly: the enema images show contrast traveling from the colon into the bladder, which is like pouring dyed water into one pipe and watching it unexpectedly appear in a neighboring pipe through a hidden "leak" (the fistula). In the **left panel (CT cystography, sagittal)**, the suspected fistula is in the area explicitly marked "**Fistulous tract region**"—the **small red arrow** points to the tract itself, and the **blue arrow** indicates contrast outside its expected compartment, supporting leakage between the bladder and adjacent bowel. In the later testing, the **right panel (water-soluble contrast enema)** still doesn't outline a crisp, tube-shaped "tunnel"; instead, the key visual is the **oval-highlighted region where the bladder fills with contrast**, which demonstrates that a connection still exists. These pictures (and the report text) **don't provide a measurable width/diameter of the tract**, so we can't reliably say how "wide" it is from what's shown here.

To make sure we're on the same page, can you describe—in your own words—what specific visual clue in the **left CT image** and what specific clue in the **right enema image** tells you there is a fistula?

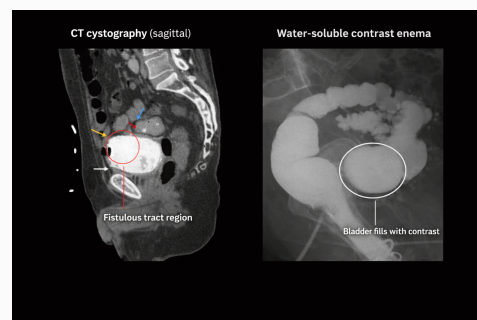


Figure 13: The image-text bundle for Section E.4 after image editing based on critic feedback from Section E.3, followed by response updating conditioned on the edited image.

Edited Doctor Response

You've understood that correctly: the enema images show contrast traveling from the colon into the bladder, which is like pouring dyed water into one pipe and watching it unexpectedly appear in a neighboring pipe through a hidden "leak" (the fistula). In the **left panel (CT cystography, sagittal)**, the suspected fistula is in the area explicitly marked "**Fistulous tract region**"—the **small red arrow is pointing to the area where the report describes the fistulous tract (the exact channel can be hard to see as a distinct tube on a single reproduced screenshot)**, and the **blue arrow** indicates contrast outside its expected compartment, supporting leakage between the bladder and adjacent bowel. In this composite image there are other colored arrows from the original figure; for the fistula, focus on the small red arrow and the nearby red-circled region. On this CT image we can localize the connection only to this general contact point between the bladder dome and adjacent sigmoid colon, not measure its diameter reliably.

In the later testing, the **right panel (water-soluble contrast enema)** still doesn't outline a crisp, tube-shaped "tunnel"; instead, the key visual is that **the white oval outlines the bladder; seeing contrast inside this outlined bladder during the enema is the key sign of a persistent connection**, which demonstrates that a connection still exists. These pictures (and the report text) **don't provide a measurable width/diameter of the tract**, so we can't reliably say how "wide" it is from what's shown here.

To make sure we're on the same page, can you describe—in your own words—what specific visual clue in the **left CT image** and what specific clue in the **right enema image** tells you there is a fistula?

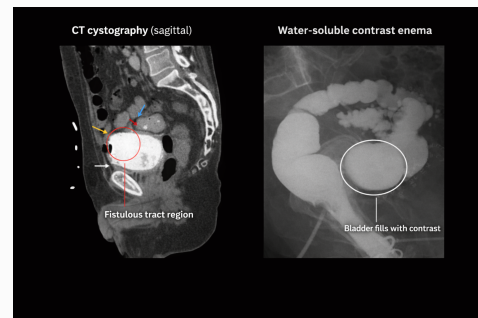


Figure 14: Image–text bundle for Section E.5 after response editing based on the critic feedback from Section E.4. The critic judged the edited image from the previous round to be sufficient, so the visual aid is unchanged and only the response text is revised.

Table 6: Evaluation breakdown for the edited doctor response and illustrative image.

Criteria	Source	Judgement	Points	Explanation	Evidence
1.1	Critic	✓ PASS	—	All highlighted elements correspond to findings explicitly described in the report and/or are neutral overlays on the provided images.	Report for Image 1 describes bladder diverticula, air-fluid level, contrast leakage into colon, and red arrow showing fistulous tract. The illustrative image overlays a red circle/label 'Fistulous tract' on Image 1 and outlines bladder filling on enema images; no new anatomy is introduced beyond bladder/colon and contrast flow shown in the originals.
1.1	Judge	✓ PASS	1	The image uses the provided base medical images and adds explanatory overlays that correspond to findings explicitly described in the report, plus generic labels.	Image 1 overlays a red circle and 'Fistulous tract' where the report states 'Red arrow shows the fistulous tract.' Images 2-5 are the provided cystography/enema images with added arrows/outline/text highlighting reported diverticula and bladder filling.
1.2	Critic	✓ PASS	—	The image does not invent new patient-specific details; it highlights regions already indicated by the report captions and original images.	Overlays (circle, arrows, outline, labels) mark the already-captioned tract/leakage and bladder filling; it does not add a new lesion, new organ, or a new connection not described in the report.
1.2	Judge	✓ PASS	1	No invented patient-specific anatomy or unsupported new pathology is added; the annotations remain at the level of highlighting reported findings.	The composite image highlights the tract on CT, diverticula on cystography/enema, and bladder filling on enema, all described in the report captions. The added note states non-measurability rather than inventing a size.
1.3	Critic	✓ PASS	—	No unsupported anatomical structure, lesion, or spatial relationship is depicted beyond the reported colovesical fistula and diverticula.	Report states colovesical fistula between sigmoid and bladder; Image 1 caption explicitly notes fistulous tract and leakage into colon. Enema images show bladder filling, consistent with persistent fistula; the illustration only emphasizes these.
1.3	Judge	✓ PASS	1	The depicted structures and relationships are supported by the report: bladder, colon, diverticula, contrast leakage/fistulous tract, and bladder filling on enema.	Report: CT showed leakage of contrast out of bladder into colon and a fistulous tract; enema showed persistence of colovesical fistula with bladder filling. The image does not add other lesions or anatomical relationships beyond these.
2.1	Critic	✓ PASS	—	The response's factual statements align with the report and basic medical knowledge.	Response states: CT/enema show contrast crossing between colon and bladder; cystogram alone may not demonstrate tract; enema shows bladder filling; tract width not measurable—these match report text: cystography could not demonstrate tract; enema showed persistence with moderate flow; no measurement given.
2.1	Judge	✓ PASS	1	The doctor's factual statements are supported by the report and basic medical knowledge.	Response says the cystogram alone did not clearly show the tract and that enema images show bladder filling proving persistence, matching the report: cystography 'could not demonstrate the fistulous tract' and enema showed persistence with bladder filling.

Criteria	Source	Judgement	Points	Explanation	Evidence
2.2	Critic	✓ PASS	—	The response avoids inventing measurements, staging, or additional findings beyond the report.	Response explicitly says width/diameter cannot be reliably measured from these images; it does not provide a numeric size or claim severity beyond what is stated.
2.2	Judge	✓ PASS	1	The response avoids unsupported extrapolation and explicitly states that tract width cannot be reliably measured from these images.	Response: 'These pictures (and the report text) don't provide a measurable width/diameter of the tract.' No unsupported measurements or additional diagnoses are given.
2.3	Critic	✓ PASS	—	No unsupported treatment plan or new diagnosis is provided; the response stays explanatory.	Response explains imaging findings and limitations; does not recommend specific interventions.
2.3	Judge	✓ PASS	1	The response does not add unsupported diagnosis, staging, severity grading, exact sub-location beyond what the report already states, or a treatment plan.	It discusses the reported colovesical fistula between sigmoid/bowel and bladder and does not propose treatment or severity beyond the report.
3.1	Critic	✓ PASS	—	The image depicts only the pathology stated in the report (colovesical fistula, diverticula, contrast leakage/filling).	Report explicitly describes fistulous tract and leakage on CT cystography and persistence on contrast enema; illustration highlights these same findings.
3.1	Judge	✓ PASS	1	The image depicts only pathologic findings explicitly stated in the report.	Fistulous tract on CT, bladder diverticula, and bladder filling on enema are all directly described in the report/captions.
3.2	Critic	✓ PASS	—	The image does not imply additional diagnoses or new abnormal findings.	Annotations focus on fistulous tract region (Image 1) and bladder opacification (enema images). No added masses, stones, tumors, or other conditions are shown.
3.2	Judge	✓ PASS	1	The image does not imply unsupported diagnoses or findings; overlays simply emphasize reported features.	No new masses, abscesses, dimensions, or alternate tract paths are drawn. The white outline in Image 5 highlights the bladder filling already shown by the asterisk in the source image.
3.3	Critic	✓ PASS	—	The response remains grounded in the report and explains uncertainty appropriately.	Response notes that enema images do not 'cleanly draw out a distinct tunnel' and uses bladder filling as proof; matches report statement that enema demonstrated maintenance of fistulous tract.
3.3	Judge	✓ PASS	1	The response remains supported by the report and basic medical knowledge throughout.	It explains fistula conceptually as abnormal passage of contrast between colon and bladder and points to bladder opacification on enema, all consistent with the report.
3.4	Critic	✓ PASS	—	No unsupported medical decision-making content is introduced.	Response does not advise surgery/antibiotics/catheter changes; it only interprets the provided report findings.
3.4	Judge	✓ PASS	1	No unsupported medical decision-making content is provided; the response appropriately limits itself to image explanation and uncertainty about measurement.	No treatment recommendation or unsupported severity statement appears. It explicitly defers on measurable width.
3.5	Critic	✓ PASS	—	Where precise identification is limited, the response uses appropriately cautious wording and acknowledges measurement limits.	Response: 'These pictures don't provide a measurable width/diameter...'; and for enema images: 'don't cleanly draw out a distinct tunnel.'

Criteria	Source	Judgement	Points	Explanation	Evidence
3.5	Judge	✓ PASS	1	Where precise measurement is not possible, both image and response use cautious wording and avoid overclaiming width.	Bottom text: 'Tract width/diameter not measurable from these images.' Response: 'we can't reliably say how "wide" it is from what's shown here.'
4.1	Critic	× FAIL	—	Base images appear to be used, but the generated composite includes added/alterd text and duplicated paneling that changes presentation; it is unclear if any underlying pixels were modified vs. only overlaying annotations, and the figure includes an apparent panel mismatch/duplication ('Image 4' shown twice), raising integrity concerns.	The illustrative image is a 2x3 collage of the report's images with added titles/labels and extra text ('Bladder fills', 'Tract width/diameter not measurable...') and includes two panels both titled 'Image 4 – Water-soluble enema'. This suggests re-rendering/recomposition rather than a simple non-destructive overlay on a single base image; the criterion requires verifying no distortion/alteration, which is not satisfied here.
4.1	Judge	× FAIL	0	Base images are used, but the composite includes modifications beyond simple overlays: the final bottom-right panel appears to reuse a different base image than the provided Image 5, and the first panel removes/replaces original colored arrows with new labels, amounting to reinterpretation rather than pure non-altering overlay-only use.	Base images used: provided Images 1-5. Violations: (a) the bottom-right panel labeled 'Image 5' appears to be the AP cystography image (similar to provided Image 3 with pelvis/hardware) rather than the provided enema Image 5, changing the underlying image/findings context; (b/c) in Image 1 the original report arrows are not preserved as simple untouched base-image overlays but are effectively replaced/reinterpreted by a red circle and text label.
5.1	Critic	✓ PASS	—	The response directly addresses the patient's two requests: locating the tract and whether its width can be assessed.	Patient asks to point out the exact tract and show how large/wide it appears; response points to circled tract in Image 1 and states width not measurable.
5.1	Judge	✓ PASS	1	The response directly addresses the patient's request to point out the tract and comment on apparent size/width.	Patient asked to point out the exact tract and show how large/wide it appears. Response identifies the circled tract in Image 1 and explains that width is not measurable from these studies.
5.2	Critic	✓ PASS	—	The image provides clear visual support for tract location on CT and for fistula persistence via bladder opacification on enema.	Image 1 has a circled region labeled 'Fistulous tract'; Image 5 outlines the bladder and indicates 'Bladder fills' with asterisk.
5.2	Judge	✓ PASS	1	The image supports the explanation by highlighting the tract on CT and the indirect evidence of persistence on the enema studies.	Red circle and 'Fistulous tract' label in Image 1 correspond to the main question; bladder filling emphasis in later images supports why the fistula persisted even when the tunnel is not distinctly outlined.
5.3	Critic	✓ PASS	—	The response references specific visual elements (red circle, blue arrow, asterisk, white outline) that are present in the image.	Response: 'red circle encloses...', 'blue arrow highlights...', 'marked by the "*" and the white outline...'. These correspond to visible annotations.
5.3	Judge	✓ PASS	1	The response references specific visual elements rather than vaguely mentioning the image.	It directs the patient to the 'red circle,' 'blue arrow,' the '*' and the 'white outline around the bladder in Image 5.'
5.4	Critic	✓ PASS	—	Key concerns are covered without speculation: exact tract is indicated where visible (CT), and measurement limits are clearly stated.	Response identifies tract on CT and explains that later studies prove persistence by bladder filling; explicitly states no measurable diameter from these images.

Criteria	Source	Judgement	Points	Explanation	Evidence
5.4	Judge	✓ PASS	1	The text covers the key concerns and the image visually supports at least one main concept without forcing unsupported speculation about tract width.	Response explains exact tract localization on CT, explains why later enema proves persistence, and explicitly states width cannot be reliably measured, which is appropriate under the report constraints.
5.5	Critic	× FAIL	—	There are unnecessary or confusing visual elements not clearly needed for the patient's question and/or not referenced precisely, including duplicate 'Image 4' panel and bottom disclaimer text embedded in the image.	Illustration shows two separate panels both labeled 'Image 4 – Water-soluble enema' and includes long bottom text 'Tract width/diameter not measurable from these images.' The response does not address the duplication, and long-form text in-image is not necessary for pointing out the tract.
5.5	Judge	× FAIL	0	The image contains unnecessary or confusing visual elements, including duplicated/incorrect panel labeling and an extra panel that is not clearly needed to answer the patient's question.	There are six panels, with two labeled 'Image 4 – Water-soluble enema' and a final panel labeled 'Image 5 – Water-soluble enema' that appears to be a different image. This duplication and mismatch are not necessary and may confuse the patient.
5.6	Critic	✓ PASS	—	Because the patient asked to point out the tract on the images, using the original images with overlays is appropriate and was done.	Illustration uses the report's actual Image 1 and Image 5 (and others) with added arrows/circle/outline.
5.6	Judge	✓ PASS	1	Because the patient asked to point out the tract on the report images, using the original medical images with overlays is appropriate and helpful.	The illustrative image is a composite of the provided original CT/cystography/enema images with annotations rather than a purely schematic replacement.
6.1	Critic	✓ PASS	—	The response is understandable for a layperson and explains both what can be seen and what cannot be measured.	Uses simple analogy ('dyed water'), defines fistula as abnormal tunnel/leak, and explains why width cannot be measured.
6.1	Judge	✓ PASS	1	The response is clear, coherent, and understandable for a layperson, with a simple pipe/leak analogy and direct image guidance.	Phrases like 'like pouring dyed water into one pipe... through a hidden leak' and explicit references to panels/arrows make the explanation accessible.
6.2	Critic	✓ PASS	—	The response's descriptions match visible elements in the illustration.	Red circle and 'Fistulous tract' label appear in Image 1; blue arrow present; asterisk and white bladder outline appear in Image 5.
6.2	Judge	× FAIL	0	The response references a specific visual element in 'Image 5' that does not match the displayed panel content and labeling reliably.	Response says Image 5 shows the '*' and 'white outline around the bladder.' In the composite, the final bottom-right panel labeled 'Image 5' appears to be an AP cystography-type image rather than the provided enema Image 5, so the description is inconsistent with what is present/labeled.
6.3	Critic	✓ PASS	—	The response does not over-interpret the image and acknowledges limitations of enema/cystogram in outlining the tunnel.	States enema does not 'cleanly draw out a distinct tunnel' and focuses on bladder opacification as proof, consistent with report.
6.3	Judge	× FAIL	0	The response partially misinterprets the illustrative image by treating the mislabeled final panel as the enema proof image.	The panel labeled 'Image 5 – Water-soluble enema' in the composite does not match the provided enema Image 5; relying on that panel for explanation misinterprets the displayed illustration.

Criteria	Source	Judgement	Points	Explanation	Evidence
6.4	Critic	✓ PASS	—	No contradictions between response, report, and illustration that affect medical meaning.	Report: cystography could not demonstrate fistulous tract; response: cystogram didn't clearly show; report: enema shows bladder filling; response: Image 5 shows bladder fills.
6.4	Judge	× FAIL	0	There is a contradiction between the composite image labeling/content and the medical report/source images.	The report's Image 5 is a water-soluble enema image showing bladder filling, but the composite's bottom-right 'Image 5' appears to use a different underlying image. Also, two panels are labeled 'Image 4,' creating inconsistency with the source set.
7.1	Critic	✓ PASS	—	Language is generally natural and patient-facing.	Clear explanations, minimal jargon, coherent structure.
7.1	Judge	✓ PASS	1	The wording is natural and conversational rather than templated or awkward.	The response uses smooth phrasing such as 'You've understood that correctly' and a natural explanatory analogy.
7.2	Critic	✓ PASS	—	Professional, respectful tone.	No dismissive or inappropriate language.
7.2	Judge	✓ PASS	1	The tone is professional, respectful, and supportive.	The response validates the patient's understanding and invites teach-back politely: 'To make sure we're on the same page...'
8.1	Critic	× FAIL	—	The illustration has an internal labeling inconsistency/duplication that may confuse readers.	Two different bottom-row panels are both titled 'Image 4 – Water-soluble enema' even though one appears to correspond to the report's Image 5 (with the asterisk) and the rightmost to Image 3 (AP cystography) with an asterisk added.
8.1	Judge	× FAIL	0	The image has internal consistency problems due to duplicated/incorrect numbering and mismatched panel content.	Two lower panels are labeled 'Image 4 – Water-soluble enema'; the final panel is labeled 'Image 5 – Water-soluble enema' but appears visually inconsistent with the provided Image 5 source. This creates contradictory visual messaging.
8.2	Critic	✓ PASS	—	Most text is legible and free of obvious typos, but there is a labeling duplication; legibility itself is acceptable.	Panel titles and labels ('Fistulous tract', 'Bladder fills') are readable; no garbled characters observed.
8.2	Judge	✓ PASS	1	The overlaid text is legible and mostly grammatically appropriate, with readable contrast and consistent styling.	Titles, 'Fistulous tract,' 'Bladder fills,' and the bottom note are readable in white text on black or dark background, without obvious typos or corrupted characters.
8.3	Critic	✓ PASS	—	No major generation artifacts (warping/mosaic) are evident; the figure appears to be a collage with overlays.	Edges and annotations look clean; no obvious AI distortion in anatomy.
8.3	Judge	✓ PASS	1	No obvious generative artifacts, distortions, or mosaic errors are visible; the image is a clean composite.	Panels retain normal radiographic appearance, overlays are crisp, and there are no malformed shapes or obvious image-generation glitches.
9.1	Critic	× FAIL	—	The image uses more complexity than needed (6 panels) to answer the patient's question, and includes multiple annotation elements and repeated panels.	The collage includes 6 panels (more than 2). The patient asked mainly to locate the tract and assess width; a focused 1–2 panel overlay would suffice.
9.1	Judge	× FAIL	0	The image exceeds the minimal-necessary complexity threshold and includes more panels than allowed.	The composite uses 6 panels, while the criterion fails if there are more than 2 panels. It also includes multiple arrows, circles, labels, outlines, and titles beyond minimal necessity.

Criteria	Source	Judgement	Points	Explanation	Evidence
9.2	Critic	× FAIL	—	The image contains long-form explanatory text that should be in the response, not embedded in the figure.	Bottom caption: 'Tract width/diameter not measurable from these images.' This is explanatory sentence text.
9.2	Judge	× FAIL	0	The image includes sentence-like explanatory text that should be kept in the response rather than embedded in the image.	Bottom text reads 'Tract width/diameter not measurable from these images.' This is explanatory sentence text rather than a short label.

Table 7: Evaluation breakdown for the edited doctor response and illustrative image.

Criteria	Source	Judgement	Points	Explanation	Evidence
1.1	Critic	✓ PASS	—	All added visual elements (circles, labels, arrows) correspond to findings explicitly described in the report (fistulous tract on CT; bladder filling on enema) and do not add new anatomy beyond basic pelvic structures.	Report Image 1 description: red arrow shows fistulous tract; blue arrow shows leakage into colon. Report Image 5: bladder filling during enema demonstrating fistula. Generated image overlays a 'Fistulous tract region' and 'Bladder fills with contrast' consistent with those descriptions.
1.1	Judge	✓ PASS	1	The image uses the provided CT and enema base images and adds explanatory overlays that correspond to findings explicitly described in the report.	Left panel is Image 1 with arrows already present in the source and an added red circle/label over the reported fistulous tract region. Right panel is Image 5 with an added white oval over the bladder filling that the report states demonstrates persistence of the fistula.
1.2	Critic	✓ PASS	—	The illustration avoids inventing tract dimensions or additional findings; it highlights regions rather than claiming pixel-precise boundaries or measurements.	Generated image uses a circled region and an arrow rather than outlining a measured tract; response explicitly states width/diameter cannot be reliably measured from the images/report.
1.2	Judge	× FAIL	0	The image adds a specific pointer implying exact tract localization on the CT beyond what can be reliably supported for patient-facing precision, and the right panel highlights a broad area rather than the tract itself while the patient asked for the exact tract.	The added label 'Fistulous tract region' and response claim that the small red arrow points to the tract itself present a precise tract depiction, but the report only says the red arrow in Image 1 shows the fistulous tract without giving measurable boundaries or width; the added circle suggests a defined region not explicitly delineated in the report.
1.3	Critic	✓ PASS	—	No unsupported lesions or relationships are depicted beyond the reported colovesical fistula concept and contrast passage.	Only two concepts are emphasized: tract region on CT and bladder opacification on enema, both reported.
1.3	Judge	✓ PASS	1	No new anatomy or new pathology is depicted beyond bladder, bowel, and the reported fistula-related leakage/filling.	The image shows the existing CT and enema images with overlays only; no extra organs, lesions, or spatial relationships beyond the reported bladder-colon communication are introduced.
2.1	Critic	✓ PASS	—	The response's factual statements align with the report: enema shows bladder filling (persistent fistula), cystography may not demonstrate the tract, CT shows leakage/tract region; and it correctly states measurement is not provided.	Report: cystography could not demonstrate fistulous tract; enema showed persistence with moderate flow and bladder filling; CT cystography shows leakage and fistulous tract. Response reflects these points and adds only a general pipe analogy.
2.1	Judge	✓ PASS	1	The response statements are supported by the report and basic medical knowledge about contrast traversing a fistulous connection.	The report states CT/enema demonstrate contrast between bladder and colon, the cystogram alone did not clearly show the tract, and the enema demonstrated persistence by bladder filling. The response repeats these points and explains them with a pipe/leak analogy.
2.2	Critic	✓ PASS	—	The response does not extrapolate a tract size or severity; it avoids quantifying diameter and frames the CT location as a marked/suspected region.	Response: 'suspected fistula is in the area... marked "Fistulous tract region"' and 'don't provide a measurable width/diameter.'

Criteria	Source	Judgement	Points	Explanation	Evidence
2.2	Judge	✓ PASS	1	The response avoids unsupported extrapolation and explicitly states that tract width cannot be reliably measured from the provided images/report.	It says the right panel does not show a crisp tube-shaped tunnel and that the images/report do not provide a measurable width/diameter.
2.3	Critic	✓ PASS	—	No new diagnosis, staging, or treatment plan is introduced beyond the already reported fistula and persistence.	Response focuses on interpretation of contrast passage and limitations; no treatment recommendations.
2.3	Judge	✓ PASS	1	The response does not add unsupported diagnosis, staging, severity grading, or treatment recommendations.	It only explains the imaging evidence for persistent colovesical fistula and declines to estimate tract width.
3.1	Critic	✓ PASS	—	The image depicts a pathology (fistula) that is explicitly stated in the report and already annotated in the source CT description.	Report explicitly: 'Red arrow shows the fistulous tract' and later enema demonstrates maintenance of the fistula.
3.1	Judge	✓ PASS	1	The image depicts only the reported fistula-related findings and bladder filling/leakage already described in the report.	Report explicitly mentions leakage of contrast into colon on CT and bladder filling on enema demonstrating maintenance of the fistulous tract; those are the concepts highlighted.
3.2	Critic	✓ PASS	—	The image does not imply additional diagnoses or unreported disease; it confines itself to highlighting the reported fistula-related regions.	No labels for diverticulitis severity, no masses, no added complications; only fistula region and bladder filling.
3.2	Judge	× FAIL	0	The image/response imply an exact tract location in the CT with a specific arrow and labeled region, although the actual tract is not sharply outlined for exact patient-facing localization and no exact width is shown.	Response says 'the small red arrow points to the tract itself' and image labels 'Fistulous tract region,' which implies precise identification. The report supports a tract in that area, but not exact measurable localization or boundaries.
3.3	Critic	✓ PASS	—	The response stays grounded in report-supported findings and uses appropriate uncertainty (doesn't claim crisp tract on later study).	Response: 'still doesn't outline a crisp, tube-shaped tunnel' and relies on bladder filling as evidence.
3.3	Judge	✓ PASS	1	The doctor's explanation is generally grounded in the report and accepted imaging principles.	The response matches the report's statement that the cystogram did not clearly demonstrate the tract while the enema filled the bladder and thus showed persistence of the connection.
3.4	Critic	✓ PASS	—	No unsupported medical decision-making or management guidance is provided.	Response contains explanation only; no recommendations for surgery/antibiotics/catheter changes.
3.4	Judge	✓ PASS	1	No unsupported medical decision-making content is provided; uncertainty about tract size is appropriately deferred.	The response explicitly states that measurable width/diameter cannot be reliably determined from what is shown.
3.5	Critic	✓ PASS	—	The response uses region-level wording and acknowledges limits where the tract is not clearly delineated.	Uses 'area' and 'suspected fistula' for CT and explicitly notes later imaging doesn't outline a crisp tract.
3.5	Judge	✓ PASS	1	The response uses uncertainty-aware wording for the right panel and does not force a precise tract measurement where the image does not allow it.	It says the right panel 'still doesn't outline a crisp, tube-shaped tunnel' and relies on bladder filling as the clue instead of naming a more exact sub-structure or giving a size.
4.1	Critic	✓ PASS	—	The illustrative image appears to reuse the provided base images with simple overlays (circles/labels) without altering modality/anatomy or introducing pixel-level changes to the underlying scans.	Left panel matches Report Image 1 CT sagittal; right panel matches Report Image 5 enema appearance; only added annotations and text are visible, with no apparent reshaping or re-rendering of the scans.

Criteria	Source	Judgement	Points	Explanation	Evidence
4.1	Judge	× FAIL	0	Base images are used, but the left CT base image is altered beyond pure overlays because the original red arrow appears replaced/repositioned and the image is cropped/reformatted; the criterion allows only overlays without altering underlying pixels.	Base images used: Image 1 (CT cystography sagittal) and Image 5 (water-soluble contrast enema). In the left panel, the source arrow set is not preserved exactly as in the original; the red arrow appears changed in shape/placement relative to Image 1, indicating alteration rather than simple overlay. Both images are also reformatted into a composite figure.
5.1	Critic	✓ PASS	—	The response addresses the patient's request to identify the tract and discuss size, and it appropriately explains limitations on measuring width.	Directly points to the CT panel's tract arrow/region and states width is not measurable from provided images/report.
5.1	Judge	✓ PASS	1	The response addresses the patient's concern about where the fistula is seen and whether its size can be determined.	It points to the left CT area, explains the right enema clue, and answers that width cannot be reliably measured from these images.
5.2	Critic	✓ PASS	—	The image directly supports the explanation by highlighting the fistula region on CT and the bladder filling on enema, which are central to the patient's question.	Left label 'Fistulous tract region' and right label 'Bladder fills with contrast' map to the key explanatory concepts.
5.2	Judge	✓ PASS	1	The image helps illustrate the two key concepts relevant to the question: the suspected tract region on CT and bladder filling on enema proving persistence.	Left panel marks the fistulous tract region; right panel highlights the bladder filling with contrast.
5.3	Critic	✓ PASS	—	The response references specific visual elements (left panel, small red arrow, blue arrow, oval highlight) rather than vague instructions.	Response: 'small red arrow points to the tract itself' and 'oval-highlighted region where the bladder fills with contrast.'
5.3	Judge	✓ PASS	1	The response references specific visual elements in the image rather than vaguely telling the patient to look at it.	It specifically mentions the left panel, the 'Fistulous tract region' label, the small red arrow, the blue arrow, and the oval-highlighted region on the right panel.
5.4	Critic	✓ PASS	—	Key concerns are covered: exact tract pointing is addressed as a region/arrow on CT, and the size question is answered with an explicit limitation consistent with the report.	Response explains why tract width can't be measured and distinguishes CT depiction vs enema indirect evidence.
5.4	Judge	✓ PASS	1	The response covers the key concerns within safety limits: it points out the imaging evidence and explicitly declines unsupported size estimation.	Patient asked for exact tract and size; response explains the best-supported visual clues and states width/diameter cannot be reliably measured from the provided images/report.
5.5	Critic	✓ PASS	—	Most visual elements are relevant to the explanation; overlays focus on fistula region and bladder filling, which are referenced in the response.	Response references the red arrow, blue arrow, and oval highlight. No major decorative elements are present.
5.5	Judge	✓ PASS	1	Most visual elements are relevant to the patient's question and referenced in the response.	The panel titles, circled tract region, arrows, and right-side oval are all discussed in the response and support the explanation.
5.6	Critic	✓ PASS	—	Because the patient asked to point out the tract on images, using the original report images with overlays is appropriate and helps the patient locate the feature.	Illustration uses actual CT/enema images rather than a standalone schematic.

Criteria	Source	Judgement	Points	Explanation	Evidence
5.6	Judge	✓ PASS	1	Because the patient asked to point out the tract on the report images, the illustration appropriately uses the original medical images with overlays.	The generated image is a composite of the original CT and enema images rather than a purely schematic replacement.
6.1	Critic	✓ PASS	—	The response is understandable for a layperson and explains the concept with a simple analogy while keeping the key points clear.	Uses 'dyed water in one pipe' analogy and then ties to specific arrows/highlights.
6.1	Judge	✓ PASS	1	The response is clear and understandable for a layperson, with a simple analogy and direct explanation.	It uses the 'dyed water into one pipe' analogy and plainly explains what each panel shows.
6.2	Critic	× FAIL	—	The response claims the 'small red arrow' points to the tract itself, but in the illustrative image the red arrow appears to point toward the circled region near the bladder dome rather than clearly delineating a discrete tract; additionally, the CT panel contains multiple colored arrows (yellow/white) not described in the response, which may confuse what is being referenced as 'small red arrow.'	Generated CT panel shows a red circle and a small red arrow near its upper right edge, but the tract is not crisply visible as a distinct tube; response states it points to 'the tract itself' (more specific than the image supports). Unmentioned yellow/white arrows are present in the CT panel but not explained in the response.
6.2	Judge	✓ PASS	1	The response's references to arrows and highlighted regions generally match what is present in the image.	Left panel contains a small red arrow, a blue arrow, and a 'Fistulous tract region' label; right panel contains an oval-highlighted region labeled 'Bladder fills with contrast'.
6.3	Critic	× FAIL	—	By stating the red arrow points to the tract 'itself,' the response over-interprets what the overlay demonstrates; a safer, accurate phrasing would describe it as the reported/annotated tract region rather than a clearly visible structure in this reproduced panel.	Report Image 1 caption says red arrow shows fistulous tract, but the generated composite/resize plus circle overlay does not ensure the tract is visually identifiable to the patient as a discrete structure; response asserts higher certainty than the displayed clarity.
6.3	Judge	✓ PASS	1	The response does not misread the right panel and appropriately notes that it shows indirect evidence rather than a sharply visualized tunnel.	It states the right panel does not outline a crisp tube-shaped tunnel and instead relies on bladder filling as the key clue.
6.4	Critic	✓ PASS	—	No direct contradiction between response and report regarding persistence and inability to measure width; limitations are acknowledged.	Response agrees cystography alone may not show tract and enema bladder filling indicates connection; matches report.
6.4	Judge	✓ PASS	1	The response is consistent with the report and image content overall.	It agrees that the cystogram alone did not clearly show the tract and that the enema demonstrated persistence by bladder filling.
7.1	Critic	✓ PASS	—	Language is generally natural and patient-facing, with a clear structure and minimal jargon.	Uses simple analogy and references panels plainly.
7.1	Judge	✓ PASS	1	The wording sounds natural and conversational rather than templated or awkward.	Phrases like 'You've understood that correctly' and the pipe/leak analogy read naturally.
7.2	Critic	✓ PASS	—	Tone is professional and respectful; the teach-back request is appropriate.	No dismissive language; asks patient to describe clues to confirm understanding.
7.2	Judge	✓ PASS	1	The tone is professional, respectful, and non-dismissive.	The response validates the patient's understanding and politely asks for teach-back.

Criteria	Source	Judgement	Points	Explanation	Evidence
8.1	Critic	✓ PASS	—	Annotations are internally consistent: the CT panel highlights a fistula region; the enema panel highlights bladder filling; no contradictory laterality markers are shown.	Single region label per panel; arrows and circles do not conflict with labels.
8.1	Judge	✓ PASS	1	The image is internally consistent; labels and highlights do not conflict with one another.	The left-panel label corresponds to the red-circled area and arrow, and the right-panel label corresponds to the white oval over the bladder region.
8.2	Critic	✓ PASS	—	Text in the image is legible, spelled correctly, and consistent in style.	Readable titles 'CT cystography (sagittal)' and 'Water-soluble contrast enema'; labels 'Fistulous tract region' and 'Bladder fills with contrast' have good contrast.
8.2	Judge	✓ PASS	1	The text in the image is legible and medically appropriate without obvious typos or corrupted characters.	Visible labels include 'CT cystography (sagittal),' 'Water-soluble contrast enema,' 'Fistulous tract region,' and 'Bladder fills with contrast,' all readable and correctly spelled.
8.3	Critic	✓ PASS	—	No obvious generation artifacts; the image looks like a clean composite with overlays.	No warping, mosaic artifacts, or corrupted areas beyond normal radiographic noise.
8.3	Judge	✓ PASS	1	No major generation artifacts or distortions are evident in the composite image.	The panels appear clean, with normal radiographic appearance and simple overlay graphics without obvious mosaic artifacts or malformed structures.
9.1	Critic	✓ PASS	—	The image uses two panels and a limited set of annotation elements and colors, staying within the simplicity constraints.	Two panels; a few arrows/circles and short labels; not overly cluttered.
9.1	Judge	✓ PASS	1	The image remains fairly simple with two panels and a limited number of annotations/colors.	There are 2 panels, about 5 annotation colors at most, and fewer than 10 total annotation elements including arrows, circles, and short labels.
9.2	Critic	✓ PASS	—	The image contains short labels only, without long explanatory paragraphs or disclaimers.	Only brief titles and two short callouts are present.
9.2	Judge	✓ PASS	1	The image uses short labels only and leaves the detailed explanation to the response text.	Embedded text is limited to panel titles and brief labels like 'Fistulous tract region' and 'Bladder fills with contrast'; there are no paragraphs or disclaimer boxes.

Table 8: Evaluation breakdown for the edited doctor response and illustrative image.

Criteria	Source	Judgement	Points	Explanation	Evidence
1.1	Critic	✓ PASS	—	The illustration uses the provided base images and adds overlays that highlight findings already described in the report (fistula region on CT; bladder filling on enema) without adding new anatomy or new pathologies.	Medical report: CT cystography shows leakage of contrast into colon and a fistulous tract; enema shows bladder filling. Generated image: left panel is the same sagittal CT with arrows and a red circle labeled 'Fistulous tract region'; right panel is the same enema image with a white oval around the bladder labeled 'Bladder fills with contrast'.
1.1	Judge	✓ PASS	1	The illustrative image uses the provided base images and adds only region-level overlays and labels supported by the report.	Left panel is based on the provided CT cystography image showing the original colored arrows; added red circle/label marks a general fistula region, which matches the report statement that the red arrow shows the fistulous tract. Right panel is based on the provided enema image where bladder filling is reported; added white oval/label highlights the bladder filling.
1.2	Critic	✓ PASS	—	The image does not visually invent a specific tract course or diameter; it marks a general region rather than drawing a new, precise tunnel.	Generated image: a circle and arrow indicate an area rather than a traced tubular tract; no measurement annotations are present.
1.2	Judge	✓ PASS	1	The image does not invent pixel-precise anatomy or an unsupported new tract course; it limits depiction to a suspected region and bladder filling sign.	The left label says 'Fistulous tract region' rather than drawing a definite tube. The right panel highlights 'Bladder fills with contrast,' which is directly described in Image 5/report. No new lesion morphology or dimensions are shown.
1.3	Critic	✓ PASS	—	No unsupported lesions or spatial relationships are introduced beyond those in the report; the overlays are consistent with a colovesical fistula concept described in the report.	Report describes sigmoid contiguous to bladder dome with obliterated plane and fistulous tract; image highlights the contact/leak region without adding additional organs or new abnormal findings.
1.3	Judge	✓ PASS	1	No unsupported anatomical structure, lesion, or spatial relationship is added beyond what is in the report and source images.	The image reuses the CT and enema figures and adds only a circle/oval and short labels. It does not add extra organs, extra fistulas, or claim a different pathway than the report's bladder-colon connection.
2.1	Critic	✓ PASS	—	Key medical statements are supported: enema demonstrates fistula by bladder filling; cystogram alone may not demonstrate the tract; CT can show leakage/region; size cannot be reliably measured from these screenshots.	Report: cystography could not demonstrate fistulous tract; enema showed persistence with moderate flow and bladder filling; CT showed leakage and fistulous tract. Response states these same points and avoids giving a diameter.
2.1	Judge	✓ PASS	1	The doctor's factual statements are supported by the report and basic medical knowledge about contrast traversing a fistula.	Response states that enema images show contrast from colon into bladder, cystogram alone did not clearly show the tract, CT localizes to a general contact point, and width cannot be reliably measured. Report says the cystography could not demonstrate the fistulous tract, while the water-soluble enema showed persistence of the fistula with bladder filling.

Criteria	Source	Judgement	Points	Explanation	Evidence
2.2	Critic	✓ PASS	—	The response does not extrapolate tract size/diameter or give unsupported quantitative claims; it explicitly defers measurement.	Response: 'not measure its diameter reliably' and 'don't provide a measurable width/diameter of the tract'.
2.2	Judge	✓ PASS	1	The response avoids unsupported extrapolation and appropriately limits certainty where the images do not allow precise identification.	The response says 'suspected fistula,' 'exact channel can be hard to see,' 'localize the connection only to this general contact point,' and 'can't reliably say how wide it is,' which avoids invention.
2.3	Critic	✓ PASS	—	No new diagnosis beyond what is in the report (colovesical fistula) and no treatment plan or staging is provided.	Response discusses 'fistula/persistent connection' consistent with report; no medication/surgical recommendations are made.
2.3	Judge	✓ PASS	1	The response does not add unsupported diagnosis details, staging, exact severity, exact sub-location beyond the report, or a treatment plan.	The response discusses only the reported colovesical fistula and inability to measure diameter. No new severity grade, diagnosis beyond the report, or treatment recommendations are introduced.
3.1	Critic	✓ PASS	—	The image depicts only the reported condition (fistula/leakage) and reported study finding (bladder fills on enema) without introducing additional pathology.	Report explicitly mentions fistulous tract and leakage; image annotations correspond to these concepts.
3.1	Judge	✓ PASS	1	The image depicts only pathology already explicitly stated in the report: a fistulous connection region and bladder filling during enema.	Report explicitly mentions a red arrow showing the fistulous tract on CT and bladder filling on enema demonstrating maintenance of the tract. The generated image highlights these same findings.
3.2	Critic	✓ PASS	—	The visualization does not imply additional diagnoses or a more specific anatomic sub-location than the report; it labels a 'region' rather than a precise tract outline.	Generated image uses 'Fistulous tract region' (region-level) and 'Bladder fills with contrast' (directly reported).
3.2	Judge	✓ PASS	1	The image does not imply an unsupported diagnosis or exact tract anatomy; it uses region-level highlighting consistent with uncertainty.	The left panel says 'Fistulous tract region' and circles a general area rather than drawing a definite complete tunnel. The right panel simply outlines the bladder filling with contrast during enema.
3.3	Critic	✓ PASS	—	The response stays within report-supported interpretation and explains limitations of visualizing a discrete tract on a single screenshot.	Response: 'exact channel can be hard to see...' and 'localize...only to this general contact point'.
3.3	Judge	✓ PASS	1	The doctor's response remains medically supported by the report and accepted medical interpretation of contrast studies.	Statements about enema proving persistent connection by bladder opacification, and inability to measure a crisp tube-shaped tract from these reproduced images, match the report and normal radiologic reasoning.
3.4	Critic	✓ PASS	—	No unsupported medical decision-making or treatment planning is included.	Response contains explanation only; no treatment advice.
3.4	Judge	✓ PASS	1	No unsupported medical decision-making or treatment advice is given.	The response does not recommend surgery, medication, or further management; it only explains imaging findings and their limitations.
3.5	Critic	✓ PASS	—	Uncertainty is handled appropriately with region-level wording and explicit limitation about not being able to measure diameter.	Response: 'suspected fistula is in the area...' and 'can localize...only to this general contact point'.

Criteria	Source	Judgement	Points	Explanation	Evidence
3.5	Judge	✓ PASS	1	The response uses uncertainty-aware, region-level wording rather than overclaiming precise identification where confidence is limited.	Phrases include 'suspected fistula,' 'area where the report describes the fistulous tract,' 'exact channel can be hard to see,' and 'general contact point... not measure its diameter reliably.'
4.1	Critic	✓ PASS	—	The generated illustrative image appears to directly reuse the provided base images (CT sagittal and enema) with non-destructive overlays (circle/oval/labels) and without altering underlying pixels in a way that changes modality/anatomy/findings.	Left panel matches Image 1 CT cystography; right panel matches Image 5 water-soluble enema view; overlays are added on top (arrows/circle/oval/text). No evidence of retouching/sharpening or content replacement.
4.1	Judge	✓ PASS	1	Provided base images are used with overlays only; modality, anatomy, and findings remain unchanged and there is no obvious pixel alteration of the underlying medical images.	Base images used: Image 1 (CT cystography sagittal) on the left and Image 5 (water-soluble contrast enema) on the right. Added elements are a red circle/line/text and a white oval/line/text; the original image content, modality, and findings are preserved with no visible distortion or reinterpretive redrawing of anatomy.
5.1	Critic	✓ PASS	—	The response addresses the patient's request to identify the tract and discuss apparent size/width, explaining why exact measurement is not possible from the provided images.	Patient asks: 'point out the exact fistulous tract... show how large/wide'. Response points to the 'Fistulous tract region' and states size cannot be reliably measured.
5.1	Judge	✓ PASS	1	The response stays on the patient's question about where the fistula is seen and whether its size can be shown.	The doctor directly points to the left CT region, explains the right enema sign of persistence, and explicitly answers that the width cannot be measured reliably from these images.
5.2	Critic	✓ PASS	—	The image supports the explanation by highlighting the suspected fistula region on CT and the bladder filling during enema—key concepts the patient asked about.	Image labels: 'Fistulous tract region' and 'Bladder fills with contrast' with visual outlines.
5.2	Judge	✓ PASS	1	The image visually supports the key explanatory concepts: approximate tract region on CT and bladder filling on enema as evidence of persistent fistula.	Left panel labels 'Fistulous tract region' near the original red arrow; right panel outlines the bladder with the label 'Bladder fills with contrast.' These directly support the doctor's explanation.
5.3	Critic	✓ PASS	—	The response references specific visual elements (small red arrow, red-circled region, white oval) and tells the patient what to focus on.	Response: 'small red arrow... nearby red-circled region' and 'white oval outlines the bladder'.
5.3	Judge	✓ PASS	1	The response references specific image elements rather than vaguely referring to the image.	It explicitly directs the patient to the 'left panel,' the 'small red arrow,' the 'nearby red-circled region,' the 'blue arrow,' and the 'white oval' on the right panel.
5.4	Critic	✓ PASS	—	Key concerns are covered: where to look for the tract on CT, what confirms persistence on enema, and why width cannot be measured from these images without speculating.	Response provides both localization guidance and a clear deferral on measurement, consistent with Exception B.

Criteria	Source	Judgement	Points	Explanation	Evidence
5.4	Judge	✓ PASS	1	The text addresses the key concerns and the image supports at least one core concern without unsafe speculation.	Patient asked for exact tract location and apparent size. The response addresses location as a general region on CT and explains why exact width cannot be reliably measured. The image supports both by highlighting the region and the bladder filling sign.
5.5	Critic	× FAIL	—	The composite CT panel contains multiple colored arrows from the original figure that are not all clearly explained as relevant vs irrelevant; this can distract from the patient's goal of identifying the fistulous tract and assessing size.	Generated image left panel includes yellow and white arrows in addition to the blue and small red arrows; response mentions 'other colored arrows' but does not explicitly state what the yellow/white arrows represent (diverticula; air-fluid level) and that they are not the tract itself.
5.5	Judge	✓ PASS	1	Most visual elements are relevant to the patient's concern and/or explicitly referenced in the response.	The added red circle, red label, and white oval are discussed in the response. The original colored arrows from the base CT figure remain visible and are acknowledged by the doctor, who tells the patient which ones to focus on.
5.6	Critic	✓ PASS	—	Because the patient asked to point out the tract on the images, using the original images with overlays is appropriate and is done here.	Generated image reuses the CT and enema base images and adds overlays rather than replacing them with a schematic.
5.6	Judge	✓ PASS	1	Because the patient asked to point out the tract on the actual images, the use of original medical images with overlays is appropriate.	The generated image uses the original CT and enema images rather than a purely schematic drawing, allowing direct visual localization on the report images.
6.1	Critic	✓ PASS	—	The explanation is largely clear for a lay audience and defines what to look for on each panel and why measurement is limited.	Response explains leakage/connection concept, points to panels, and explains limitations.
6.1	Judge	✓ PASS	1	The response is clear and understandable for a layperson, with a simple pipe/leak analogy and straightforward explanation of imaging limitations.	The doctor explains the fistula as 'pouring dyed water into one pipe' and then clearly distinguishes what each panel shows and what cannot be measured.
6.2	Critic	× FAIL	—	The response states the blue arrow indicates contrast outside its expected compartment; however, in the shown CT panel the blue arrow appears to point to a bowel loop/region adjacent to the bladder, and the screenshot alone does not unambiguously demonstrate 'contrast outside' without relying on the original caption/report. The wording risks overclaiming what is visually evident in this reproduction.	Generated image: blue arrow is present but the exact contrast leak is not crisply delineated on this downsampled screenshot; report caption (Image 1) defines blue arrow as leakage into colon, but the response wording presents it as directly evident in the screenshot.
6.2	Judge	✓ PASS	1	The response accurately describes visual elements that are present and does not contradict their location or appearance.	The image contains the small red arrow, blue arrow, red-circled region on the left, and white oval on the right exactly as referenced in the response. No left-right or front-back contradictions are present.
6.3	Critic	✓ PASS	—	The response does not misread the added overlays themselves (circle/oval/arrows) and generally keeps interpretation tied to the report.	Response correctly uses the circle/oval as highlighting aids and notes limitations of seeing a distinct tube.

Criteria	Source	Judgement	Points	Explanation	Evidence
6.3	Judge	✓ PASS	1	The response does not misread the image; it interprets the overlays and original findings consistently with the report.	It correctly treats the left CT as localizing only a suspected region and the right enema as showing bladder opacification rather than a crisp tubular tract.
6.4	Critic	✓ PASS	—	No contradictions with the report are present; the response aligns with the report's statements about which study demonstrated what.	Report: cystography didn't show tract; enema showed persistence; response matches.
6.4	Judge	✓ PASS	1	There is no contradiction between the response, the report, and the generated image.	Report states CT showed the fistulous tract and enema demonstrated persistent fistula by filling of bladder; response states the same and also notes cystography alone did not clearly show the tract, consistent with the report.
7.1	Critic	✓ PASS	—	Language is natural and patient-facing, with a helpful analogy and clear guidance.	Response uses an understandable 'dyed water' analogy and step-by-step references to panels.
7.1	Judge	✓ PASS	1	The response reads naturally and conversationally, without awkward templating.	It uses smooth phrasing such as 'You've understood that correctly' and a clear explanatory flow tied to the two panels.
7.2	Critic	✓ PASS	—	Tone is professional and respectful.	No dismissive or inappropriate phrasing.
7.2	Judge	✓ PASS	1	The tone is professional, respectful, and patient-centered.	The response validates the patient's understanding, explains limitations carefully, and asks a teach-back question politely.
8.1	Critic	✓ PASS	—	Annotations are internally consistent: the 'Fistulous tract region' label points to the circled area; the 'Bladder fills with contrast' label points to the circled bladder on the enema image.	Visual alignment between labels and their leader lines/ovals is consistent.
8.1	Judge	✓ PASS	1	The image's annotations are internally consistent and do not send contradictory visual messages.	The left panel's red circle/line/label all refer to the same region, and the right panel's white oval/line/label all refer to the bladder filling area. No inconsistent laterality or duplicate conflicting labels are present.
8.2	Critic	✓ PASS	—	Text is legible, medically appropriate, and free of typos.	Labels: 'CT cystography (sagittal)', 'Water-soluble contrast enema', 'Fistulous tract region', 'Bladder fills with contrast' are readable and correct.
8.2	Judge	✓ PASS	1	The annotation text is legible, appropriate, and free of obvious errors.	Visible labels read 'CT cystography (sagittal)', 'Water-soluble contrast enema,' 'Fistulous tract region,' and 'Bladder fills with contrast.' Text is high-contrast white on dark background and readable.
8.3	Critic	✓ PASS	—	No obvious generation artifacts (warping, corrupted text, incoherent shapes) are present; overlays are clean.	Edges of circles/ovals and text rendering are crisp; base images look unchanged.
8.3	Judge	✓ PASS	1	There are no obvious image-generation artifacts that impair interpretation.	Both panels appear as clean reproductions of the original medical images with simple overlays; no warped anatomy, mosaic corruption, or malformed annotation shapes are evident.
9.1	Critic	✓ PASS	—	The image remains within minimal complexity (2 panels; limited colors; annotations not excessive), though some arrows are potentially extraneous.	Two panels only; annotations include a circle, an oval, a few arrows, and short labels; appears under the stated limits.

Criteria	Source	Judgement	Points	Explanation	Evidence
9.1	Judge	× FAIL	0	The image exceeds the benchmark's limit on total annotation elements.	There are 2 panel titles, 1 red circle, 1 red leader line, 1 'Fistulous tract region' label, 1 white oval, 1 white leader line, and 1 'Bladder fills with contrast' label, totaling at least 8 added annotations; combined with the 4 original colored arrows retained in the CT base image, the composite visibly contains more than 10 total annotation elements.
9.2	Critic	✓ PASS	—	Image text is short-label style and does not contain long-form explanation or disclaimers.	Only brief headings and labels; no paragraphs in the image.
9.2	Judge	✓ PASS	1	The image contains only short labels, not long explanatory sentences or text boxes.	Embedded text is limited to brief titles and labels such as 'Fistulous tract region' and 'Bladder fills with contrast'; the explanatory detail remains in the doctor's response.