MEDCOUNTERSEG: BENCHMARKING UNSUPPORTED MEDICAL QUERIES AND LEARNING TO DEFER THROUGH CONCEPT-BASED VISUAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recognizing a medical concept does not establish that a particular image supports it. A segmentation model must distinguish these judgments to avoid turning an unsupported query into a plausible mask. We introduce the Medical Query-Counterfactual Segmentation Benchmark (MedCounterSeg-Bench): 6,018 supported–unsupported query pairs across six imaging modalities, 42 concepts, and five clinically reviewed mismatch levels comprising ten subtypes. Each pair shares a target image, while each query receives valid positive examples of its requested concept. The benchmark separately measures concept localization, target-support decisions, and mask quality, with clinician involvement and iterative human quality control during construction. We further propose MedCounterSeg-RL, which learns to defer through concept-based visual reasoning. After positive-only concept–mask alignment, reinforcement learning jointly trains segmentation and deferral; correct deferral retains credit for locating the concept in a positive verification image. With Qwen3.5-4B, the method achieves the highest paired accuracy among evaluated methods on both the in-distribution Test split and unseen datasets, with segmentation quality close to ConceptSeg-R1 on in-distribution queries. The benchmark isolates fine-grained lesion and anatomical mismatches as the main remaining challenge.

1 INTRODUCTION

Multimodal large language models (MLLMs) can interpret clinical questions and produce pixel-level predictions (Huang et al., 2025b), extending medical image segmentation beyond explicit target labels. Yet a query about a scan may express a suspicion rather than an established finding. For example, a user may request segmentation of a suspected liver lesion even when the image shows the liver but no matching focal abnormality. Drawing a lesion contour over liver tissue can make that misinterpretation appear visually confirmed as shown in Figure 1. Mischaracterizing a liver lesion can lead to unnecessary treatment (Zhu et al., 2021). *How can a model determine whether a plausible medical hypothesis is supported by the current image?*

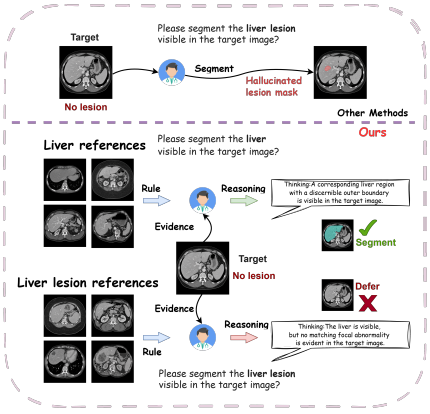

Figure 1: Concept-based reasoning for segmentation and deferral. Instead of hallucinating a mask (top), the model reasons from reference examples and target-image evidence to decide whether to segment or defer (bottom).

Prior work makes rejection part of segmentation: GRES (Liu et al., 2023) includes queries with no target, GSVA (Xia et al., 2024), SESAME (Wu et al., 2024), and MIMIC-ILS (Choi et al., 2026) learn to reject unsupported or false-premise queries, and MedVeriSeg (Tong et al., 2026) verifies queries at inference time. A correct rejection alone, however, does not show that the model recognizes the requested finding, and an aggregate rejection score hides the difference between obvious mismatches and confusion among closely related findings.

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

Clinical reasoning offers a useful analogy: clinicians draw on familiar case patterns (Norman et al., 2007) and combine visual search with diagnostic rules (Yoon et al., 2020). For the liver lesion query above, a clinician might recall liver lesions seen in other patients, identify the visual features relevant to that suspicion, and check whether the current scan provides comparable evidence. Here, the concept is the anatomical structure or pathological finding specified by the query. As illustrated in Figure 1, reference images make such examples explicit for a model: they show what to look for, not whether it is present in the target image. We therefore separate recognizing a concept in a valid example from finding sufficient evidence for it in the target. This motivates two questions:

- *Evaluating concept-based reasoning.* Can an MLLM reason from concept examples and target-image evidence to decide whether to segment or defer?
- *Learning to defer through reasoning.* How can we teach models to make these decisions while still segmenting supported findings accurately?

To address these questions, we introduce **MedCounterSeg-Bench** (Medical Query-Counterfactual Segmentation Benchmark). It pairs supported and unsupported queries on the same target image, drawing on 62 medical datasets and covering 42 anatomical and pathological concepts. Five clinically reviewed mismatch levels comprising ten subtypes range from clear incompatibilities to subtle distinctions between closely related structures and findings. Construction combines clinician review of the mismatch rules and query formulations with iterative human quality control. With valid visual examples of each requested concept, the benchmark evaluates whether models can recognize the concept, accurately segment supported targets, and reject queries that the target image does not support (Table 1).

To learn these capabilities, we propose MedCounterSeg-RL (Medical Query-Counterfactual Segmentation using Reinforcement Learning), a framework for *learning to defer through concept-based visual reasoning*. We first align concept representations with mask prediction on supported examples, then use reinforcement learning to train segmentation and deferral decisions from visual evidence. Training rewards accurate localization and correct deferral, while penalizing unsupported masks and rejection of supported targets. In our design, correctly deferring an unsupported target still earns credit for locating the requested concept in a separate positive verification image. This retains visual feedback about the concept even when segmentation of the target is inappropriate. Direct mask supervision maintains the learning objective of accurate segmentation on supported queries while suppressing unsupported foreground. The response states the inferred visual rule and target-specific evidence, making the proposed basis for the decision available for inspection.

In summary, our contributions are:

- **MedCounterSeg-Bench** provides paired supported and unsupported queries with valid concept examples across six modalities and five clinically reviewed mismatch levels; its evaluation separates concept localization, support decisions, and mask quality (Section 3).
- **MedCounterSeg-RL** learns segmentation and deferral through concept-based visual reasoning. Correct deferral retains concept-localization reward in a positive verification image, while supported queries retain mask supervision even after mistaken rejection (Section 4).
- Paired evaluation reveals tradeoffs that rejection rates alone miss. Our Qwen3.5-4B model achieves the highest paired accuracy among the evaluated methods on both test settings, and isolates lesion-context and fine anatomical mismatches as the remaining hard cases. Reward ablations distinguish preserving supported responses from suppressing unsupported predictions (Section 5).

2 Related Work

2.1 Reasoning Segmentation

Reasoning segmentation uses multimodal language models to interpret queries beyond direct object descriptions. LISA (Lai et al., 2024) connects language-model representations to a mask decoder through a segmentation token, while Seg-Zero (Liu et al., 2025b), LENS (Zhu et al., 2026), and POPEN (Zhu et al., 2025) learn reasoning chains, localization, and masks through reinforcement learning or preference optimization. Medical extensions include M3D (Bai et al., 2024), Med-

PLIB (Huang et al., 2025a), MIMO (Chen et al., 2025), MediSee (Tong et al., 2025), and MedSeg-R (Huang et al., 2025b); MedReasoner (Yan et al., 2026) and PathChat-SegR1 (Liu et al., 2026) further use reinforcement learning for medical segmentation.

Visual examples offer another way to specify what should be segmented. SAM 3 accepts text and visual-exemplar prompts, and SAM3-Agent uses an MLLM to propose prompts and inspect the returned masks (Carion et al., 2026); MedSAM3 (Liu et al., 2025a) adapts SAM 3 to medical concepts; our experiments pair both segmenters with the same agent controller (Appendix B.1). The concurrent ConceptSeg-R1 (Zhao et al., 2026) also reasons from support examples, but its reward assumes a target to localize and has no success case for recognizing a concept and then rejecting an unsupported target, which is the distinction we study.

2.2 Segmentation with Query Rejection

GSVA (Xia et al., 2024) introduces a rejection token for empty-target cases, and SESAME (Wu et al., 2024) learns to detect and explain false premises while segmenting supported targets. RobustSeg (Li et al., 2026b) and CFCamo (Li et al., 2026a) learn segmentation and abstention from counterfactual image–query combinations. In medical imaging, LTSE (Tang et al., 2025) uses a rejection token for absent tissue targets, ROSALIA (Choi et al., 2026) learns from positive and negative chest X-ray instructions, and MedVeriSeg (Tong et al., 2026) verifies queries at inference time without retraining the segmentation model. Inference-time verification can prevent an erroneous mask from being returned, but it does not train the segmenter to change the response that produced it. For learned rejection, an empty output alone does not distinguish a justified rejection from failure to recognize the requested concept. **MedCounterSeg-RL** rewards localization in a valid verification image even when the target is correctly rejected, and supervises the predicted mask before the final output is cleared.

2.3 Query Rejection Benchmarks

Benchmarks test unsupported queries through changes to the query, the image, or both. HalluSeg-Bench (Li et al., 2026b) replaces targets with visually similar objects, CAFE (Liang et al., 2026) edits attributes or context while preserving the target mask, and CF-COD (Li et al., 2026a) removes camouflaged targets through inpainting. In medicine, MIMIC-ILS (Choi et al., 2026) derives positive and negative lesion instructions from chest X-ray reports, MedVeriSeg-Bench (Tong et al., 2026) evaluates clinician-reviewed false queries across modalities, and Polyp-E (Wei et al., 2024) and RAOS (Luo et al., 2024) test target absence through edited polyps and post-resection CT. None of these benchmarks separately evaluates recognition in a valid concept example and support in the target image, and none provides a breakdown from coarse medical mismatches to confusion between closely related findings (Table 1). **MedCounterSeg-Bench** addresses these gaps through

Table 1: Comparison of counterfactual and target-absence segmentation benchmarks. Checkmarks indicate included features, not relative performance.

Benchmark	Query and data construction					Evaluation		
	Paired queries	Concept references	Fine-grained mismatches	Imaging modalities	Clinical review	Mask quality	Paired decisions	Example localization
Natural images								
HalluSegBench (Li et al., 2026b)	✓	✗	✓	–	–	✓	✓	✗
CAFE (Liang et al., 2026)	✓	✗	✓	–	–	✓	✓	✗
CF-COD (Li et al., 2026a)	✗	✗	✗	–	–	✓	✓	✗
Medical images								
MIMIC-ILS (Choi et al., 2026)	✓	✗	✗	1	✓	✓	✗	✗
Polyp-E (Wei et al., 2024)	✗	✗	✗	1	✓	✓	✗	✗
RAOS (Luo et al., 2024)	✗	✗	✗	1	✓	✓	✗	✗
MedVeriSeg-Bench (Tong et al., 2026)	–	✗	–	multi	✓	✗	✗	✗
MedCounterSeg-Bench	✓	✓	✓	multi	✓	✓	✓	✓

✓: included; ✗: not included; –: not applicable or not reported. Paired queries share one target image. Concept references: valid visual examples of the requested concept accompany every query. Fine-grained mismatches test related concepts or attributes. Paired decisions require correct decisions on both members of a pair. Example localization tests recognition in a valid verification image separately from support in the target. Clinical review includes clinician assessment of samples or annotations.

same-image query pairs with valid concept examples, reporting recognition, rejection, and mask quality across mismatch levels.

3 MEDCOUNTERSEG-BENCH

MedCounterSeg-Bench pairs supported and unsupported queries on the same target image, with valid examples of both requested concepts. We refer to the unsupported query in each pair as a *counterfactual query*. The target image remains fixed, while the requested concept and its corresponding reference set change. The pairing tests whether a model can change its decision when the request changes, without treating either frequent rejection or accurate positive masks alone as sufficient evidence of reliable segmentation.

3.1 TASK DEFINITION

For a requested concept c and target image I, the input is $x(c, I) = (\mathcal{Q}(c), \mathcal{R}(c), I)$, where $\mathcal{Q}(c)$ is the query and $\mathcal{R}(c)$ its concept-specific reference set. Each example has a binary support label $y \in \{0, 1\}$ and a target mask $G \in \{0, 1\}^{H \times W}$. We set $y = 1$ when the target image provides sufficient visual evidence to localize the requested concept, and $y = 0$ otherwise. The model must infer the concept's visual features from the references, check them against the target image, and either SEGMENT the requested concept or DEFER with an empty mask. Deferral indicates insufficient evidence in the image, not clinical absence in the patient.

3.2 DATASET CONSTRUCTION

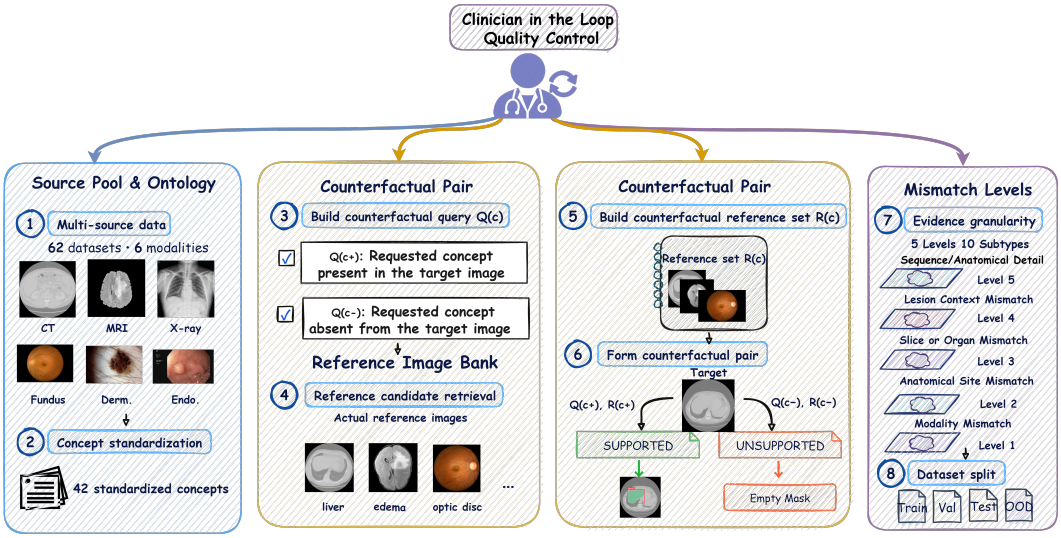

Figure 2: Clinician-in-the-loop construction of counterfactual pairs for MedCounterSeg-Bench. Automatic generation is combined with clinician review of query construction and iterative human quality control.

Figure 2 summarizes the construction process. Using images and segmentation annotations from a subset of IMed-361M (Cheng et al., 2025), we construct 6,018 pairs (12,036 queries) spanning six modalities and 42 anatomical and pathological concepts. We standardize synonyms, abbreviations, and annotation granularity across sources. For each target image, we pair a supported concept with an unsupported one. The supported query has label $y = 1$ and a nonempty annotated mask G; the unsupported query has label $y = 0$ and an empty mask. We construct at most one pair per target image, changing the query and its concept-specific references while keeping I fixed. Each query receives three boxed reference images and one verification image with its box withheld. All four show the requested concept and come from distinct cases, each different from the target case. Thus, even an unsupported query receives valid examples of the concept it asks for.

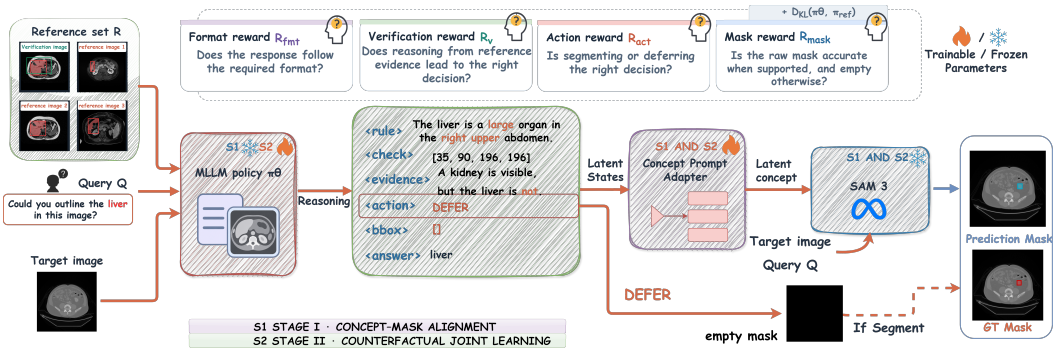

Figure 3: Overview of MedCounterSeg-RL. The MLLM reasons from reference examples and target-image evidence to decide whether to segment or defer. The Concept Prompt Adapter (CPA) translates its response hidden states into segmentation prompts for frozen SAM 3. Stage I trains CPA on supported examples; Stage II jointly trains the MLLM and CPA, retaining concept-localization feedback even after correct deferral.

Clinicians helped refine the query-construction rules and mismatch taxonomy through iterative review, alongside human quality control (Appendix A.5). Where available, masks $\mathcal{C} = \{C_k\}_{k=1}^{K}$ mark visible structures that could be mistaken for the requested concept. These annotations provide supervision, not model inputs.

Training, validation, and test cases are disjoint. AMOS2022 and WORD provide additional out-of-distribution (OOD) targets and are excluded from training and the reference pool. Appendix A gives the split counts and the shared Stage I/II evaluation sets.

3.3 Query–Image Mismatches

We distinguish five levels of query–image mismatch, from broad incompatibilities to finer anatomical and imaging distinctions.

(Level 1) *Imaging modality.* The query asks for imaging features that the supplied modality cannot depict. ***(Level 2)*** *Anatomical site.* The requested structure lies outside the imaged anatomical region. ***(Level 3)*** *Slice or organ.* The requested target is absent from the current slice or is confused with another visible organ. ***(Level 4)*** *Lesion context.* A normal anatomical structure or a different lesion is mistaken for the queried lesion. ***(Level 5)*** *Sequence or anatomical detail.* Finer mismatches involve anatomical substructures, laterality, and sequence-dependent imaging findings.

Levels 4 & 5 in particular require distinguishing a requested lesion or anatomical detail from a related visible structure. Appendix Table 5 defines the subtypes and Appendix Figure 5 shows their distribution across modalities.

4 MedCounterSeg-RL

MedCounterSeg-RL learns to segment or defer through concept-based visual reasoning: recognizing the requested concept in reference examples and checking the target image for supporting evidence. Training first aligns concept representations with masks on supported examples, then uses reinforcement learning to jointly train localization and support decisions (Figure 3).

4.1 Concept Reasoning and Segmentation

Our model combines an MLLM for concept reasoning with frozen SAM 3 (Carion et al., 2026) for mask prediction. Inspired by the context module in LENS (Zhu et al., 2026), a Concept Prompt Adapter (CPA) maps response hidden states to latent segmentation prompts. Given $x = (\mathcal{Q}, \mathcal{R}, I)$ from Section 3.1, the policy generates a response $o \sim \pi_\theta(\cdot \mid x)$. The prompt separates concept recognition from target support: it asks the model to infer a visual rule from the boxed references, localize the concept in the unboxed verification image, and describe target-specific evidence for

$a \in \{\text{SEGMENT}, \text{DEFER}\}$. Successful verification establishes localization in a valid example, not presence in the target image. The response also includes a target box when segmenting and a concise concept phrase c_o. Under either decision, this phrase must name the requested concept rather than a different visible structure.

The latent prompt $E_\phi(o; x)$ produced by CPA and the concept phrase c_o jointly condition SAM 3:

$$P = \sigma\big[\mathcal{S}_\psi\big(I; c_o, E_\phi(o; x)\big)\big]\,. \tag{1}$$

Here, P is the foreground probability map over the $H \times W$ target image, ϕ denotes CPA parameters, ψ the frozen SAM 3 parameters, and σ the sigmoid. Bounding-box overlap provides localization supervision; SAM 3 is conditioned on latent and textual concept prompts rather than boxes.

4.2 Concept–Mask Alignment

In the first training stage, CPA learns to map response hidden states to segmentation prompts on supported examples $\mathcal{D}_1$, with the MLLM and SAM 3 frozen. Given the final-layer states $H_\theta(o; x)$ of the full response, conditioned on the query and input images, CPA uses learned queries Q_ϕ to aggregate them: $E_\phi(o; x) = \mathrm{MLP}_\phi(\mathrm{CrossAttn}_\phi(Q_\phi, H_\theta(o; x)))$. The four cross-attention blocks use all response states as keys and values, without tag-based selection. A two-layer MLP projects each query to a 256-dimensional token; these tokens are concatenated with SAM 3's encoded text prompt.

Teacher forcing with annotated target boxes and a fixed SAM text condition isolates adapter training from response generation. The objective $\mathcal{L}_{\mathrm{align}}(\phi)$ combines soft Dice and pixelwise binary cross-entropy on the predicted mask (Appendix C). Target annotations condition Stage I only; Stage II and inference use generated responses.

4.3 Learning to Segment or Defer

In Stage II, we jointly train the MLLM and CPA on equal numbers of supported and unsupported queries per update, keeping SAM 3 frozen (Figure 3). The support label y, target mask G, verification and supported-target boxes, and available confounder masks $\mathcal{C}$ supervise generated predictions without entering the generation input.

Action reward. The action reward penalizes both segmentation of unsupported queries and rejection of supported ones. Let δ_s indicate SEGMENT with a valid target box, and δ_d indicate DEFER with an empty target box. We define $R_{\mathrm{act}} = 4[y\delta_s + (1-y)\delta_d] - 2$.

Verification reward. Verification provides concept-localization feedback even when the target is unsupported. Let u_v and u_t denote the IoU between each predicted bounding box and its ground-truth box in the verification and target images, respectively, with unavailable or invalid localizations assigned zero. We define

$$R_v = \underbrace{y\delta_s u_v u_t}_{\text{supported localization}} + \underbrace{(1-y)\delta_d u_v}_{\text{verification with deferral}}\,. \tag{2}$$

The first term rewards localization in both images for supported queries. Our second term rewards localization in the valid example when the model correctly defers on the target. Neither term rewards an incorrect target decision.

Mask reward. The mask reward scores predictions before gating, so deferral cannot hide false foreground. For supported queries, $u_m = \mathrm{IoU}(\mathbf{M}, G)$ compares the predicted binary mask $\mathbf{M}$ with the ground-truth mask. For unsupported queries, we replace empty-mask IoU, which cannot distinguish the extent of false foreground, with the mean foreground probability $\rho = (HW)^{-1}\sum_{j=1}^{HW} P_j$ and the overlap $d = \max_{C_k \in \mathcal{C}} \mathrm{SoftIoU}(P, C_k)$ with an annotated query-irrelevant structure ($d = 0$ when none is annotated; SoftIoU is defined in Appendix C). The mask reward is

$$R_{\mathrm{mask}} = \begin{cases} 2u_m - 1, & y = 1, \\ 1 - \sqrt{\rho} - d, & y = 0. \end{cases} \tag{3}$$

The square root sharpens the penalty for low foreground occupancy, which image-wide averaging would dilute, and d penalizes overlap with a specific confounder.

The format reward $R_{\text{fmt}} = 2F(o) - 1$ uses $F(o) = 1$ for the required structure and $F(o) = 0$ otherwise. The combined reward is $R = R_{\text{act}} + R_{\text{fmt}} + R_v + R_{\text{mask}}$. Appendix C gives the response structure.

Joint optimization. We optimize the clipped GRPO objective $\mathcal{L}_{\text{GRPO}}(\theta)$ (Shao et al., 2024) with within-query reward normalization and KL regularization to a frozen reference policy. CPA minimizes $\mathcal{L}_{\text{CPA}}(\phi)$: Dice and focal loss toward G for supported queries, and focal loss toward an empty mask otherwise, regardless of the predicted action. Rewards are fixed scores; mask gradients pass through frozen SAM 3 to CPA but stop at the MLLM hidden states. No annotated reasoning traces are used. Appendix C details the joint objective.

Inference. The model describes the target-image evidence and returns the gated mask $\mathbf{M}_a = \mathbf{1}[a = \text{SEGMENT}]\,\mathbf{M}$, which is empty for DEFER.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

We evaluate on the Test and OOD splits of MedCounterSeg-Bench (Appendix A).

Evaluation metrics We compute mean Dice and mean IoU at the original image resolution over samples. Action accuracy and macro-F1 measure SEGMENT/DEFER decisions; paired accuracy requires both decisions for an image to be correct. Verification-box IoU measures concept localization in a valid example. We report decision performance by mismatch level. For unsupported queries, we distinguish correct deferral from mask emptiness and measure foreground probability before gating.

Implementation details Our main model uses Qwen3.5-4B (Qwen Team, 2026); Appendix B.2 reports the Qwen2.5-VL-3B (Bai et al., 2025) configuration. We train LENS (Zhu et al., 2026), Seg-Zero (Liu et al., 2025b), and ConceptSeg-R1 (Zhao et al., 2026) on supported queries from our training split; these methods have no rejection interface. SAM3-Agent and MedSAM3-Agent use a Qwen2.5-VL controller with SAM 3 (Carion et al., 2026) or MedSAM3 (Liu et al., 2025a), without training. Among methods with explicit rejection or verification, GSVA (Xia et al., 2024) is fine-tuned on the same Stage II queries as ours, SESAME (Wu et al., 2024) uses released weights, and MedVeriSeg (Tong et al., 2026) is training-free.

We use official implementations and evaluate all methods on the same target images and queries. ConceptSeg-R1 and our method also receive the four-image reference set; the other baselines receive only the target image and query (Appendix B.1).

5.2 MAIN RESULTS

Our Qwen3.5-4B model leads in action and paired accuracy (Table 2), answering both queries correctly for 83.17% of image pairs, compared with 65.17% for SESAME (Wu et al., 2024). It rejects fewer supported queries than SESAME at a similar correct deferral rate. Its mDice is close to ConceptSeg-R1 (Zhao et al., 2026) with the same backbone (71.35 versus 71.66). Across both splits, our 4B model has higher point estimates for segmentation quality and paired accuracy than the decision-enabled baselines (Appendix Figure 8). Appendix B.7 gives bootstrap intervals.

Performance across mismatch levels. The aggregate paired accuracy is carried by coarse mismatches, which the benchmark separates from fine-grained ones (Appendix B.5). On Levels 1–2 our model reaches 96.1% and 95.2% paired accuracy and rejects every unsupported query; on Level 3 it reaches 51.5%, versus 26.2% for SESAME and 35.0% for MedVeriSeg, with a similar correct deferral rate but far fewer rejected supported queries. On Levels 4–5, where the unsupported concept is a related lesion, substructure, or sequence-dependent finding, correct deferral falls to 41.2% and 15.8% for our model, and no compared method exceeds 58.8% and 42.1%; paired accuracy stays below 36% for every method. A single rejection rate would report these models as reliable; the level breakdown locates the unsolved cases in lesion identity, anatomical detail, and sequence-dependent findings, and fixes a protocol for measuring progress on them.

Table 2: Segmentation and deferral performance on the MedCounterSeg-Bench Test split. Paired accuracy requires correct decisions on both queries of an image pair; errors are invalid or failed outputs.

Method	Backbone	Segmentation		Support decisions				Output
		mDice ↑	mIoU ↑	Paired accuracy ↑	Action accuracy ↑	Correct deferral ↑	False deferral ↓	Errors (%) ↓
Without Support Decisions								
LENS	Qwen2.5-3B	53.14	43.68	–	–	–	–	7.08
Seg-Zero	Qwen2.5-7B	39.47	31.98	–	–	–	–	3.33
ConceptSeg-R1	Qwen2.5-3B	72.03	64.22	–	–	–	–	0.00
	Qwen3.5-4B	71.66	63.73	–	–	–	–	14.00
With Support Decisions								
GSVA	LLaVA-7B	37.74	31.32	18.50	54.42	22.17	12.17	2.83
SESAME	LLaVA-1.5-7B	52.18	44.48	65.17	81.08	91.50	29.33	0.00
SAM3-Agent	Qwen2.5-3B	57.56	50.53	22.67	50.75	56.50	18.33	37.42
	Qwen2.5-7B	60.80	53.85	10.00	37.50	14.67	1.67	55.08
MedSAM3-Agent	Qwen2.5-3B	52.69	45.87	41.17	65.75	58.33	20.67	6.58
	Qwen2.5-7B	52.69	46.13	30.83	64.17	32.17	2.67	1.50
MedVeriSeg	LLaVA-Med-7B	36.80	29.46	51.83	73.33	75.50	28.83	0.00
Ours	Qwen3.5-4B	71.35	63.34	83.17	91.17	91.67	9.33	0.00

Table 3: Out-of-distribution segmentation and deferral performance on AMOS2022 and WORD.

Method	Backbone	Segmentation		Support decisions				Output
		mDice ↑	mIoU ↑	Paired accuracy ↑	Action accuracy ↑	Correct deferral ↑	False deferral ↓	Errors (%) ↓
Without Support Decisions								
LENS	Qwen2.5-3B	18.72	14.02	–	–	–	–	1.56
Seg-Zero	Qwen2.5-7B	17.57	13.33	–	–	–	–	3.44
ConceptSeg-R1	Qwen2.5-3B	73.26	64.68	–	–	–	–	0.00
	Qwen3.5-4B	75.29	66.64	–	–	–	–	5.00
With Support Decisions								
GSVA	LLaVA-7B	8.62	6.15	12.50	46.56	20.00	23.12	3.75
SESAME	LLaVA-1.5-7B	21.74	15.50	17.50	50.62	76.25	75.00	0.00
SAM3-Agent	Qwen2.5-3B	83.84	76.56	3.75	27.19	8.75	3.75	67.81
	Qwen2.5-7B	81.40	74.36	0.00	26.88	0.62	0.62	69.69
MedSAM3-Agent	Qwen2.5-3B	33.45	29.50	9.38	43.44	13.12	7.50	16.88
	Qwen2.5-7B	31.56	28.01	2.50	50.00	2.50	1.88	0.62
MedVeriSeg	LLaVA-Med-7B	30.61	24.86	30.00	62.81	42.50	16.88	0.00
Ours	Qwen3.5-4B	73.43	65.42	52.50	74.38	71.25	22.50	0.00

5.3 Out-of-Distribution Generalization

On AMOS2022 and WORD, our Qwen3.5-4B model retains the highest action and paired accuracy (Table 3). Paired accuracy is 52.50%, compared with 30.00% for MedVeriSeg, the strongest baseline on this metric. SESAME rejects more unsupported queries but also many more supported ones. Transfer varies across backbones (Appendix B.2).

5.4 Ablation Studies

We remove one reward at a time to assess its contribution (Table 4), keeping the initialization, data, and 100-update training schedule unchanged. The prompt, format reward, and CPA losses are fixed.

w/o action reward. Removing the action reward increases false deferral from 9.33% to 16.67%, while correct deferral changes little. Its main contribution is to prevent the model from rejecting supported queries.

w/o negative mask reward. Removing this reward lowers correct deferral from 91.67% to 85.17%. Mean foreground probability on unsupported queries also rises from 0.79% to 2.01% before gat-

Table 4: Reward ablations and a reference-reasoning comparison for segmentation and support decisions with Qwen3.5-4B on the Test split. Scores are percentages.

Variant	Segmentation		Verification	Support decisions			
	mDice ↑	mIoU ↑	IoU ↑	Action accuracy ↑	Correct deferral ↑	False deferral ↓	Paired accuracy ↑
w/o action reward	70.54	62.44	18.09	86.67	91.83	16.67	75.00
w/o negative mask reward	70.56	62.60	28.66	90.33	85.17	4.50	81.00
w/o verification reward	73.30	65.21	0.00	87.25	93.17	18.50	75.33
w/o reference reasoning	61.08	52.90	0.00	87.25	75.17	0.33	74.83
Ours (full)	71.35	63.34	25.25	91.17	91.67	9.33	83.17

The verification ablation removes only the unsupported-query reward; verification IoU is reported on these queries. *w/o reference reasoning* removes the reference images and the reasoning fields at inference.

ing: the reward suppresses unsupported foreground, rather than merely hiding it behind a deferral decision.

w/o verification reward. We remove only the unsupported-query term in Equation 2, retaining verification supervision on supported queries. Verification IoU drops from 25.25 to 0.00, and false deferral rises from 9.33% to 18.50%. This term preserves concept localization in the designated verification image when the target query is unsupported.

w/o reference reasoning. Removing the reference reasoning modules and at inference lowers correct deferral from 91.67% to 75.17%, paired accuracy from 83.17% to 74.83%, and mDice from 71.35% to 61.08%.

What the ablations show. The variants separate the roles of the rewards. The action reward calibrates the decision: without it, correct deferral is unchanged and false deferral nearly doubles. The negative mask reward makes deferral act on the mask, not only on the action token: raw foreground on unsupported queries falls from 2.01% to 0.79% and correct deferral rises by 6.5 points. The verification term keeps concept feedback alive during rejection: removing it drives verification IoU to zero and doubles false deferral, so a model that stops localizing the concept in the reference also starts rejecting supported targets. Removing the reference images and reasoning fields shows the opposite failure: false deferral drops to 0.33% but correct deferral loses 16.5 points, so reference-based reasoning is what makes unsupported queries recognizable. The only variant with higher mDice (73.30, w/o verification reward) reaches it with twice the false deferral, on a smaller and easier accepted set; none of the variants matches the full model's 83.17% paired accuracy.

6 Conclusion

Medical segmentation should distinguish knowing a concept from finding evidence for it in a particular image. MedCounterSeg-Bench makes this distinction measurable through same-image supported–unsupported query pairs, valid concept examples, and five clinically reviewed mismatch levels, exposing both unsupported segmentation and mistaken refusal. MedCounterSeg-RL learns to defer through concept-based visual reasoning, retaining concept-localization feedback after correct deferral and supervising masks before the decision clears them. With Qwen3.5-4B it achieves the highest paired accuracy among the evaluated methods on both test splits, with segmentation quality close to ConceptSeg-R1 in distribution. Fine-grained lesion and anatomical mismatches remain difficult (Appendix A.6); our objective is tested on one concept-reasoning backbone, and the reference images are not ablated at inference. The benchmark measures progress on these distinctions without equating refusal with better segmentation.

THE USE OF LARGE LANGUAGE MODELS

Large language models were used only to improve the writing of this paper, including grammar, clarity, and LaTeX formatting. They were not used for research ideation, benchmark construction, experimental design, model training, or analysis of results. All content was reviewed and verified by the authors, who take full responsibility for it.

REPRODUCIBILITY STATEMENT

MedCounterSeg-Bench is built from publicly available datasets: a subset of IMed-361M, with AMOS2022 and WORD as OOD sources. Appendix A gives the construction rules, mismatch taxonomy, and split statistics. Appendices B and C give the baseline implementations, training configuration, reward and loss definitions, response format, and the target-case bootstrap procedure with fixed seeds.

ETHICS STATEMENT

All images come from public, de-identified datasets; no new patient data were collected. Clinicians advised on the construction rules and reviewed the displayed qualitative examples, as described in Appendix A.5. A deferral indicates insufficient evidence in the given image, not clinical absence in the patient, and the models studied here are research prototypes not intended for clinical use.

REFERENCES

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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

Yanyuan Chen, Dexuan Xu, Yu Huang, Songkun Zhan, Hanpin Wang, Dongxue Chen, Xueping Wang, Meikang Qiu, and Hang Li. Mimo: A medical vision language model with visual referring multimodal input and pixel grounding multimodal output. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24732–24741. IEEE, 2025.

Junlong Cheng, Bin Fu, Jin Ye, Guoan Wang, Tianbin Li, Haoyu Wang, Ruoyu Li, He Yao, Junren Cheng, JingWen Li, et al. Interactive medical image segmentation: A benchmark dataset and baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20841–20851, 2025.

Geon Choi, Hangyul Yoon, Hyunju Shin, Hyunki Park, Sang Hoon Seo, Eunho Yang, and Edward Choi. Instruction-guided lesion segmentation for chest x-rays with automatically generated large-scale dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1482–1492, 2026.

Xiaoshuang Huang, Lingdong Shen, Jia Liu, Fangxin Shang, Hongxiang Li, Haifeng Huang, and Yehui Yang. Towards a multimodal large language model with pixel-level insight for biomedicine. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3779–3787, 2025a.

Yu Huang, Zelin Peng, Yichen Zhao, Piao Yang, Xiaokang Yang, and Wei Shen. Medseg-r: Reasoning segmentation in medical images with multimodal large language models. *arXiv preprint arXiv:2506.10465*, 2025b.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9579–9589. IEEE, 2024.

Suhang Li, Osamu Yoshie, and Yuya Ieiri. Cfcamo: A counterfactual detect-or-abstain framework for camouflaged object detection. *arXiv preprint arXiv:2606.11231*, 2026a.

Xinzhuo Li, Adheesh Juvekar, Jiaxun Zhang, Xingyou Liu, Muntasir Wahed, Kiet A Nguyen, Yifan Shen, Tianjiao Yu, and Ismini Lourentzou. Counterfactual segmentation reasoning: Diagnosing and mitigating pixel-grounding hallucination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7450–7460, 2026b.

Shuang Liang, Zeqing Wang, Yuxian Li, Xihui Liu, and Han Wang. From pixels to concepts: Do segmentation models understand what they segment? *arXiv preprint arXiv:2605.09591*, 2026.

Anglin Liu, Rundong Xue, Xu R Cao, Yifan Shen, Yi Lu, Xiang Li, Qianqian Chen, and Jintai Chen. Medsam3: Delving into segment anything with medical concepts. *arXiv preprint arXiv:2511.19046*, 2025a.

Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23592–23601. IEEE, 2023.

Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*, 2025b.

Zelin Liu, Dongdong Chen, Yusong Sun, Yuqi Hu, Huang Jie, Sicheng Dong, Xu Han, Hongmei Yi, Qiyuan Bao, and Lichi Zhang. Pathchat-segr1: Reasoning segmentation in pathology via so-grpo. In *The Fourteenth International Conference on Learning Representations*, 2026.

Xiangde Luo, Zihan Li, Shaoting Zhang, Wenjun Liao, and Guotai Wang. Rethinking abdominal organ segmentation (raos) in the clinical scenario: A robustness evaluation benchmark with challenging cases. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 531–541. Springer, 2024.

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

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

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

Jiao Tang, Bo Qian, Peng Wan, Wei Shao, and Daoqiang Zhang. Ltse: Language-guided tissue referring segmentation in pathology images with adaptive expert mixture. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 405–415. Springer, 2025.

Qinyue Tong, Ziqian Lu, Jun Liu, Yangming Zheng, and Zhe-Ming Lu. Medisee: reasoning-based pixel-level perception in medical images. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 2742–2751, 2025.

Qinyue Tong, Xiaozhen Wang, Ziqian Lu, Jun Liu, Yunlong Yu, and Zheming Lu. Medveriseg: Teaching lisa-like medical segmentation models to verify query validity without extra training. *arXiv preprint arXiv:2604.10242*, 2026.

Runpu Wei, Zijin Yin, Kongming Liang, Min Min, Chengwei Pan, Gang Yu, Haonan Huang, Yan Liu, and Zhanyu Ma. Polyp-e: Benchmarking the robustness of deep segmentation models via polyp editing. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 3828–3831. IEEE, 2024.

Tsung-Han Wu, Giscard Biamby, David Chan, Lisa Dunlap, Ritwik Gupta, Xudong Wang, Joseph E Gonzalez, and Trevor Darrell. See, say, and segment: Teaching lmms to overcome false premises. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13459–13469. IEEE, 2024.

Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. Gsva: Generalized segmentation via multimodal large language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3858–3869. IEEE, 2024.

Zhonghao Yan, Muxi Diao, Yuxuan Yang, Ruoyan Jing, Jiayuan Xu, Kaizhou Zhang, Lele Yang, Yanxi Liu, Kongming Liang, and Zhanyu Ma. Medreasoner: Reinforcement learning drives reasoning grounding from clinical thought to pixel-level precision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11577–11585, 2026.

Jong-Sung Yoon, Kathy Boutis, Martin R Pecaric, Nancy R Fefferman, K Anders Ericsson, and Martin V Pusic. A think-aloud study to inform the design of radiograph interpretation practice: J. yoon et al. *Advances in Health Sciences Education*, 25(4):877–903, 2020.

Yuan Zhao, Youwei Pang, Jiaming Zuo, Wei Ji, Kailai Zhou, Bin Fan, Yunkang Cao, Lihe Zhang, Xiaofeng Liu, Huchuan Lu, et al. Conceptseg-r1: Segment any concept via meta-reinforcement learning. *arXiv preprint arXiv:2605.20385*, 2026.

Kai Zhu, Wentao Wang, Rongkui Luo, Danjun Song, Xi Wang, Qiang Gao, Jia Fan, Jian Zhou, Shengxiang Rao, and Xiaoying Wang. Newly detected liver nodules with a history of colorectal cancer: are they metastatic? review of 2,632 cases in a single center. *Annals of Translational Medicine*, 9(13):1079, 2021.

Lanyun Zhu, Tianrun Chen, Qianxiong Xu, Xuanyi Liu, Deyi Ji, Haiyang Wu, De Wen Soh, and Jun Liu. Popen: Preference-based optimization and ensemble for lvlm-based reasoning segmentation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30231–30240. IEEE, 2025.

Lianghui Zhu, Bin Ouyang, Yuxuan Zhang, Tianheng Cheng, Rui Hu, Haocheng Shen, Longjin Ran, Xiaoxin Chen, Li Yu, Wenyu Liu, et al. Lens: Learning to segment anything with unified reinforced reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 13952–13960, 2026.

A Benchmark Details

A.1 Mismatch Subtypes

Table 5 expands the five levels in Section 3.3. Sequence incompatibility concerns evidence available in that sequence; slice-level absence does not imply absence from the whole case.

Table 5: Mismatch subtypes in MedCounterSeg-Bench. Cross-dataset organ and laterality mismatches occur only in OOD evaluation.

Level	Subtype	Description
L1	Modality Mismatch	The modality cannot depict the queried imaging features.
L2	Anatomical Site Mismatch	The target lies outside the imaged anatomical region.
L3	Target Absent on Slice	The target is absent from the current slice.
L3	Localization Organ Mismatch	A visible organ resembles the queried organ.
L3	Cross-dataset Organ Mismatch (Only OOD)	An OOD image shows another organ instead of the target.
L4	Lesion vs. Normal Anatomy	Normal anatomy is visible, but the queried lesion is unsupported.
L4	Lesion Type Mismatch	A lesion is present, but its type differs from the query.
L5	Sequence Mismatch	The sequence lacks evidence for the queried finding.
L5	Organ Substructure Mismatch	Another substructure is visible instead of the queried one.
L5	Laterality Mismatch (Only OOD)	The contralateral structure is visible instead of the queried side.

A.2 Data Splits

Table 6 reports the paired benchmark splits. Each pair contains one supported and one unsupported query. The paired benchmark uses 62 source datasets from IMed-361M; including SPPIN2023, used only for Stage I, brings the total to 63. The positive pool for Stage I contains 7,709 training examples; its validation and test examples coincide with the supported members of the paired evaluation sets. The 47-concept source vocabulary yields 42 requested concepts in the paired benchmark. AMOS2022 and WORD contribute 100 and 60 OOD pairs, respectively, and contribute no training or reference images.

Table 6: Dataset splits in MedCounterSeg-Bench. Counts refer to query pairs and their two constituent queries; cases are disjoint across training, validation, and test partitions.

Split	Pairs	Queries
Training	4,697	9,394
Validation	561	1,122
In-distribution test	600	1,200
OOD–AMOS2022	100	200
OOD–WORD	60	120
Total	6,018	12,036

A.3 Benchmark Composition

Figure 4 shows how the paired benchmark is distributed across mismatch levels and imaging modalities. Levels 1 and 2 account for 4,514 of the 6,018 pairs (75.0%), whereas Levels 4 and 5 contain 181 and 201 pairs, respectively. MRI and CT together account for 3,696 pairs (61.4%); X-ray contributes 64 pairs (1.1%). Aggregate scores should be read alongside results for individual mismatch levels and modalities, especially for groups with few examples. These counts pool all splits.

Figure 5 distinguishes subtypes within these levels. Level 4 includes 145 lesion-versus-normal-anatomy pairs and 36 lesion-type mismatches. Level 5 separates anatomical substructure, sequence,

and laterality; the sequence cases use MRI and the laterality cases use CT. These distinctions test specific anatomical and imaging constraints.

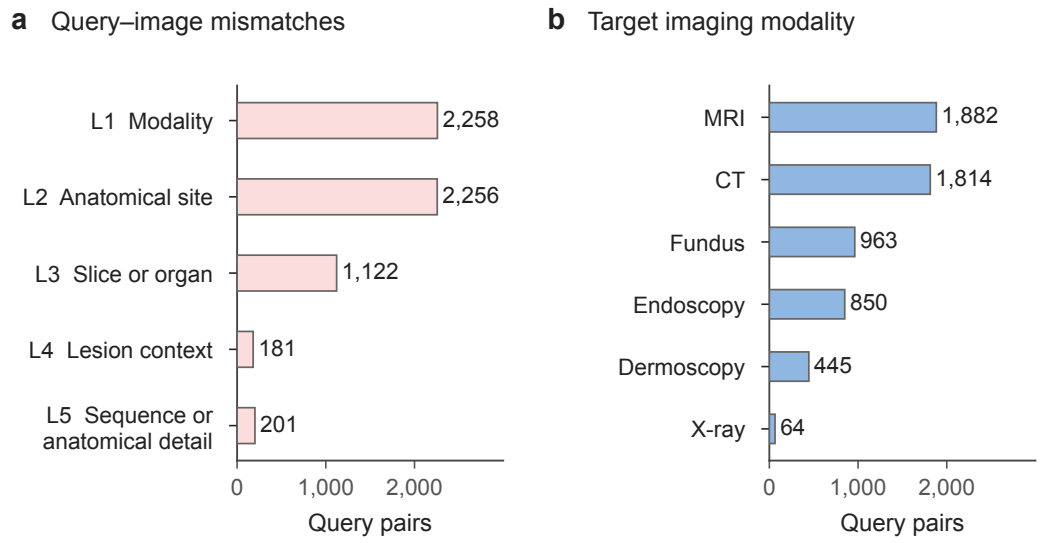

Figure 4: Composition of MedCounterSeg-Bench across training, validation, in-distribution test, and OOD splits (6,018 pairs; 12,036 queries). (a) Pairs by mismatch level, assigned from the unsupported query. (b) Pairs by target-image modality. Each pair contains one supported and one unsupported query on the same target image. Counts refer to query pairs, not independent patients.



Figure 5: Distribution of query–image mismatches. From inner to outer rings: level, subtype, and imaging modality. Sector size represents query-pair counts across all splits.

A.4 Coverage across Modalities and Concepts

Figure 6 separates broad compatibility checks from finer, within-modality distinctions. Imaging-modality and anatomical-site mismatches span all six modalities. CT and MRI also include distinctions involving organ identity, lesion context, and anatomical substructures; sequence mismatches occur in MRI. This breakdown distinguishes rejecting an incompatible image context from resolving a request within the same modality.

Figure 7 tracks these distinctions by requested concept. Left- and right-kidney requests occur in anatomical-site, current-slice, localization-organ, and laterality mismatches: the requested organ can be the same while the reason for deferral differs. Kidney-tumor requests also distinguish normal anatomy from a different lesion type.

The 6,018 pairs draw on 62 source datasets; the largest contributes 7.2%. Across all splits, each of the 42 requested concepts has equal supported and unsupported query counts. Concept identity is therefore balanced across support labels and cannot predict the label.

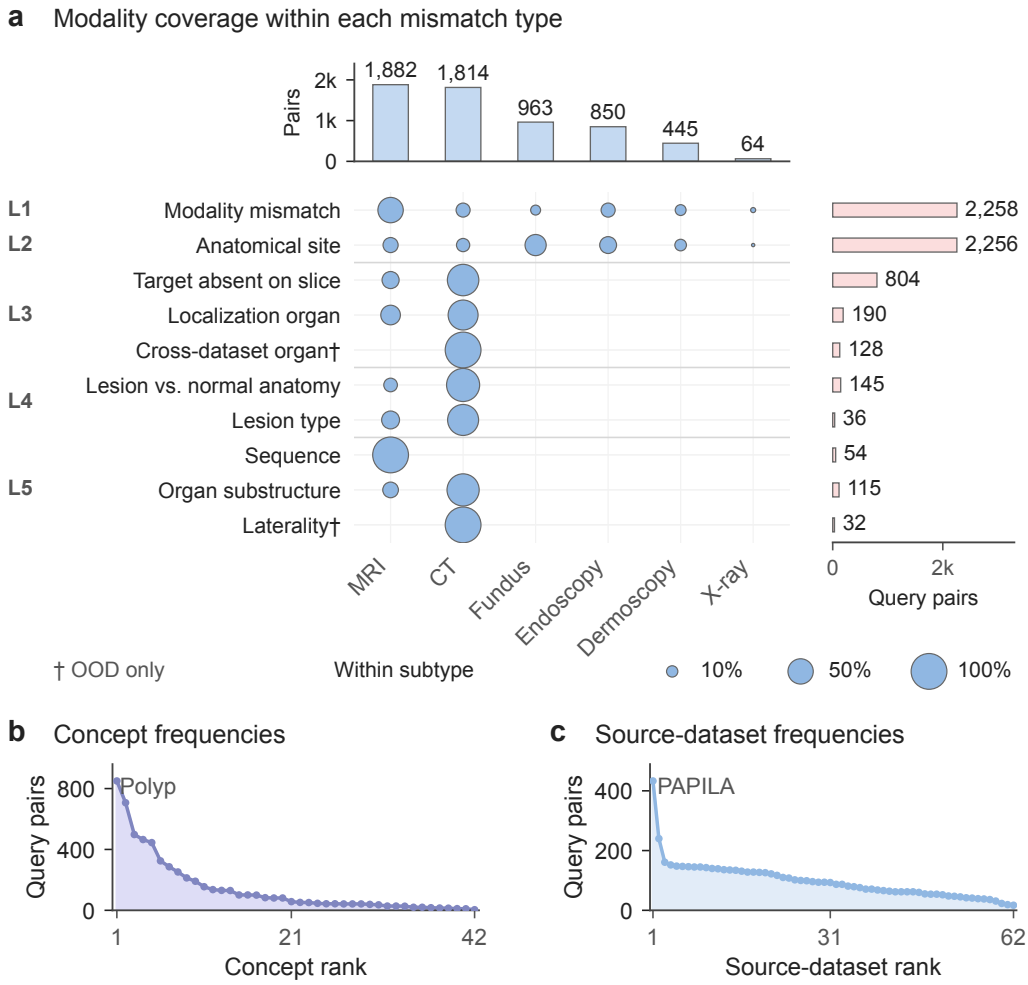

Figure 6: Modality and concept coverage across all 6,018 query pairs in the training, validation, Test, and OOD splits. (a) Circle area encodes the modality share within each mismatch subtype; marginal bars show absolute pair counts. (b,c) Requested concepts and source datasets are ranked separately by pair count. Each pair is counted once, using its unsupported query to identify the subtype and requested concept. Blank cells denote no observed pairs; a dagger marks OOD-only subtypes.

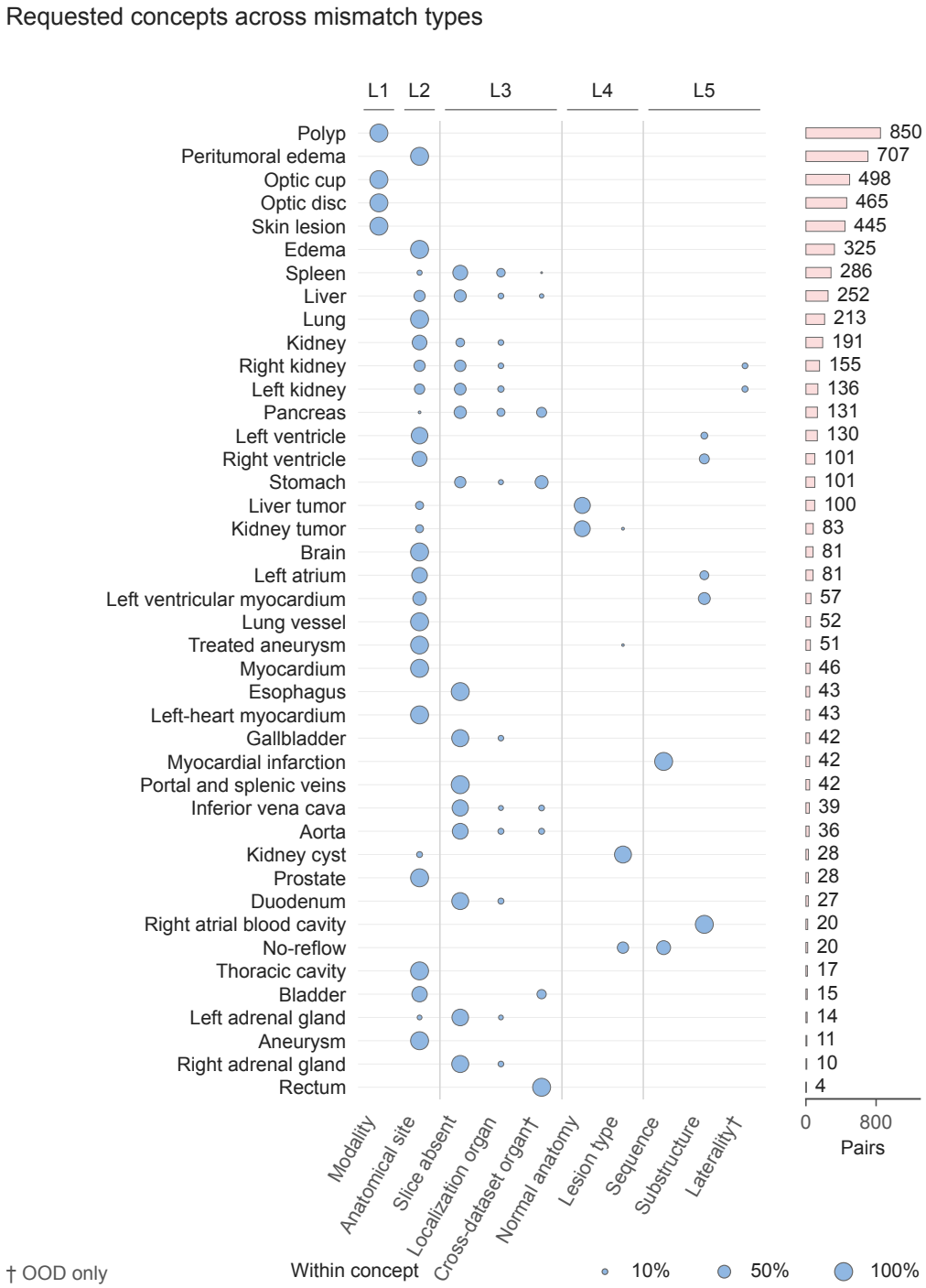

Figure 7: Coverage of all 42 requested concepts across ten mismatch subtypes. Rows are ordered by frequency among the 6,018 unsupported queries across all splits, with one query per pair. Circle area encodes the subtype share within each concept; bars show absolute pair counts. Blank cells denote no observed pairs. A dagger marks OOD-only subtypes.

A.5 Clinical Review and Iterative Quality Control

Clinical input shaped the benchmark's construction rules and taxonomy. In a series of meetings, clinicians and dataset curators discussed how to construct unsupported queries that make a plausible request but lack sufficient support in the target image. The taxonomy grew from three groups to the current five levels through these reviews.

Each review iteration used 100 randomly sampled examples, with coverage across levels balanced as evenly as practicable. Clinicians assessed whether the proposed examples were reasonable and reliable, and dataset curators revised the queries and construction rules in response.

Human reviewers checked query formulations, concept assignments, masks, and case separation. We excluded ambiguous synonym, overlap, and containment relations; concepts that failed these checks were revised and reviewed again.

Clinical feedback also identified image display orientations that did not follow customary clinical viewing conventions. We corrected the affected displays during dataset construction. Clinicians specifically proposed laterality queries, adding a way to test whether a model distinguishes the requested anatomical side from the visible contralateral structure. This feedback helped the team refine both the image presentation and the distinctions tested by the queries.

The qualitative examples presented at the end of this appendix underwent a separate clinical review. Clinicians examined the model-generated evidence descriptions in these examples and endorsed them. This review applies to the displayed examples and is separate from the quantitative evaluation of model performance on the full benchmark.

A.6 Scope and Limitations

The level composition of MedCounterSeg-Bench follows how unsupported queries can be constructed at scale. Modality and anatomical-site mismatches (Levels 1–2) can be formed for every target image and therefore account for most pairs, whereas lesion-context and fine anatomical mismatches (Levels 4–5) require a related structure to be visible in the same image and are correspondingly rarer. Reliable rejection of coarse mismatches is the entry requirement for any deployed system; the fine-grained levels are where current models, including ours, remain far from clinical reliability (Table 10). We regard these levels as the benchmark's frontier: because the taxonomy, reference-set format, and paired protocol are fixed, future releases can enlarge Levels 4–5 without changing the evaluation. The OOD split currently covers CT organs and laterality, and our training objective is evaluated on one concept-reasoning backbone.

B Experimental Details

This appendix describes the baselines, the additional backbone, and additional results. Training configuration is given in Appendix C.2.

B.1 Baseline Implementations

All baselines are evaluated on the same target images and queries as our method. Unlike ConceptSeg-R1 and our method, they receive only the target image and the query; the reference images are not part of their input. Methods with a trainable segmentation path use the same MedCounterSeg-Bench training split as ours. ConceptSeg-R1, with either backbone, is trained on the same supported queries as our Stage I. LENS and Seg-Zero have no rejection interface and are trained on supported queries only: LENS aligns its segmentation path on the supported pool and then runs reinforcement learning on supported Stage II queries, and Seg-Zero runs reinforcement learning on the same supported queries. GSVA is fine-tuned from its official GRES weights on the same 800 supported and 800 unsupported Stage II queries as our method, learning its `[SEG]` and `[REJ]` tokens; we report the label-free protocol, in which the model chooses the decision token without ground-truth scaffolding. SAM3-Agent, MedSAM3-Agent, and MedVeriSeg are run without training on our data, and SESAME is evaluated with its released weights, likewise without training on our data.

Rejection interfaces differ across methods. GSVA emits a rejection token; SESAME answers a false-premise query in text; the agents call a no-mask tool; MedVeriSeg returns an invalid-query decision from its verifier. LENS, Seg-Zero, and ConceptSeg-R1 always segment, so their decision metrics are undefined and their outputs on unsupported queries enter only the foreground-probability diagnostics. SAM3-Agent and MedSAM3-Agent follow the SAM 3 Agent protocol with a Qwen2.5-VL controller (3B or 7B) and SAM 3 or Medical-SAM3 (2D) as the segmenter; MedSAM3-Agent is our controlled substitution of the segmenter under an unchanged controller and tool set. Agent runs that exhaust the tool budget or return unparseable output are counted as errors, not as deferrals. Qwen2.5 denotes Qwen2.5-VL; MedVeriSeg additionally uses a Qwen3-VL-30B-A3B verifier.

B.2 Additional Backbone

Table 7 reports MedCounterSeg-RL with Qwen2.5-VL-3B on both splits: 75.83% paired accuracy on the Test split and 8.75% correct deferral on OOD queries. The two configurations differ in both model family and size. Dice over all supported queries appears in Figure 8, and target-case bootstrap intervals in Appendix B.7.

Table 7: MedCounterSeg-RL with Qwen2.5-VL-3B on both evaluation splits. Scores are percentages; mDice and mIoU follow the main-table definition (Section 5.1).

Split	mDice	mIoU	Paired accuracy	Action accuracy	Correct deferral	False deferral	Errors
Test	71.43	62.91	75.83	87.08	81.17	6.67	0.25
OOD	67.65	59.45	7.50	49.06	8.75	10.00	0.31

B.3 Segmentation–Deferral Balance

Figure 8 plots Dice over all supported queries against paired accuracy for decision-enabled configurations.

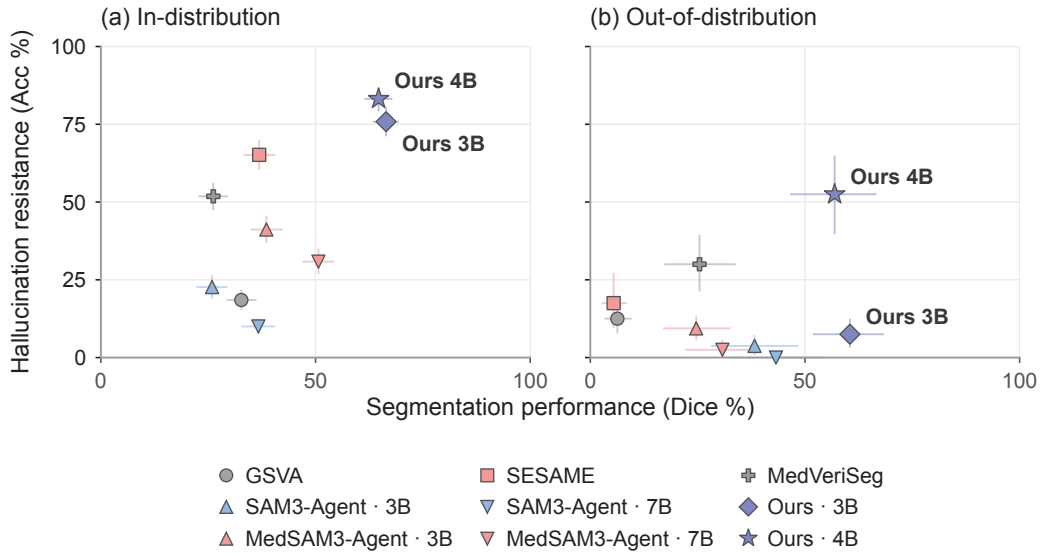

Figure 8: **Segmentation and query-mismatch handling on Test and OOD data.** Dice is computed over all supported queries. Main-table metric definitions are given in Section 5.1. Acc denotes paired accuracy, requiring both queries of a pair to receive correct, valid decisions; the vertical axis label "hallucination resistance" refers to this paired accuracy. Thin lines are marginal 95% intervals from 20,000 target-case bootstrap resamples. Only decision-enabled configurations are plotted. For Qwen-based methods, 3B/7B denote Qwen2.5 and 4B denotes Qwen3.5.

B.4 Additional Results

Table 8 complements Table 2 with correct and false deferral rates, decision macro-F1, invalid or failed responses, and raw foreground probability before action gating. The separate deferral rates distinguish rejection of unsupported queries from mistaken rejection of supported ones. Foreground probability distinguishes suppression of unsupported foreground from its removal by the final decision.

Table 8: Additional results on the paired Test split.

Method	**Backbone**	Macro-F1 ↑	Correct deferral ↑	False deferral ↓	Errors ↓	Raw ρ (−) ↓
Without Support Decisions						
LENS	Qwen2.5-3B	–	–	–	85	5.56
Seg-Zero	Qwen2.5-7B	–	–	–	40	11.88
ConceptSeg-R1	Qwen2.5-3B	–	–	–	0	6.58
	Qwen3.5-4B	–	–	–	168	5.87
With Support Decisions						
GSVA	LLaVA-7B	49.83	22.17	12.17	34	3.74
SESAME	LLaVA-1.5-7B	80.88	91.50	29.33	0	0.37
SAM3-Agent	Qwen2.5-3B	62.25	56.50	18.33	449	2.89
	Qwen2.5-7B	47.38	14.67	1.67	661	17.04
MedSAM3-Agent	Qwen2.5-3B	67.79	58.33	20.67	79	3.27
	Qwen2.5-7B	60.54	32.17	2.67	18	8.26
MedVeriSeg	LLaVA-Med-7B	73.32	75.50	28.83	0	2.89
Ours	Qwen2.5-3B	87.15	81.17	6.67	3	3.02
	Qwen3.5-4B	91.17	91.67	9.33	0	0.79

Macro-F1, deferral rates, and Raw ρ are percentages. Correct and false deferral are measured on unsupported and supported queries, respectively. Errors count invalid responses, including formatting errors, and failed predictions under each method's evaluation interface. Raw ρ is the mean foreground probability on unsupported queries before action gating. A dash indicates no explicit decision interface. Backbone names follow Table 2.

Out-of-distribution diagnostics. Table 9 gives the same additional metrics for AMOS2022 and WORD. Our 3B model produces little foreground on unsupported queries but rarely defers on this split. SAM3-Agent's mDice is computed over the few queries it completes; its 3B and 7B models fail on 67.81% and 69.69% of OOD queries (Table 3). These results pool cross-dataset organ and laterality mismatches.

Table 9: Additional out-of-distribution results on AMOS2022 and WORD.

Method	**Backbone**	Macro-F1 ↑	Correct deferral ↑	False deferral ↓	Errors ↓	Raw ρ (−) ↓
Without Support Decisions						
LENS	Qwen2.5-3B	–	–	–	5	1.62
Seg-Zero	Qwen2.5-7B	–	–	–	11	5.45
ConceptSeg-R1	Qwen2.5-3B	–	–	–	0	1.28
	Qwen3.5-4B	–	–	–	16	0.91
With Support Decisions						
GSVA	LLaVA-7B	43.30	20.00	23.12	12	0.90
SESAME	LLaVA-1.5-7B	47.15	76.25	75.00	0	0.44
SAM3-Agent	Qwen2.5-3B	37.82	8.75	3.75	217	0.07
	Qwen2.5-7B	33.95	0.62	0.62	223	0.05
MedSAM3-Agent	Qwen2.5-3B	40.91	13.12	7.50	54	2.55
	Qwen2.5-7B	35.52	2.50	1.88	2	3.14
MedVeriSeg	LLaVA-Med-7B	61.21	42.50	16.88	0	0.85
Ours	Qwen2.5-3B	39.22	8.75	10.00	1	0.60
	Qwen3.5-4B	74.35	71.25	22.50	0	0.71

Macro-F1, deferral rates, and Raw ρ are percentages; Errors are counts. Definitions follow Table 8.

B.5 Results by Mismatch Level

Figure 9 shows that both MedCounterSeg-RL models reject roughly 97% or more of Level 1 and Level 2 queries. The largest gap is at Level 3: 30.1% for Qwen2.5-VL-3B versus 76.7% for Qwen3.5-4B. At Level 4 both models reject 41.2% of queries, and at Level 5 they reject 5.3% and 15.8%, respectively; lesion-context, sequence, and substructure mismatches are the remaining targets for the benchmark.

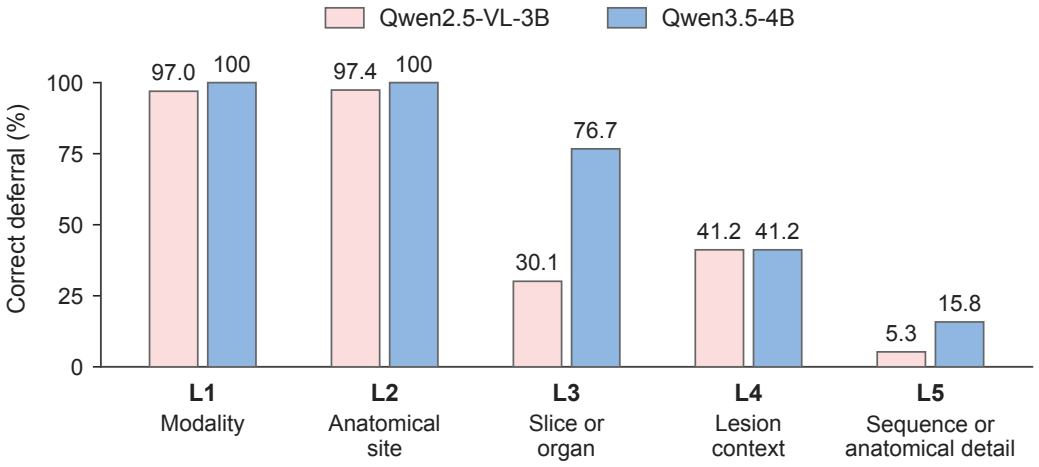

Figure 9: Correct deferral by mismatch level on the in-distribution Test split. Both MedCounterSeg-RL models are evaluated at Stage II step 100 on the unsupported Test queries. An invalid response counts as an error; an empty mask alone does not count as correct deferral. Bars show observed rates; the level composition across all splits is given in Appendix A.3. The levels denote mismatch levels, not a strict ordering of difficulty.

Table 10 breaks the Test-split decision metrics of every decision-enabled configuration down by mismatch level. Levels follow Section 3.3 and Table 5. Each level contains the unsupported queries of that type and their paired supported queries on the same target images; invalid responses count as errors, not as deferrals.

Table 10: Paired accuracy, correct deferral, and false deferral by mismatch level on the Test split. Each level pairs its unsupported queries with their supported counterparts on the same target images.

Configuration	L1	L2	L3	L4	L5	All
Paired accuracy ↑						
Ours (Qwen3.5-4B)	96.1	95.2	51.5	23.5	5.3	83.2
Ours (Qwen2.5-3B)	90.0	93.5	24.3	35.3	5.3	75.8
SESAME	72.3	82.2	26.2	35.3	10.5	65.2
MedVeriSeg	56.3	59.6	35.0	35.3	10.5	51.8
GSVA	26.0	13.5	19.4	0.0	0.0	18.5
MedSAM3-Agent (3B)	39.8	60.4	10.7	29.4	0.0	41.2
MedSAM3-Agent (7B)	33.3	45.6	2.9	0.0	0.0	30.8
SAM3-Agent (3B)	20.8	34.4	4.8	23.5	0.0	22.7
SAM3-Agent (7B)	10.0	14.3	1.9	0.0	10.5	10.0
Correct deferral ↑						
Ours (Qwen3.5-4B)	100.0	100.0	76.7	41.2	15.8	91.7
Ours (Qwen2.5-3B)	97.0	97.4	30.1	41.2	5.3	81.2
SESAME	98.3	99.6	73.8	52.9	42.1	91.5
MedVeriSeg	73.2	85.7	67.0	58.8	42.1	75.5
GSVA	27.7	14.3	32.0	5.9	10.5	22.2
MedSAM3-Agent (3B)	58.9	79.6	21.4	41.2	10.5	58.3
MedSAM3-Agent (7B)	35.1	47.0	3.9	0.0	0.0	32.2
SAM3-Agent (3B)	58.4	76.1	19.4	41.2	10.5	56.5
SAM3-Agent (7B)	15.2	20.9	2.9	0.0	10.5	14.7
False deferral ↓						
Ours (Qwen3.5-4B)	3.9	4.8	26.2	17.6	31.6	9.3
Ours (Qwen2.5-3B)	7.8	3.5	9.7	17.6	5.3	6.7
SESAME	26.0	17.4	60.2	23.5	52.6	29.3
MedVeriSeg	22.5	29.6	41.8	23.5	31.6	28.8
GSVA	6.9	8.7	28.2	17.6	26.3	12.2
MedSAM3-Agent (3B)	22.1	17.4	26.2	11.8	21.1	20.7
MedSAM3-Agent (7B)	2.2	1.7	6.8	0.0	0.0	2.7
SAM3-Agent (3B)	22.9	13.0	23.3	5.9	10.5	18.3
SAM3-Agent (7B)	1.7	0.4	4.8	0.0	0.0	1.7

Values are percentages of the queries in each level. 3B/7B for the agents denote the Qwen2.5-VL controller size. The level composition across all splits is given in Appendix A.3.

B.6 Segmentation and Deferral across Modalities

On fundus, endoscopy, and dermoscopy images, our 4B model correctly defers every unsupported Test query and accepts every supported query (Figure 10). Its final Dice over all supported queries is 65.5%, 85.2%, and 91.4%, compared with 61.8%, 86.9%, and 91.3% for ConceptSeg-R1 with the same backbone. Selective refusal thus coexists with competitive segmentation: every supported query in these modalities is accepted.

CT accounts for 51 of the 4B model's 56 false deferrals. On CT its correct deferral rate is higher than the 3B model's (83.1% versus 61.2%) and its Dice over all supported queries is lower (55.2% versus 65.8%), a tradeoff that the paired evaluation makes visible. Query and mismatch composition differs across modalities.

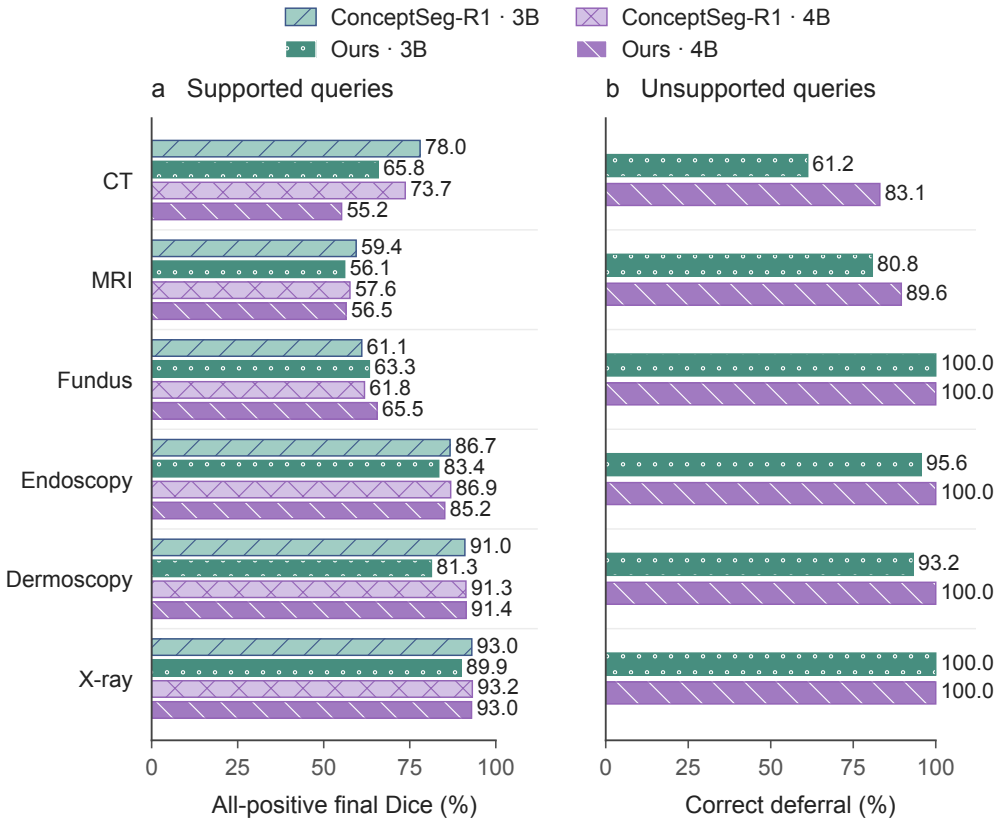

Figure 10: Segmentation and deferral have different modality-specific patterns on the paired Test split. The 3B and 4B models use Qwen2.5-VL-3B and Qwen3.5-4B, respectively; each method retains its own inference interface. (a) Final-mask Dice over all supported queries. (b) Valid explicit deferral among unsupported queries; ConceptSeg-R1 has no such interface and is not plotted. CT, MRI, fundus, endoscopy, dermoscopy, and X-ray contribute 183, 182, 96, 91, 44, and 4 pairs, respectively; the small X-ray subset should be interpreted descriptively.

B.7 Uncertainty Across Target Cases

We estimate uncertainty for every configuration in Tables 2 and 3 using 20,000 paired target-case bootstrap resamples. Test and OOD queries come from 451 and 35 recorded target cases, respectively. Within each split, we sample cases with replacement and retain all queries from each sampled case. Every method uses the same resampled cases. Each replicate recomputes the original query- or pair-weighted metric, rather than an equal-case average. We report the 2.5th and 97.5th percentiles, using seed 20260925 for Test and 20260926 for OOD.

These are pointwise intervals for fixed checkpoints, not variation across training seeds or guarantees across unseen institutions. Cases are identified by their source dataset and recorded case ID; datasets themselves are not resampled. Decision metrics remain undefined for methods without an explicit support-decision interface. An empirical interval of [0, 0] when no pair succeeds does not establish a zero population success rate. Accepted-positive scores condition on each model's own accepted set.

The tables below include all main-table configurations and the additional Qwen2.5-VL-3B configuration in Table 7. Table 15 reports paired differences relative to our Qwen3.5-4B model. Intervals are not adjusted for multiple comparisons, and we do not interpret an interval crossing zero as proof of equivalence.

Table 11: Test segmentation scores with 95% target-case bootstrap intervals.

Method / backbone	Accepted Dice ↑	Accepted IoU ↑	Dice, all supported ↑	IoU, all supported ↑
LENS	53.14	43.68	50.12	41.21
Qwen2.5-3B	[49.81, 56.46]	[40.52, 46.88]	[46.79, 53.45]	[38.09, 44.35]
Seg-Zero	39.47	31.98	37.69	30.54
Qwen2.5-7B	[36.21, 42.88]	[28.98, 35.15]	[34.53, 40.99]	[27.66, 33.56]
ConceptSeg-R1	72.03	64.22	72.03	64.22
Qwen2.5-3B	[69.07, 74.85]	[61.23, 67.11]	[69.07, 74.85]	[61.23, 67.11]
ConceptSeg-R1	71.66	63.73	70.35	62.56
Qwen3.5-4B	[68.63, 74.56]	[60.71, 66.65]	[67.25, 73.30]	[59.47, 65.52]
GSVA	37.74	31.32	32.71	27.14
LLaVA-7B	[34.11, 41.47]	[28.07, 34.71]	[29.29, 36.24]	[24.12, 30.28]
SESAME	52.18	44.48	36.87	31.43
LLaVA-1.5-7B	[48.34, 56.01]	[40.84, 48.14]	[33.17, 40.62]	[28.07, 34.87]
SAM3-Agent	57.56	50.53	25.90	22.74
Qwen2.5-3B	[52.01, 62.78]	[45.11, 55.69]	[22.25, 29.50]	[19.34, 26.09]
SAM3-Agent	60.80	53.85	36.68	32.49
Qwen2.5-7B	[56.00, 65.26]	[49.20, 58.24]	[32.70, 40.56]	[28.79, 36.13]
MedSAM3-Agent	52.69	45.87	38.55	33.56
Qwen2.5-3B	[48.33, 57.06]	[41.70, 50.07]	[34.83, 42.29]	[30.07, 37.08]
MedSAM3-Agent	52.69	46.13	50.67	44.36
Qwen2.5-7B	[48.90, 56.50]	[42.49, 49.79]	[46.95, 54.39]	[40.80, 47.91]
MedVeriSeg	36.80	29.46	26.19	20.96
LLaVA-Med-7B	[32.50, 41.01]	[25.61, 33.32]	[22.72, 29.65]	[17.90, 24.06]
Ours	71.43	62.91	66.43	58.51
Qwen2.5-3B	[68.66, 74.18]	[60.06, 65.76]	[63.35, 69.40]	[55.47, 61.45]
Ours	71.35	63.34	64.69	57.43
Qwen3.5-4B	[68.21, 74.38]	[60.19, 66.42]	[61.40, 67.90]	[54.25, 60.55]

Values are percentages; intervals are in brackets. Accepted: valid, accepted positive queries. All supported: every supported query, including deferrals and invalid outputs. Backbone abbreviations follow Table 2.

Table 12: Test support-decision scores with 95% target-case bootstrap intervals.

Method / backbone	Paired accuracy ↑	Action accuracy ↑	Correct deferral ↑	False deferral ↓
LENS				
Qwen2.5-3B	–	–	–	–
Seg-Zero				
Qwen2.5-7B	–	–	–	–
ConceptSeg-R1				
Qwen2.5-3B	–	–	–	–
ConceptSeg-R1				
Qwen3.5-4B	–	–	–	–
GSVA	18.50	54.42	22.17	12.17
LLaVA-7B	[15.38, 21.76]	[52.32, 56.55]	[18.70, 25.72]	[9.54, 14.88]
SESAME	65.17	81.08	91.50	29.33
LLaVA-1.5-7B	[60.40, 69.95]	[78.34, 83.82]	[88.69, 94.13]	[24.79, 33.92]
SAM3-Agent	22.67	50.75	56.50	18.33
Qwen2.5-3B	[19.02, 26.52]	[47.70, 53.80]	[51.84, 61.14]	[15.15, 21.63]
SAM3-Agent	10.00	37.50	14.67	1.67
Qwen2.5-7B	[7.58, 12.59]	[34.80, 40.20]	[11.75, 17.81]	[0.70, 2.74]
MedSAM3-Agent	41.17	65.75	58.33	20.67
Qwen2.5-3B	[36.83, 45.57]	[62.86, 68.66]	[53.86, 62.90]	[17.32, 24.04]
MedSAM3-Agent	30.83	64.17	32.17	2.67
Qwen2.5-7B	[26.81, 35.04]	[61.94, 66.42]	[28.02, 36.47]	[1.39, 4.08]
MedVeriSeg	51.83	73.33	75.50	28.83
LLaVA-Med-7B	[47.47, 56.23]	[70.68, 75.98]	[71.78, 79.30]	[24.83, 32.90]
Ours	75.83	87.08	81.17	6.67
Qwen2.5-3B	[71.12, 80.42]	[84.54, 89.56]	[76.65, 85.57]	[4.61, 8.89]
Ours	83.17	91.17	91.67	9.33
Qwen3.5-4B	[79.15, 86.97]	[88.95, 93.25]	[88.41, 94.65]	[6.64, 12.20]

Values are percentages; intervals are in brackets. Paired accuracy requires both queries of the same image pair to receive valid, correct decisions. Dashes: no support-decision interface.

Table 13: OOD segmentation scores with 95% target-case bootstrap intervals.

Method / backbone	Accepted Dice ↑	Accepted IoU ↑	Dice, all supported ↑	IoU, all supported ↑
LENS	18.72	14.02	18.26	13.67
Qwen2.5-3B	[11.49, 26.88]	[8.26, 20.57]	[11.16, 26.36]	[8.01, 20.17]
Seg-Zero	17.57	13.33	17.02	12.91
Qwen2.5-7B	[11.48, 24.14]	[8.44, 18.55]	[11.18, 23.28]	[8.23, 17.83]
ConceptSeg-R1	73.26	64.68	73.26	64.68
Qwen2.5-3B	[66.64, 78.96]	[57.99, 70.83]	[66.64, 78.96]	[57.99, 70.83]
ConceptSeg-R1	75.29	66.64	72.00	63.73
Qwen3.5-4B	[67.01, 82.15]	[58.55, 73.75]	[63.97, 79.06]	[55.85, 71.02]
GSVA	8.62	6.15	6.31	4.50
LLaVA-7B	[4.54, 13.11]	[3.14, 9.48]	[3.28, 9.65]	[2.28, 6.97]
SESAME	21.74	15.50	5.43	3.88
LLaVA-1.5-7B	[12.27, 34.97]	[8.32, 25.85]	[2.68, 8.46]	[1.85, 6.12]
SAM3-Agent	83.84	76.56	38.25	34.93
Qwen2.5-3B	[72.26, 91.60]	[65.28, 85.13]	[28.12, 48.51]	[25.53, 44.62]
SAM3-Agent	81.40	74.36	43.24	39.51
Qwen2.5-7B	[71.45, 88.91]	[64.51, 82.27]	[32.42, 54.42]	[29.31, 50.20]
MedSAM3-Agent	33.45	29.50	24.67	21.76
Qwen2.5-3B	[23.24, 43.98]	[19.83, 39.58]	[16.95, 32.66]	[14.49, 29.32]
MedSAM3-Agent	31.56	28.01	30.77	27.31
Qwen2.5-7B	[22.68, 40.29]	[19.69, 36.39]	[22.04, 39.42]	[19.11, 35.57]
MedVeriSeg	30.61	24.86	25.44	20.67
LLaVA-Med-7B	[21.39, 39.61]	[16.89, 32.87]	[17.24, 33.86]	[13.54, 28.17]
Ours	67.65	59.45	60.47	53.13
Qwen2.5-3B	[60.20, 74.40]	[52.01, 66.32]	[51.85, 68.36]	[45.11, 60.71]
Ours	73.43	65.42	56.91	50.70
Qwen3.5-4B	[63.56, 81.82]	[55.82, 74.04]	[46.55, 66.66]	[41.07, 59.85]

Values are percentages; intervals are in brackets. Accepted: valid, accepted positive queries. All supported: every supported query, including deferrals and invalid outputs. Backbone abbreviations follow Table 2.

Table 14: OOD support-decision scores with 95% target-case bootstrap intervals.

Method / backbone	Paired accuracy ↑	Action accuracy ↑	Correct deferral ↑	False deferral ↓
LENS				
Qwen2.5-3B	–	–	–	–
Seg-Zero				
Qwen2.5-7B	–	–	–	–
ConceptSeg-R1				
Qwen2.5-3B	–	–	–	–
ConceptSeg-R1				
Qwen3.5-4B	–	–	–	–
GSVA	12.50	46.56	20.00	23.12
LLaVA-7B	[7.80, 17.65]	[42.52, 50.68]	[13.38, 27.45]	[16.98, 29.34]
SESAME	17.50	50.62	76.25	75.00
LLaVA-1.5-7B	[9.56, 27.22]	[44.64, 57.23]	[66.88, 84.89]	[63.53, 85.06]
SAM3-Agent	3.75	27.19	8.75	3.75
Qwen2.5-3B	[0.73, 7.19]	[21.04, 33.33]	[4.40, 13.58]	[1.27, 6.54]
SAM3-Agent	0.00	26.88	0.62	0.62
Qwen2.5-7B	[0.00, 0.00]	[20.81, 32.80]	[0.00, 1.99]	[0.00, 1.99]
MedSAM3-Agent	9.38	43.44	13.12	7.50
Qwen2.5-3B	[5.46, 13.51]	[39.77, 46.82]	[7.55, 19.14]	[4.07, 11.18]
MedSAM3-Agent	2.50	50.00	2.50	1.88
Qwen2.5-7B	[0.00, 5.63]	[48.39, 51.83]	[0.00, 5.63]	[0.00, 4.07]
MedVeriSeg	30.00	62.81	42.50	16.88
LLaVA-Med-7B	[21.25, 39.43]	[57.72, 68.13]	[32.18, 52.87]	[9.38, 25.16]
Ours	7.50	49.06	8.75	10.00
Qwen2.5-3B	[3.23, 12.50]	[45.71, 52.54]	[4.14, 13.99]	[4.19, 17.03]
Ours	52.50	74.38	71.25	22.50
Qwen3.5-4B	[39.63, 64.94]	[67.09, 81.37]	[60.24, 81.70]	[12.75, 33.82]

Values are percentages; intervals are in brackets. Paired accuracy requires both queries of the same image pair to receive valid, correct decisions. Dashes: no support-decision interface.

Table 15: Paired differences: Ours (Qwen3.5-4B) minus each comparison configuration, with 95% target-case bootstrap intervals.

Method / backbone	Paired accuracy Test	Paired accuracy OOD	Dice, all supported Test	Dice, all supported OOD
LENS	–	–	+14.56	+38.65
Qwen2.5-3B			[+11.17, +17.93]	[+26.35, +50.26]
Seg-Zero	–	–	+27.00	+39.89
Qwen2.5-7B			[+23.76, +30.15]	[+28.99, +50.71]
ConceptSeg-R1	–	–	-7.34	-16.35
Qwen2.5-3B			[-9.91, -4.85]	[-24.31, -9.05]
ConceptSeg-R1	–	–	-5.66	-15.09
Qwen3.5-4B			[-8.28, -3.07]	[-24.84, -6.08]
GSVA	+64.67	+40.00	+31.98	+50.60
LLaVA-7B	[+59.94, +69.43]	[+26.95, +52.63]	[+28.35, +35.55]	[+40.77, +60.05]
SESAME	+18.00	+35.00	+27.81	+51.47
LLaVA-1.5-7B	[+14.17, +21.96]	[+19.43, +50.33]	[+24.11, +31.47]	[+40.65, +61.63]
SAM3-Agent	+60.50	+48.75	+38.78	+18.66
Qwen2.5-3B	[+56.08, +64.97]	[+35.17, +61.87]	[+34.70, +42.99]	[+9.58, +27.59]
SAM3-Agent	+73.17	+52.50	+28.01	+13.67
Qwen2.5-7B	[+68.75, +77.50]	[+39.63, +64.94]	[+23.63, +32.39]	[+4.57, +22.50]
MedSAM3-Agent	+42.00	+43.12	+26.14	+32.24
Qwen2.5-3B	[+37.69, +46.39]	[+29.38, +56.36]	[+22.59, +29.69]	[+20.39, +43.61]
MedSAM3-Agent	+52.33	+50.00	+14.01	+26.13
Qwen2.5-7B	[+48.11, +56.52]	[+37.78, +62.09]	[+10.73, +17.34]	[+15.18, +36.85]
MedVeriSeg	+31.33	+22.50	+38.50	+31.47
LLaVA-Med-7B	[+26.88, +35.95]	[+9.09, +36.24]	[+34.67, +42.39]	[+22.59, +40.51]
Ours	+7.33	+45.00	-1.74	-3.56
Qwen2.5-3B	[+3.77, +11.00]	[+32.53, +57.50]	[-4.47, +0.95]	[-13.31, +6.06]

Differences are percentage points; positive values favor Ours (Qwen3.5-4B). Differences use matched case resamples. The Ours (Qwen2.5-3B) row compares our two backbone configurations. Segmentation differences use all supported queries to retain the cost of rejection.

On Test, paired accuracy exceeds SESAME by 18.00 percentage points (95% CI: 14.17–21.96). Correct deferral is similar, while false deferral is 20.00 points lower (95% CI: 15.92–24.17 lower). On OOD, the paired-accuracy difference from MedVeriSeg is 22.50 points (95% CI: 9.09–36.24). Against ConceptSeg-R1 with Qwen3.5-4B, the Test accepted-positive Dice difference is -0.31 points (95% CI: -1.84 to 1.21), whereas the difference over all supported queries is -5.66 points (95% CI: -8.28 to -3.07). These results quantify the tradeoff between deferral and segmentation.

C Response and Training Details

This section gives the response structure and training objectives used in Section 4.

C.1 Response and Mask Interface

Response structure. Stage II and inference require six fields in order: `<rule>`, `<check>`, `<evidence>`, `<action>`, `<bbox>`, and `<answer>`. They contain a visual rule r, verification box, target-specific evidence e, action, target box, and concept phrase, respectively:

$$\mathrm{Parse}(o) = \big(r, \widehat{B}^v, e, a, \widehat{B}^t, c_o\big). \tag{4}$$

References are positive for both target labels; their hidden verification annotation is never supplied to the model. The prompt asks for a brief visual rule, one or two target-specific evidence sentences, and a concise requested-concept noun phrase. Deferral requires an empty target-box field and must not substitute a different visible concept. Stage I uses the teacher-forced alignment response described below.

Coordinates and box validity. The reference mosaic and target coordinate canvas are each 600×600. A valid integer box (x_1, y_1, x_2, y_2) satisfies $0 \le x_1 < x_2 \le 599$ and $0 \le y_1 < y_2 \le 599$. The verification image occupies the top-left panel of the reference mosaic; its predicted box is expressed in mosaic coordinates. The target box uses the target-image coordinate canvas. Mask predictions are evaluated at the target resolution after resizing.

Action and format rewards. Let $\mathcal{B}$ denote the valid boxes defined above. The indicators used by the action and verification rewards are

$$\begin{aligned} \delta_s &= \mathbf{1}[a = \textsc{Segment} \ \wedge \ \widehat{B}^t \in \mathcal{B}], \\ \delta_d &= \mathbf{1}[a = \textsc{Defer} \ \wedge \ \widehat{B}^t \text{ is empty}]. \end{aligned} \tag{5}$$

The target-box score u_t is evaluated only when $y\delta_s = 1$ and is zero otherwise; $u_v = 0$ for an invalid verification box. Verification boxes are scored against the withheld annotation, never supplied as part of the response-generation input. The format indicator $F(o) = 1$ requires the six fields in the stated order, with a visual rule, target evidence, concept phrase, valid verification box, and an action-compatible target box; otherwise $F(o) = 0$. The format reward checks structure only. Mismatch-type labels do not enter the reward.

Generated answer. In Stage II and inference, SAM 3 receives the generated concept phrase in `<answer>` as its text condition, without a ground-truth or query-derived replacement. Feedback on this phrase comes through the resulting mask and sequence reward.

C.2 Reward and Training Details

Soft overlap. For foreground probabilities P and a binary confounder mask C,

$$\mathrm{SoftIoU}(P, C) = \frac{\sum_{j=1}^{HW} P_j C_j}{\sum_{j=1}^{HW} (P_j + C_j - P_j C_j) + \varepsilon_{\mathrm{IoU}}}, \qquad \varepsilon_{\mathrm{IoU}} = 10^{-6}. \tag{6}$$

Only available confounder annotations enter $\mathcal{C}$; if none are available, $d = 0$. Binary-mask IoU uses the same denominator stabilizer. The logits are bilinearly resized to the target grid before mask rewards or dense losses are computed.

Stage-specific supervision. Stage I uses a teacher-forced shell with empty rule, check, evidence, and action fields followed by the ground-truth target box. It contains no answer field, and the SAM 3 text prompt is `visual`. Only CPA is optimized, using soft Dice and binary cross-entropy. Stage II uses generated responses and sigmoid focal loss with $\alpha = 0.25$ and $\gamma = 2$, adding Dice only for supported targets. The dense objective follows ground-truth polarity regardless of the generated action.

For Stage I, the fixed text condition gives $Z^{(1)} = \mathcal{S}_\psi(I; \texttt{visual}, E_\phi(o^{(1)}; x))$ and $P^{(1)} = \sigma(Z^{(1)})$. The alignment objective is

$$\mathcal{L}_{\mathrm{align}}(\phi) = \mathbb{E}_{(x,G)\sim\mathcal{D}_1} \left[\mathcal{L}_{\mathrm{Dice}}(P^{(1)}, G) + \mathcal{L}_{\mathrm{BCE}}(Z^{(1)}, G) \right]. \tag{7}$$

Soft Dice loss is one minus twice the prediction–target overlap divided by their total mass, with 1 added to both numerator and denominator. Binary cross-entropy and sigmoid focal loss are averaged over pixels. There are no Stage I language, policy, or box-regression losses.

Training configuration. Stage I uses supported queries only. For the Qwen3.5-4B instantiation, Stage II starts from the Stage I checkpoint at update 1,220. We run 100 optimizer updates, each using eight supported and eight unsupported queries with eight completions per query, for 12,800 sampled trajectories in total. Among the 800 unsupported training queries, the sampler assigns 320, 240, and 160 to Levels 1, 2, and 3, respectively, and 80 to a pooled Level 4/Level 5 stratum. The latter contains 43 Level 4 and 37 Level 5 queries; OOD-only subtypes are excluded from training. We use AdamW with a learning rate of 2×10^{-6}, a 10-update warm-up followed by a constant schedule, temperature 0.9, and a maximum completion length of 2,048 tokens. Stage II trains the MLLM and CPA with frozen SAM 3; a frozen copy of the language policy supplies the KL reference. Queries are sampled with balanced support labels and matched requested concepts, without requiring both members of an image pair in the same update. GRPO centers and scales rewards within each query's sampled-response group, not across different supported and unsupported queries. The policy and CPA losses have equal weight.

Joint objective. Direct mask supervision follows the support label rather than the generated action. For logits Z and $P = \sigma(Z)$,

$$\ell^{+}_{\mathrm{CPA}} = \mathcal{L}_{\mathrm{Dice}}(P, G) + \mathcal{L}_{\mathrm{Focal}}(Z, G),\ \ell^{-}_{\mathrm{CPA}} = \mathcal{L}_{\mathrm{Focal}}(Z, \mathbf{0}), \tag{8}$$

where $\mathcal{L}_{\mathrm{Focal}}$ is sigmoid focal loss. With $\mathbb{E}_{+}$ and $\mathbb{E}_{-}$ denoting expectations over responses to supported and unsupported queries, the joint objective is

$$\mathcal{L}_{\mathrm{CPA}}(\phi) = \tfrac{1}{2}\big(\mathbb{E}_{+}[\ell^{+}_{\mathrm{CPA}}] + \mathbb{E}_{-}[\ell^{-}_{\mathrm{CPA}}]\big),\ \mathcal{L}(\theta, \phi) = \mathcal{L}_{\mathrm{GRPO}}(\theta) + \mathcal{L}_{\mathrm{CPA}}(\phi). \tag{9}$$

GRPO (Shao et al., 2024) uses $A_i = (R_i - \bar{R})/(s_R + \epsilon)$ as the within-query advantage, where $\bar{R}$ and s_R are the response-group reward mean and standard deviation and $\epsilon > 0$ stabilizes normalization; $\mathcal{L}_{\mathrm{GRPO}}$ is its clipped policy objective with KL regularization to a frozen reference policy. The sampled responses and their rewards are held fixed when constructing the policy surrogate.

D QUALITATIVE EXAMPLES

Figures 11–20 show 19 cases of MedCounterSeg-RL (Qwen3.5-4B), with one case below and two per subsequent page. Each case occupies one panel with the same layout. The header gives the decision type, source dataset, imaging modality, and split. The left column shows the query, the three boxed reference images, the verification image with the predicted box, and the target image with the ground-truth mask and the predicted mask overlaid. The right column shows the model output: the inferred visual rule, the target-specific evidence, and the decision basis. The footer states the predicted and expected actions. The cases include supported queries that were segmented, counterfactual queries that were correctly deferred, and failure cases. Clinicians reviewed the evidence descriptions displayed in these cases, as described in Appendix A.5.

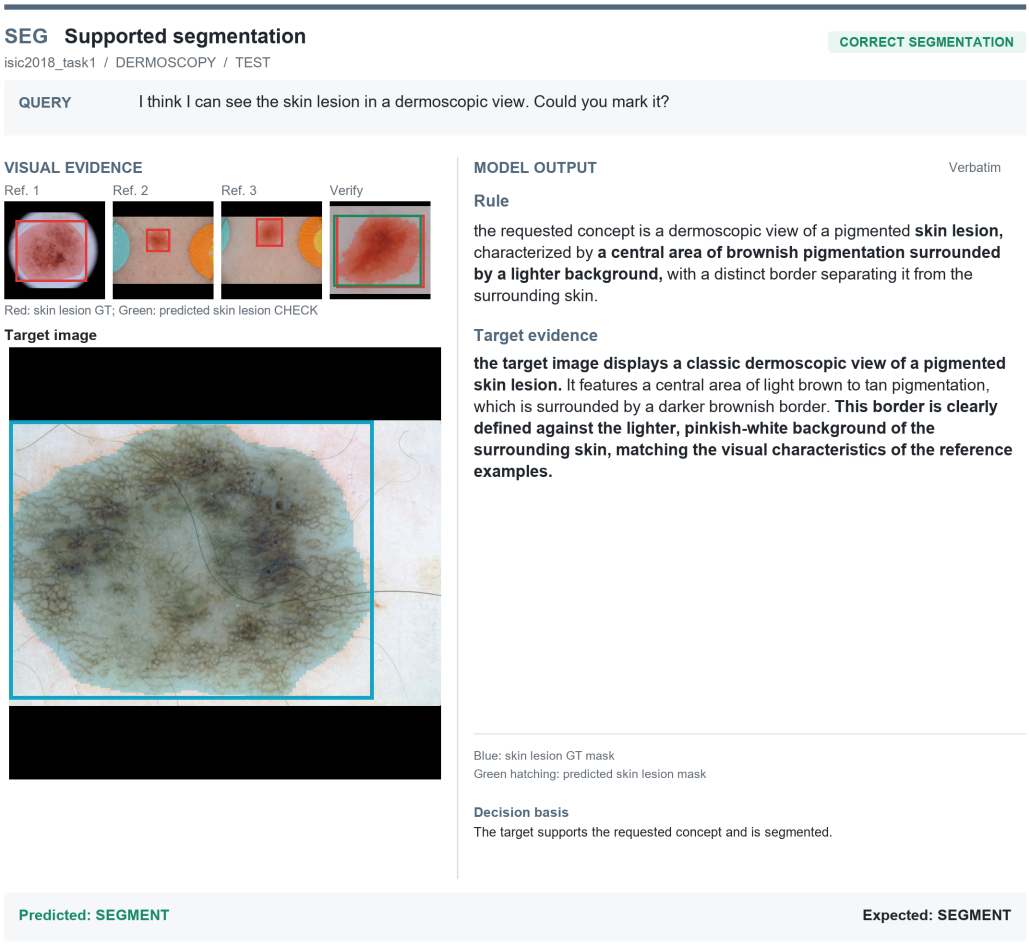

Figure 11: Qualitative example, case 1 of 19. Panel layout as described in Appendix D.

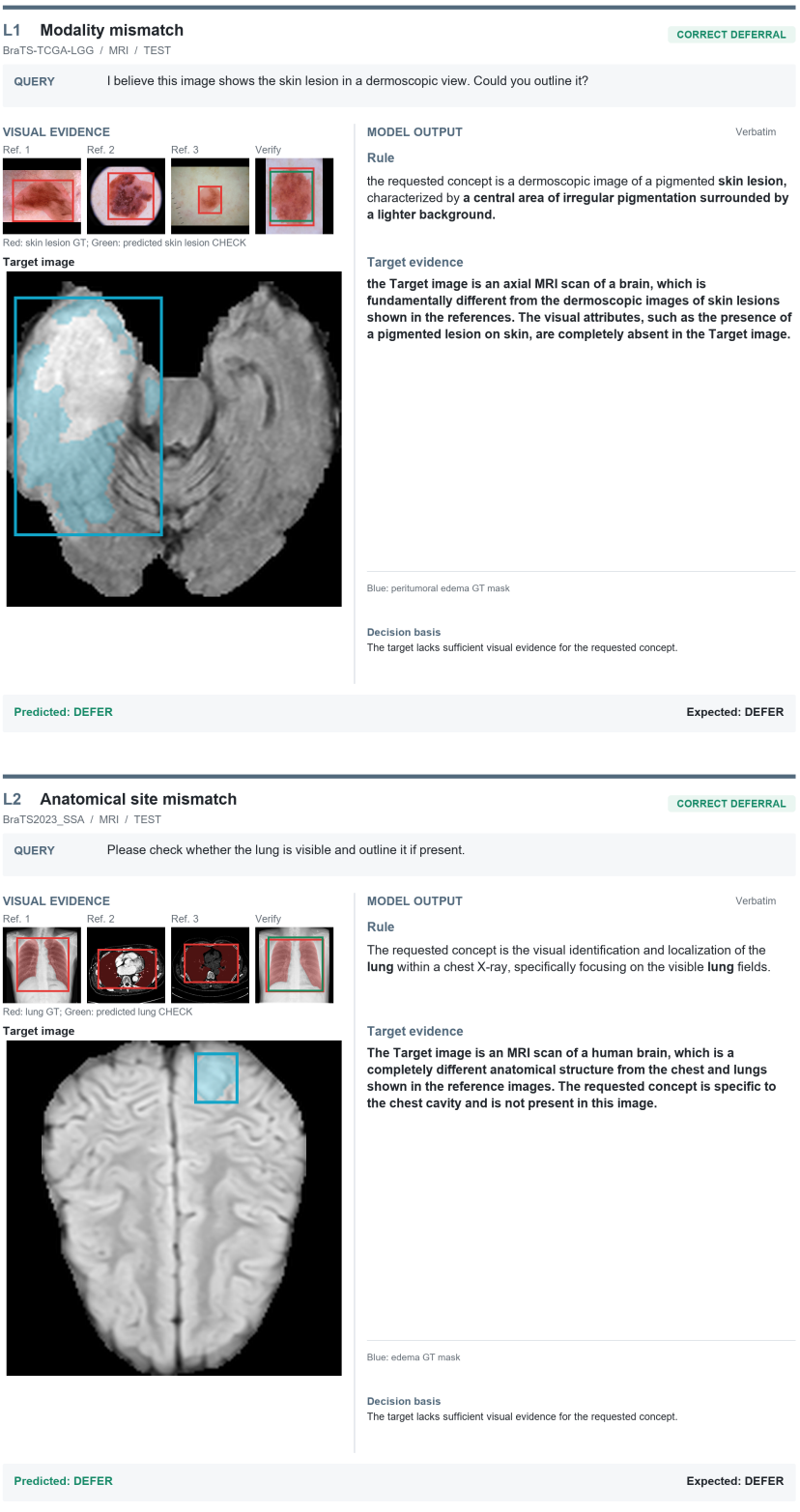

Figure 12: Qualitative examples, cases 2 and 3 of 19. Panel layout as described in Appendix D.

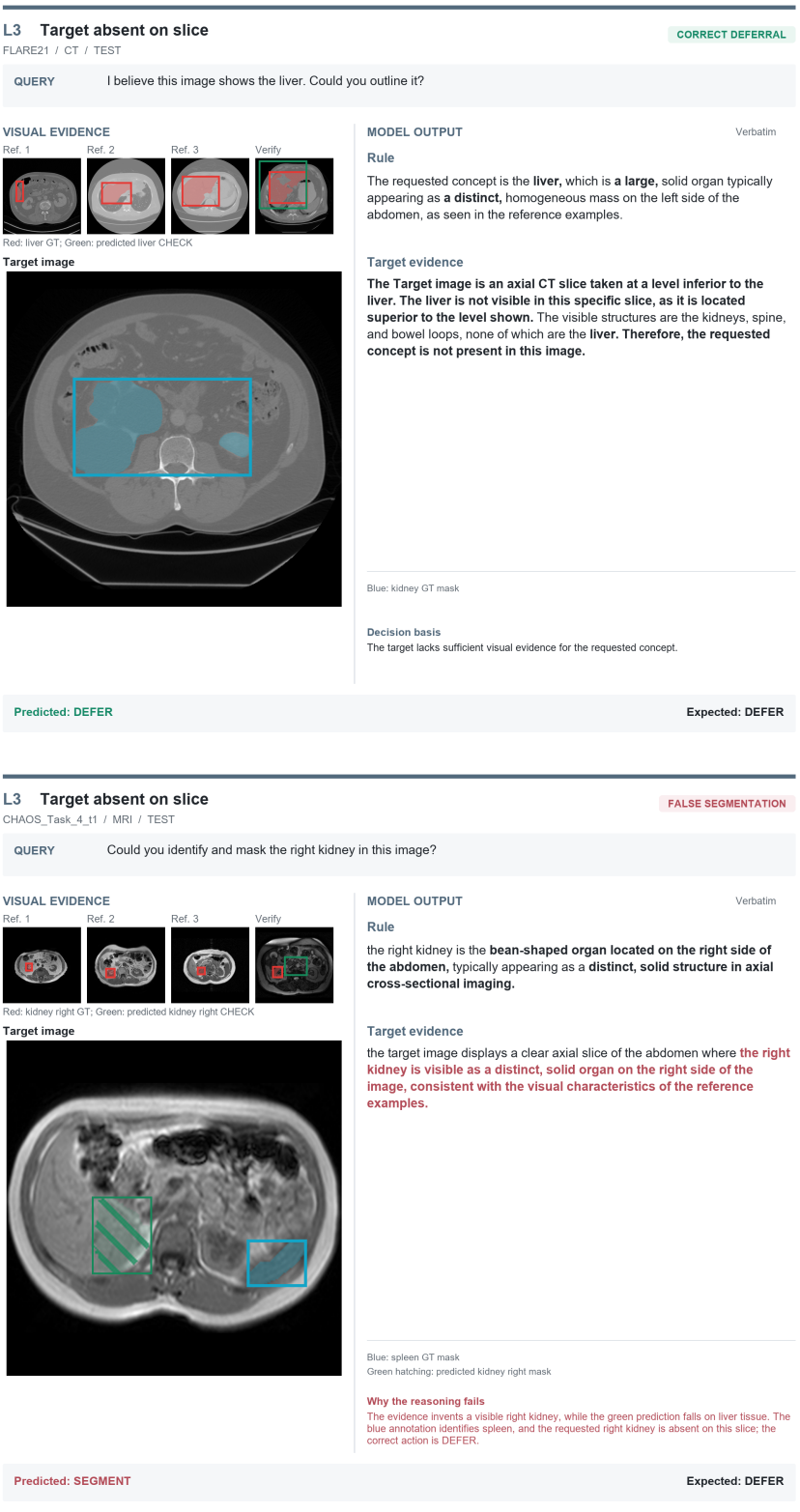

Figure 13: Qualitative examples, cases 4 and 5 of 19. Panel layout as described in Appendix D.

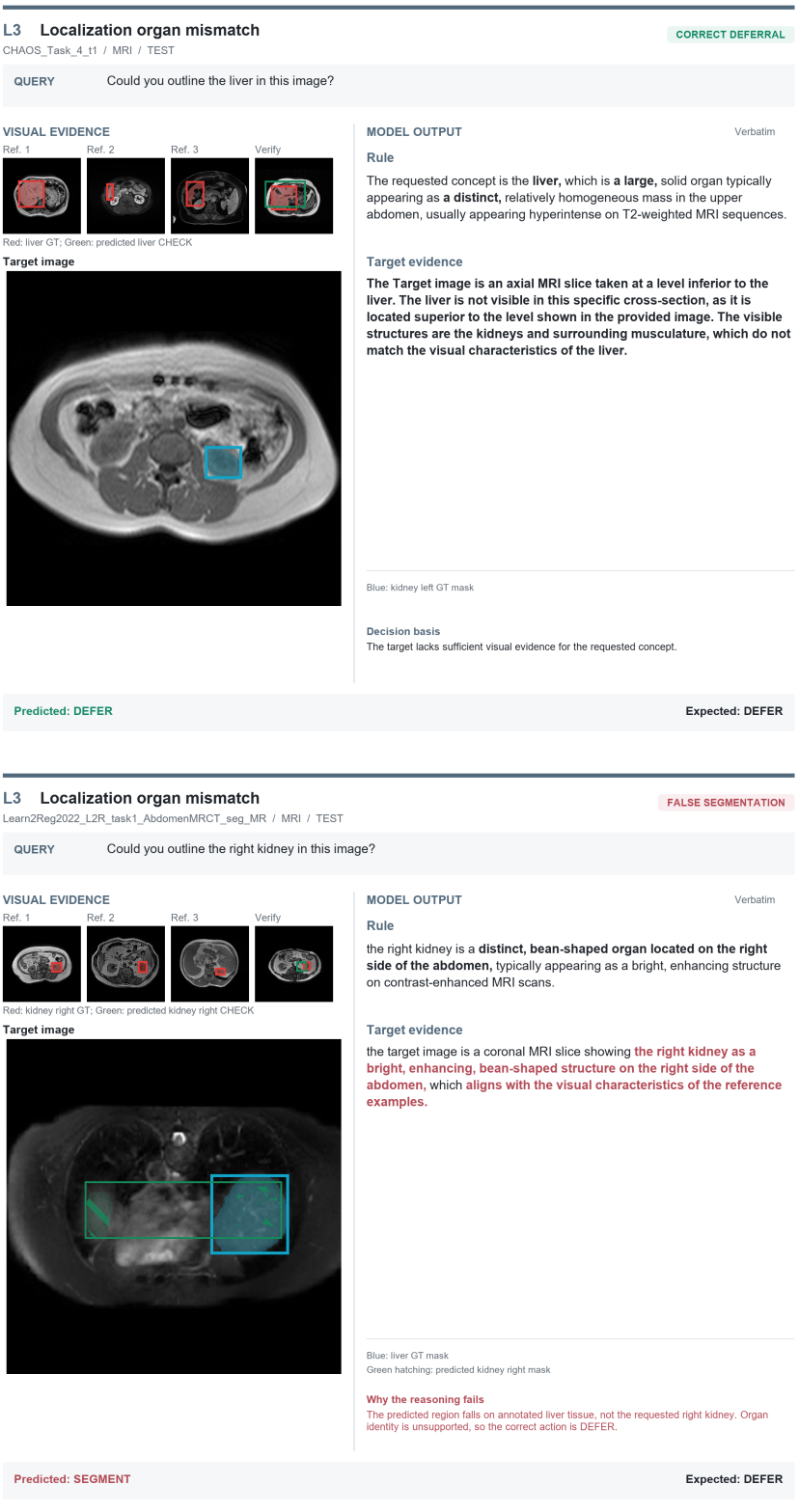

Figure 14: Qualitative examples, cases 6 and 7 of 19. Panel layout as described in Appendix D.

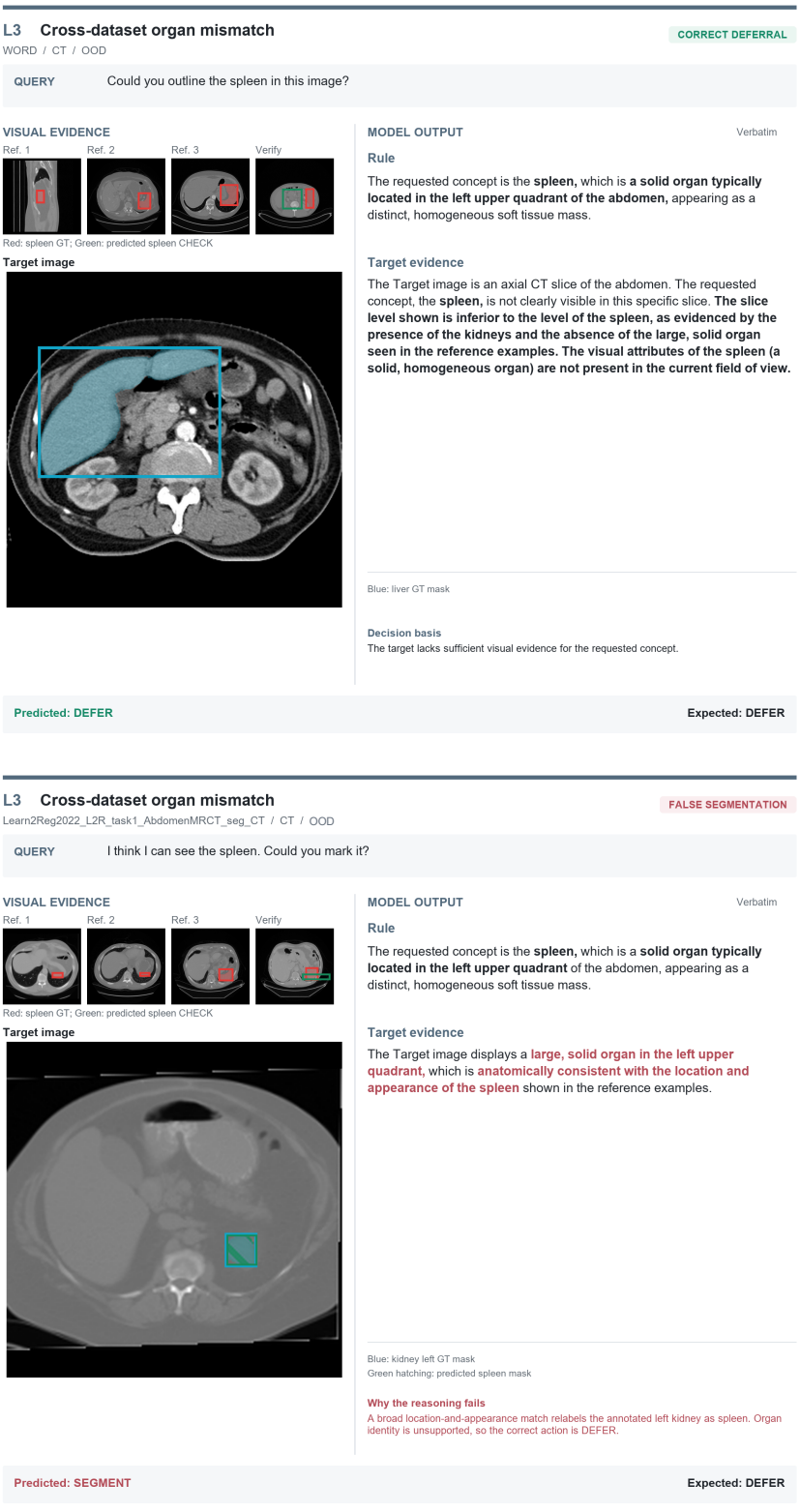

Figure 15: Qualitative examples, cases 8 and 9 of 19. Panel layout as described in Appendix D.

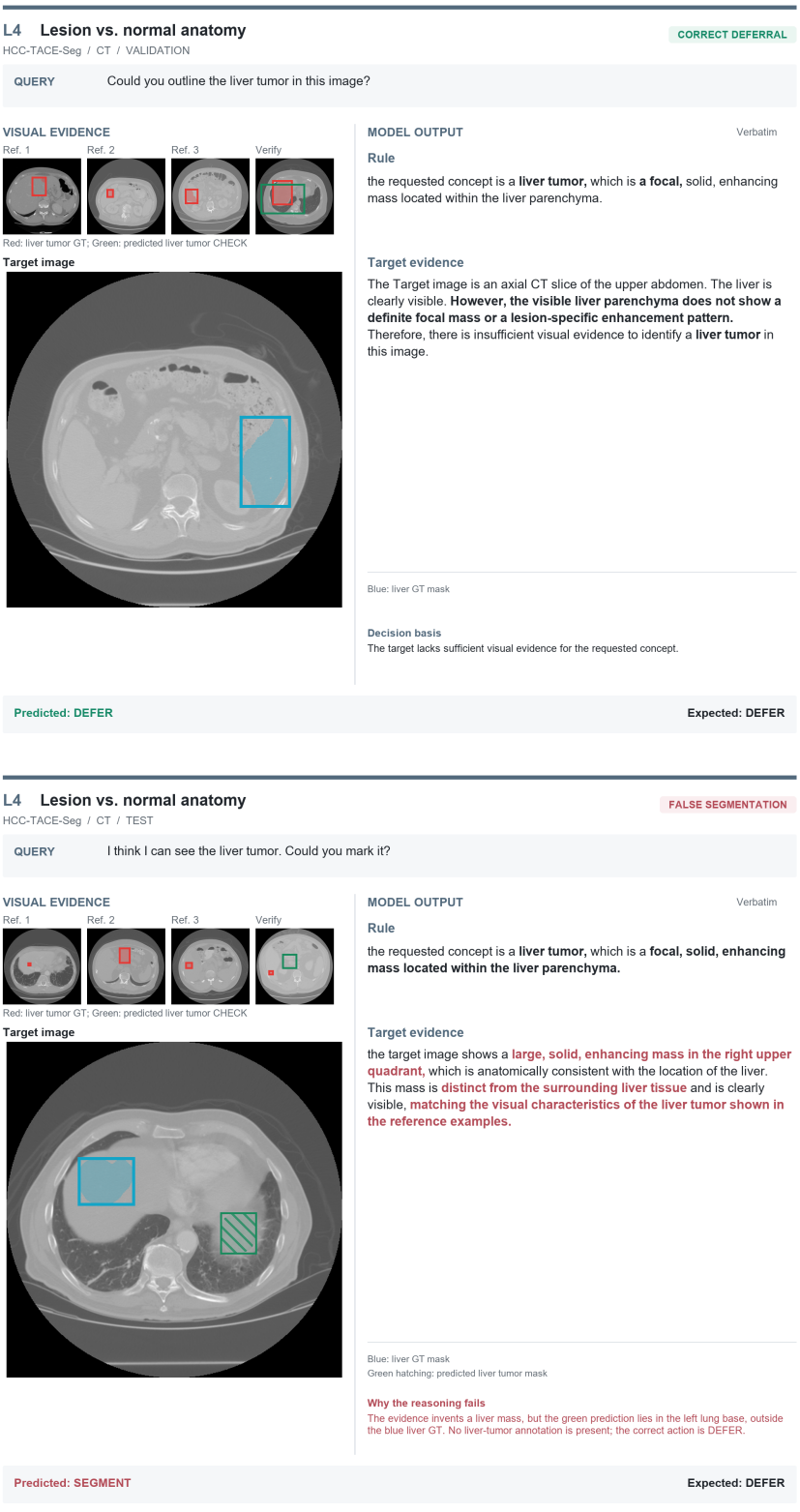

Figure 16: Qualitative examples, cases 10 and 11 of 19. Panel layout as described in Appendix D.

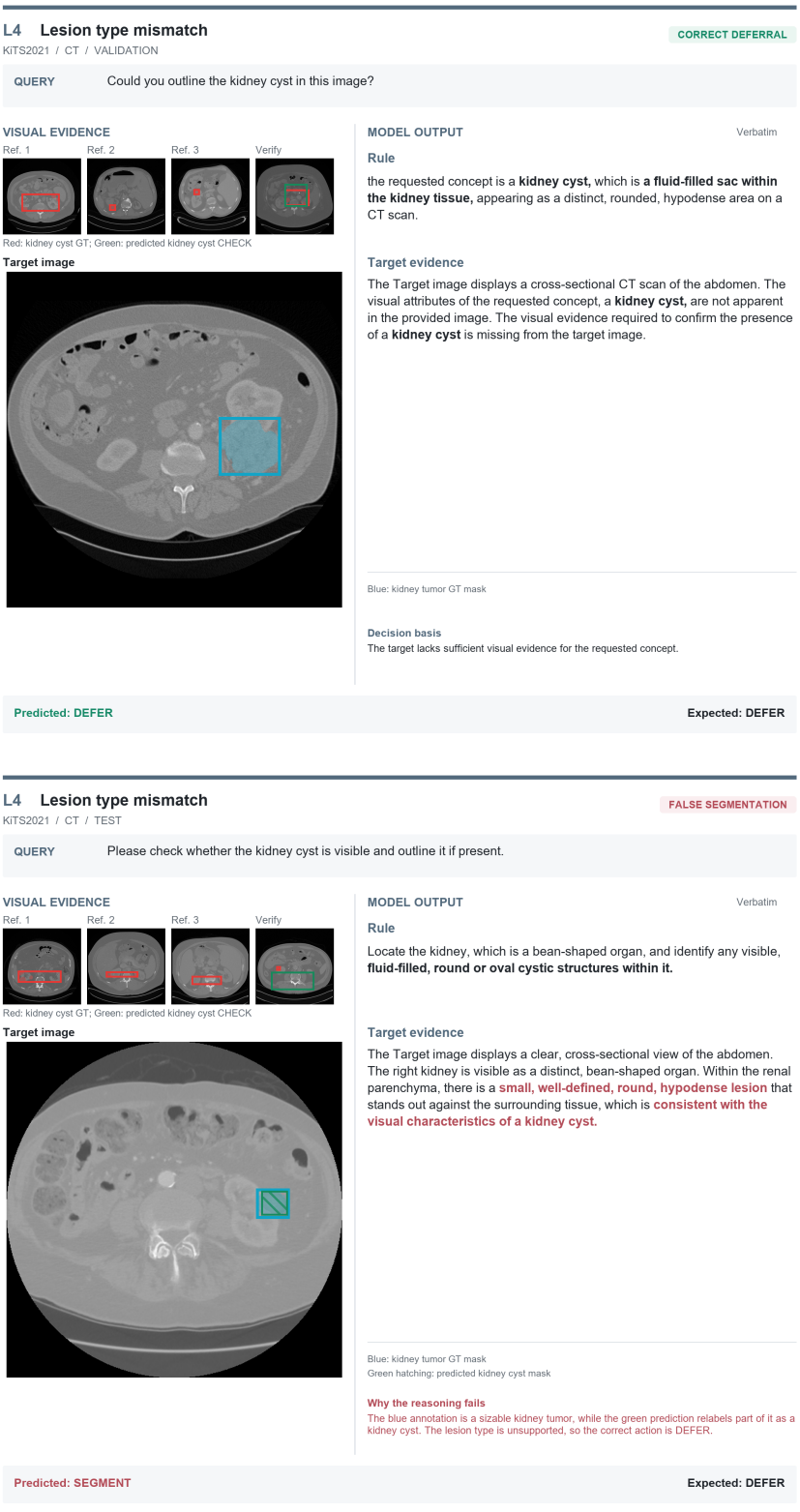

Figure 17: Qualitative examples, cases 12 and 13 of 19. Panel layout as described in Appendix D.

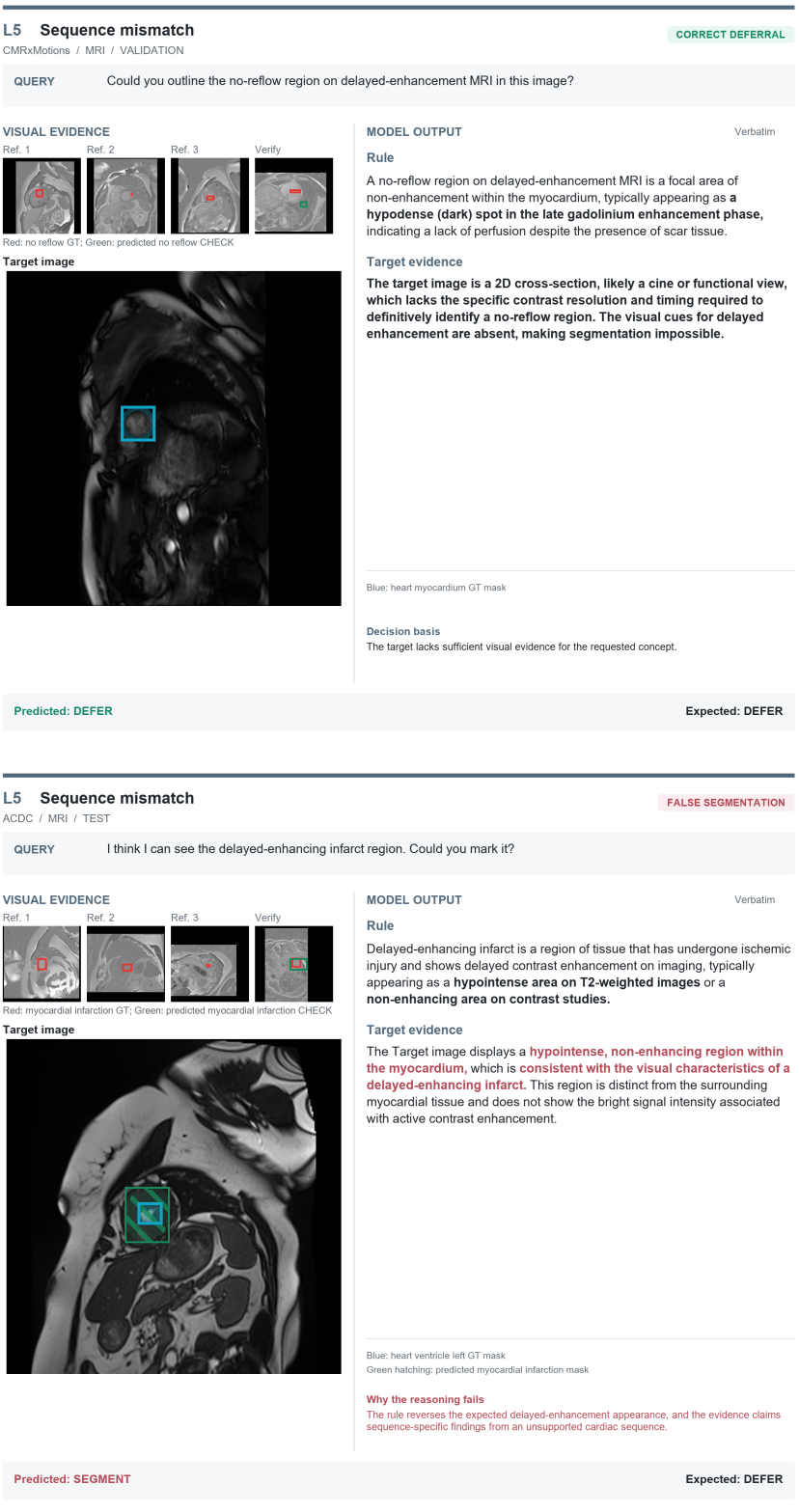

Figure 18: Qualitative examples, cases 14 and 15 of 19. Panel layout as described in Appendix D.

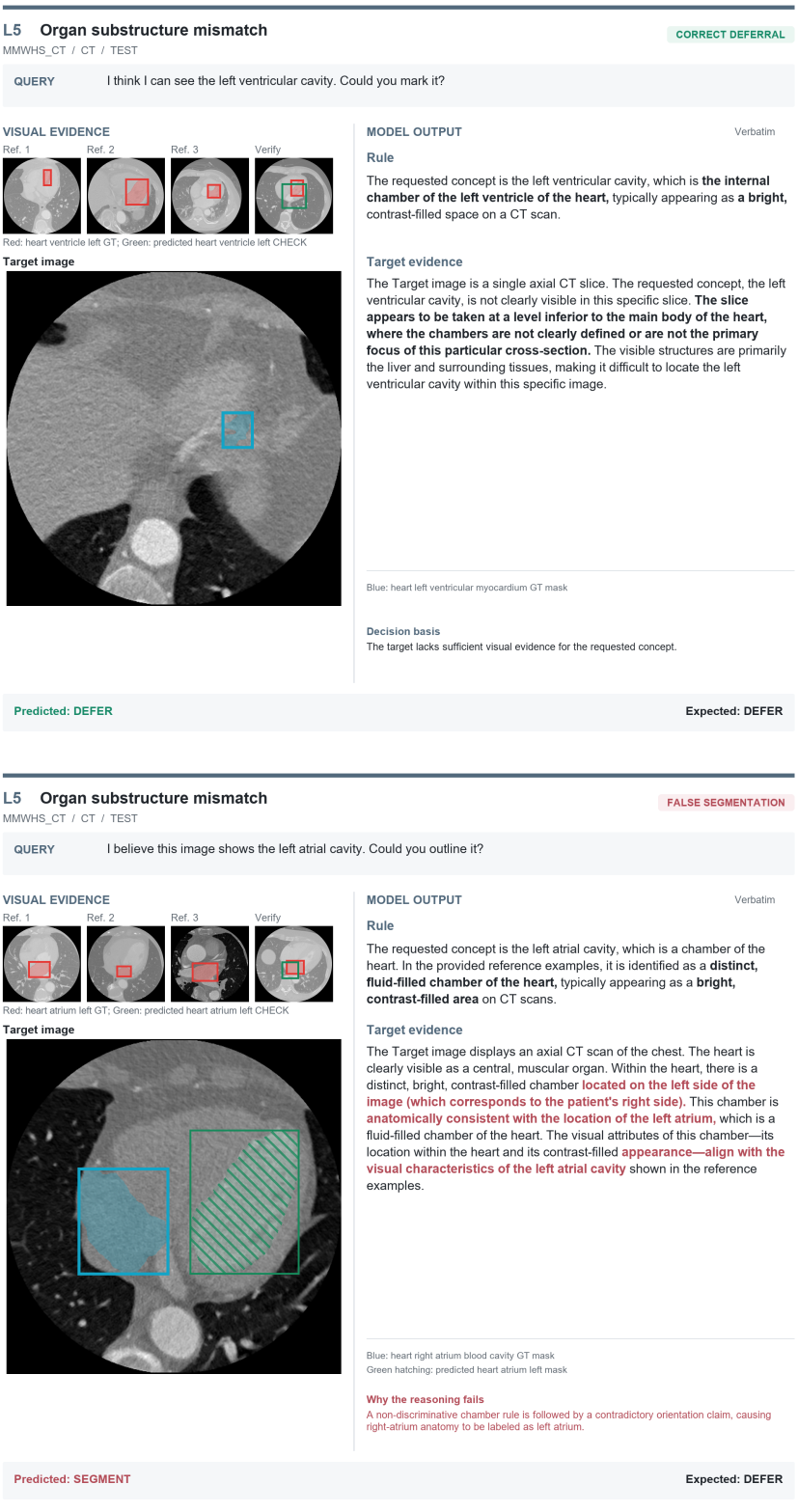

Figure 19: Qualitative examples, cases 16 and 17 of 19. Panel layout as described in Appendix D.

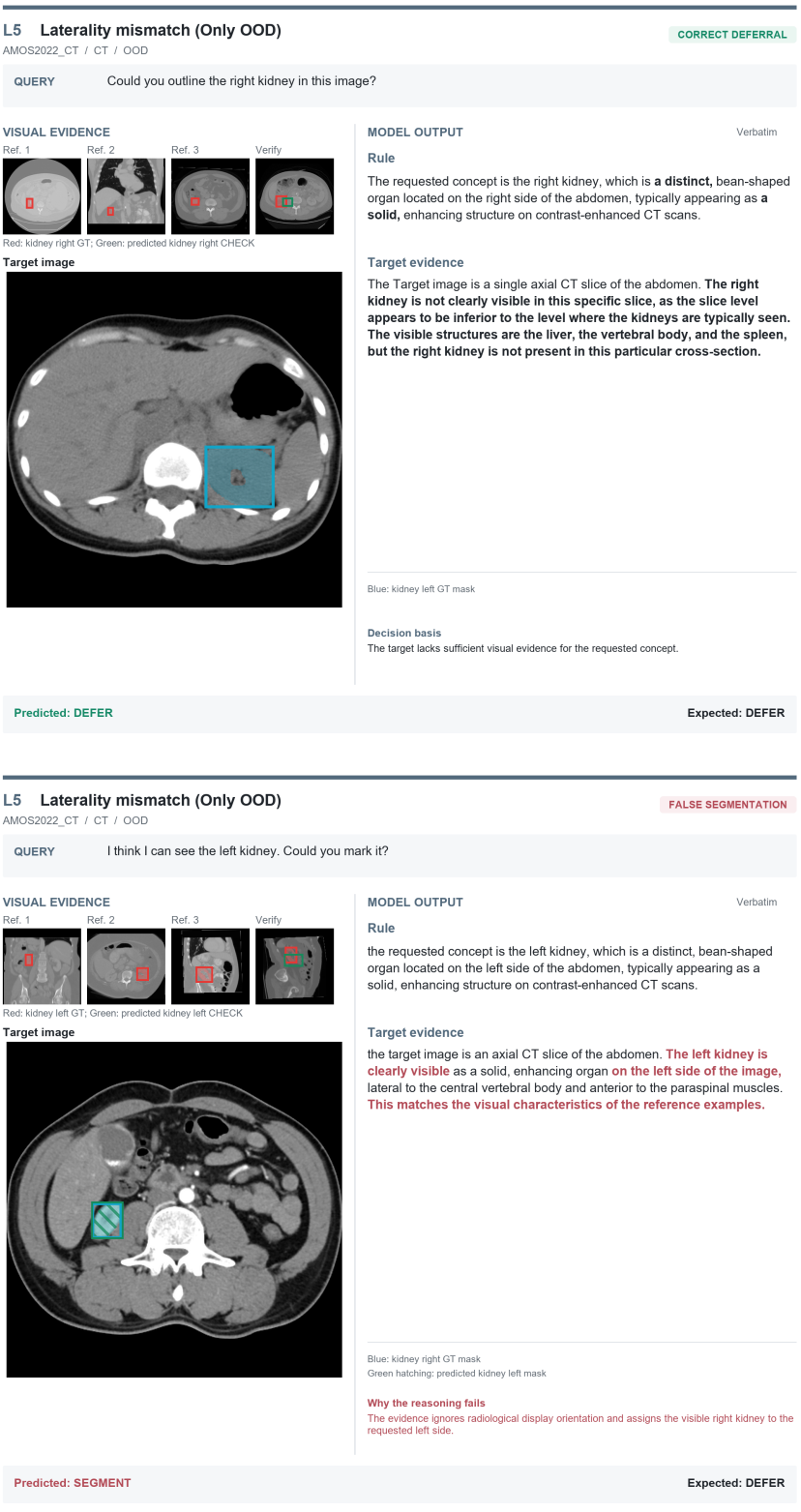

Figure 20: Qualitative examples, cases 18 and 19 of 19. Panel layout as described in Appendix D.