XFLOW: A WORKFLOW MODEL FOR INSTRUCTION-GUIDED LESION SEGMENTATION IN CHEST X-RAYS

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing text-guided segmentation models in the medical domain cover only a narrow set of anatomical structures and lesions in chest X-rays (CXRs), and most of them assume that the queried target is always present in the image. Instruction-guided lesion segmentation (ILS) was introduced to overcome these limitations by segmenting diverse lesion types from simple user instructions while also recognizing when the queried lesion is absent, and ROSALIA was proposed as the first model for this task. However, the masks produced by ROSALIA remain of limited quality, often carrying scattered noise. Moreover, ROSALIA predicts the mask in a single shot, which differs fundamentally from how radiologists perceive and delineate lesions in practice. A radiologist first surveys the entire thorax, then localizes the approximate region of abnormality, and only then refines the lesion contour. Motivated by this coarse-to-fine, multi-level perception process, we present XFlow, a workflow model for ILS that combines box-based localization with multi-turn point refinement. XFlow detects the lungs, decides whether the queried finding is present in each of them, and prompts a fine-tuned SAM with the lesion box for an initial mask. It then corrects that mask through point prompts until its boundary follows the lesion, leaving every intermediate decision visible. Our experiments show that XFlow achieves the best segmentation quality on both internal and external evaluation. Notably, it surpasses ROSALIA in segmentation quality even when the two are trained on the same lesion annotations.

1 INTRODUCTION

Medical imaging plays a crucial role in modern medicine as a primary source of visual evidence. Among the available modalities, chest X-ray (CXR) is the most widely used examination for rapidly assessing a patient's overall condition (Broder, 2011). Compared with other modalities such as CT or MRI, however, CXR offers limited spatial resolution and projects three-dimensional anatomy onto a single plane, so that multiple structures overlap within the same region. Accurate interpretation therefore demands substantial radiological expertise. In routine clinical practice, however, radiologists must read a large volume of studies each day and are often expected to interpret an individual case within minutes. Among the required reading tasks, lesion segmentation requires pixel-level judgment rather than a mere image-level impression, and this step consumes the largest share of a radiologist's reading time.

A text-guided segmentation model could ease this burden, yet existing medical models cover only a limited range of anatomical structures and lesion types in CXRs (Huang et al., 2024; Li et al., 2023), since collecting large-scale, high-quality annotations from radiologists across diverse findings is time-intensive. Moreover, most assume the queried target is present in the image, whereas radiologists must first determine whether a lesion exists at all. To address these limitations, Choi et al. (2026c) introduced instruction-guided lesion segmentation (ILS), which segments diverse CXR lesions from simple instructions while also recognizing when the queried lesion is absent. Along with the task, they released MIMIC-ILS, a large-scale lesion segmentation dataset (Choi et al., 2026b), and proposed ROSALIA, the first vision-language model (VLM) for ILS. That work, however, centers on the dataset, and ROSALIA's masks remain of limited quality, often carrying scattered noise.

We attribute these failures to ROSALIA's architecture: it adopts LISA (Lai et al., 2024), which couples a VLM with the Segment Anything Model (SAM) (Kirillov et al., 2023) for referring image

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.

segmentation (RIS) in the general domain, and fine-tunes it on MIMIC-ILS without further modification. Like ROSALIA, most text-guided segmentation models in the medical domain couple a VLM with a segmentation module and predict the mask in a single shot (Zhao et al., 2024; Huang et al., 2025). Such a design offers the user no view of how the mask was reached, and it is overly simplistic for lesion segmentation, which demands clinical expertise.

Recent work such as IBISAgent (Jiang et al., 2026) moves beyond a single shot by having a VLM place point prompts on SAM over multiple turns, yet it relies on points alone, carving the mask at a fine-grained level from the start and thus requiring many steps to complete it. Radiologists in fact proceed through several levels of perception in sequence to delineate a lesion. They first survey the whole thorax to form a global impression, then narrow their attention to the approximate region where an abnormality is likely to reside, and only at the last stage refine the exact boundary at a fine-grained level. Single-shot models collapse these levels into one step, and point-only refinement covers only the last.

Following this observation, we develop XFlow, a workflow model for ILS in CXRs. XFlow performs segmentation as a sequence of two stages that separate coarse-level from fine-level perception (Figure 1). In the initial segmentation stage, the model reads the cardiopulmonary field and first detects the lungs, then examines each of them in turn to decide whether the queried finding is there, and draws a box around the lesion, which it hands to a fine-tuned SAM as a prompt to obtain the first mask. The subsequent segmentation revision stage inspects that mask and corrects it through point prompts to the SAM, step by step, until the boundary follows the lesion. Every intermediate decision is written out explicitly, so the user can trace how the final mask was reached.

In our experiments, XFlow outperforms all compared models in segmentation quality on both internal and external evaluation, and an ablation study shows that each stage of the workflow, together with its training scheme, contributes to the gain. Trained on the same MIMIC-ILS cases as ROSALIA, XFlow improves segmentation by a clear margin, indicating that the gain does not come from additional lesion annotations.

Our contributions are summarized as follows:

- We present XFlow, a workflow model for CXR lesion segmentation, which decomposes the task into two stages following the order in which radiologists read an image.
- Trained on the same lesion annotations as ROSALIA, XFlow outperforms it by a clear margin on both internal and external evaluation, and an ablation shows that each component of the decomposition contributes to the gain.
- In a blinded reader study with three medical experts, masks from XFlow are ranked close to the ground truth and are preferred over those of ROSALIA about twice as often.

2 Related Works

Segmentation with VLMs and SAM. To segment a target described in free-form text, general-domain work couples a VLM with SAM, and the resulting models differ mainly in how the VLM communicates with the segmenter. LISA (Lai et al., 2024) feeds the hidden state of a special `[SEG]` token to the SAM decoder, whereas later models have the VLM emit geometric prompts that SAM consumes directly. Seg-Zero (Liu et al., 2025) predicts a box and points together and prompts SAM with them in a single step, while SegAgent (Zhu et al., 2025) relies on point clicks alone, placing them over multiple turns in imitation of a human annotator. In the medical domain, IBISAgent (Jiang et al., 2026) follows the latter design, with a VLM issuing point prompts across turns.

Both designs, however, have limitations. A single-step prompt cannot correct the mask once it is produced, while a point-only procedure forgoes the box and must carve out the mask click by click, taking IBISAgent 8.69 steps on average. Moreover, IBISAgent is supervised on individual revision steps, and although it later applies RL to its point prompts, it still optimizes each turn in isolation, so the policy never learns how its clicks should work together across the trajectory. XFlow combines the two designs, first localizing the lesion with a box and then refining its boundary through successive point prompts within at most four steps, and trains the revision policy with a reward defined over the entire multi-turn trajectory.

Instruction-Guided Lesion Segmentation in Chest X-rays. Earlier models for text-guided lesion segmentation in CXRs are confined to a single target, largely because annotated data is

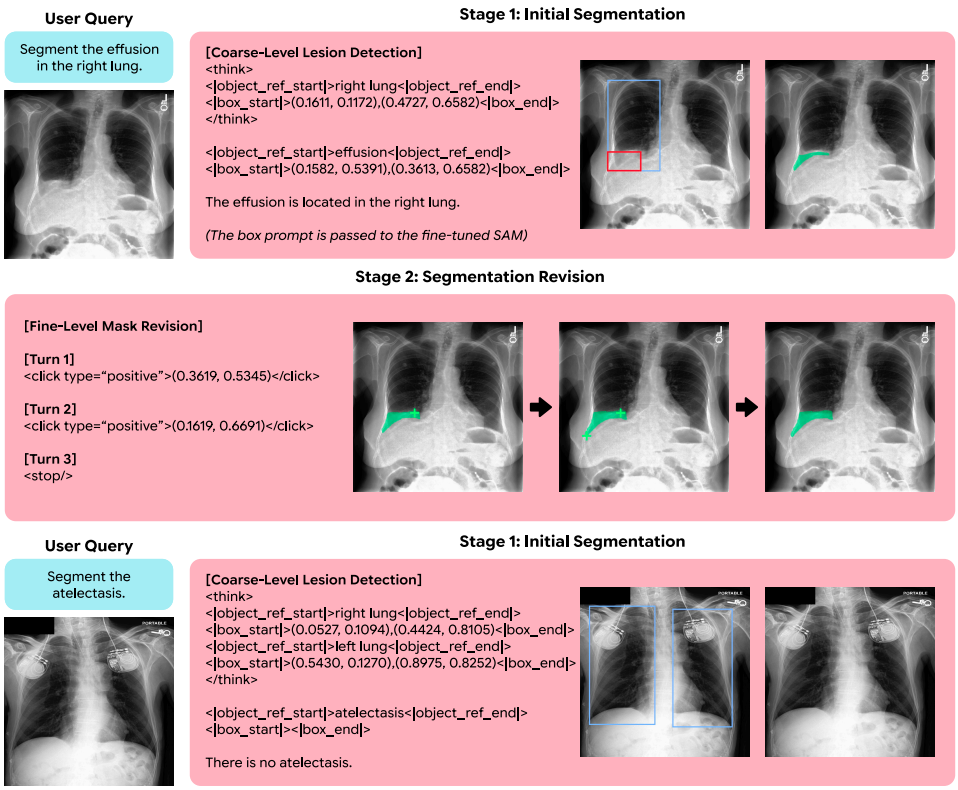


Figure 1: Overview of XFlow. (Top) In Stage 1, XFlow grounds the lung regions, localizes the finding as a bounding box, and prompts a fine-tuned SAM with that box to obtain an initial mask. In Stage 2, it inspects the mask and issues positive or negative point prompts that expand or shrink the predicted region, repeating this until the mask is satisfactory. (Bottom) If the queried finding is absent, XFlow terminates at Stage 1 without producing a mask.

scarce (Huang et al., 2024; Li et al., 2023). CheXlocalize (Saporta et al., 2022) provides radiologist-drawn masks for CXR findings, but it is far too small to train on. To overcome this, MIMIC-ILS introduced a large-scale lesion segmentation dataset built through an automated framework, together with ROSALIA, a LISA-based model trained on it (Choi et al., 2026c;b). ROSALIA accepts a text instruction such as "Segment the opacity" and covers seven clinically important findings, namely cardiomegaly, pneumonia, atelectasis, opacity, consolidation, edema, and effusion. Its masks, however, are often imprecise, and because it predicts them in one shot, it neither follows the coarse-to-fine workflow of a radiologist nor shows the user how the final delineation was reached. XFlow addresses both: it covers the same findings through this workflow, and because every intermediate decision is written out in text, each of them remains inspectable by the user.

3 Task Formulation

We address instruction-guided lesion segmentation (ILS), the task introduced in MIMIC-ILS. An instruction $Q = (c, \ell)$ names a target lesion c and a location ℓ at which to look, where ℓ ranges from the whole thorax ("Segment the pleural effusion.") through one lung ("…in the left lung.") to a specific zone ("…in the left lung base."). Given a chest X-ray I and such an instruction, a model f returns a binary mask over the image,

$$M^{\text{pred}} = f(I, Q), \qquad M^{\text{pred}} \in \{0,1\}^{H \times W}, \tag{1}$$

covering the region the instruction refers to.

ILS is a medical counterpart of referring image segmentation (RIS), with one difference that follows from how the queries arise. In RIS, the referring expression describes an object the image is known to contain, so the target mask is nonempty by construction, whereas a radiologist is as often asked

whether a finding is there at all. When c is absent from ℓ, the ground truth is the empty mask, $M^{\mathrm{GT}} = \mathbf{0}$. Whatever the instruction specifies, then, the model must arrive at its own verdict on the region in question rather than take the instruction as evidence that something is there.

4 XFLOW

4.1 ARCHITECTURE OVERVIEW

Notation. XFlow and a single promptable segmenter F_{seg}, built on SAM and shared by both. The two policies interact with F_{seg} through its geometric prompts (boxes and points) rather than predicting the mask directly, and are applied in sequence.

Every region the model refers to is represented as a grounded box $g = (e, b)$, where e is a referring expression naming the region (e.g., right lung, pneumonia) and $b = (x_1, y_1, x_2, y_2) \in [0, 1]^4$ is its bounding box in normalized image coordinates. We denote by $\mathcal{L}$ and $\mathcal{B}$ the sets of grounded boxes for the lungs and for the lesion, respectively.

Stage 1: Initial segmentation. The first policy π_{init} follows the order in which a radiologist reads the image. It first grounds the lung regions to establish the anatomical context in which the lesion should be sought, then judges whether the queried lesion is present, and only then localizes it. These three decisions are emitted as a single sequence,

$$(\mathcal{L}, \mathcal{B}) \sim \pi_{\text{init}}(\cdot \mid I, Q), \tag{2}$$

where the lung boxes $\mathcal{L}$ precede the lesion boxes $\mathcal{B}$. The scope of $\mathcal{L}$ follows the instruction. A query about the thorax as a whole grounds both lungs, $|\mathcal{L}| = 2$, whereas a query restricted to one side grounds only that lung, $|\mathcal{L}| = 1$. When the lesion is absent, the policy emits no lesion box, $\mathcal{B} = \emptyset$, and the instruction is answered with an empty mask. Otherwise each lesion box prompts the segmenter separately, yielding a set of initial masks

$$\mathcal{M}_0 = \{\, F_{\text{seg}}(I, b) \mid b \in \mathcal{B} \,\}. \tag{3}$$

Stage 2: Segmentation revision. Each mask in $\mathcal{M}_0$ is revised independently, and the union of the revised masks forms the final prediction. Below we describe the procedure for one such mask, denoted M_0. The second policy π_{rev} refines M_0 over at most $T = 4$ steps. At step t, it observes the image overlaid with the current mask and the point prompts issued so far, denoted o_t, together with the history of preceding steps $P_{<t}$, and produces an action a_t,

$$a_t \sim \pi_{\text{rev}}(\cdot \mid I, Q, o_t, P_{<t}). \tag{4}$$

The action space consists of point prompting and termination. A point prompt is a pair (s_t, p_t) with polarity $s_t \in \{+1, -1\}$ and relative coordinate $p_t \in [0, 1]^2$, where a positive point expands and a negative point shrinks the predicted region. Since the box prompt confines the region the segmenter can cover, we enlarge it to the smallest box enclosing both b_{t-1} and p_t whenever a positive point falls outside it, and leave it unchanged otherwise, with b_0 set to the box that produced M_0. Executing the prompt yields the next mask

$$M_t = F_{\text{seg}}(I, b_t, \{(s_i, p_i)\}_{i \le t}, M_{t-1}), \tag{5}$$

which is rendered as the observation o_{t+1} and appended to the trajectory. The loop terminates when the policy judges the mask to be sufficient or when $t = T$. Because the entire revision proceeds within a single context, every step is conditioned on the full history of the preceding ones. Applying this procedure to every mask in $\mathcal{M}_0$ and taking the union of the results gives the final prediction M^{pred}.

4.2 TRAINING DATA

Training XFlow involves three components. The segmenter F_{seg} is fine-tuned for CXR lesions, and each of the two policies is trained through supervised fine-tuning (SFT) followed by reinforcement learning (RL). All three are trained on data constructed automatically from MIMIC-ILS (Choi et al., 2026c;b), since every field is obtained from the masks and metadata the dataset already provides, so building it requires no additional human annotation. We divide MIMIC-ILS into disjoint splits for SFT and RL (D_{SFT} and D_{RL}), both of which are also disjoint from the evaluation split D_{eval}. For RL and evaluation we draw both sets from the cases that medical experts in the CheXpercept study (Choi et al., 2026a) identified as the most reliable within MIMIC-ILS in terms of mask quality.

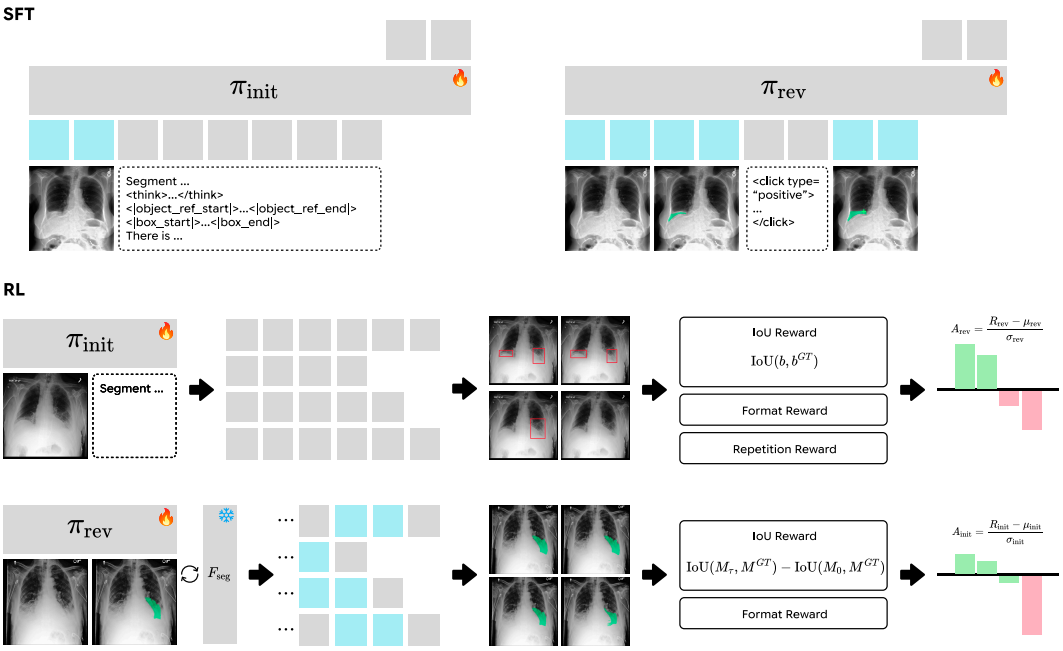


Figure 2: Training the two policies. We first apply SFT to π_{init} and π_{rev} on automatically constructed initial segmentation and segmentation revision scenarios, and then train each policy further with GRPO. A_{init} and A_{rev} are the advantages obtained by normalizing R_{init} and R_{rev} within each group of sampled generations. Text tokens are shown in gray and image tokens in light blue.

4.3 SAM Training

Within XFlow, F_{seg} is invoked by the policies, so the segmenter must be able to respond to prompts that point at lesions. Neither the original SAM nor its medical variants such as MedSAM are trained on CXR lesions, so we obtain F_{seg} by fine-tuning SAM3 (Carion et al., 2026) on the SFT split D_{SFT}. From each ground-truth mask M^{GT} we derive the prompts used during training, sampling a box b that encloses the annotated region together with points (s, p) drawn from inside and outside it, so that the prompt configuration seen during training matches the one the policies produce at inference time.

4.4 VLM Training

The two policies are trained separately, and executed in sequence at inference time. SFT alone optimizes only the token-level likelihood of the target sequence. The model learns to imitate the format of the target sequence, but receives no direct feedback on whether the decisions it contains are correct. We therefore continue with RL in both stages, using rewards that measure the quality of the outcome. Throughout, F_{seg} is kept frozen, so the policies learn to steer a fixed segmenter. Prompt templates and example SFT targets are deferred to Appendix A.3.

4.4.1 Initial Segmentation Policy

Each SFT example for π_{init} pairs an image and an instruction with the sequence of decisions the stage requires, the lung boxes, the presence decision, and the lesion boxes, written out in the format shown in Figure 1. Lung boxes are obtained by running an off-the-shelf lung segmentation model (Seibold et al., 2023; Cosarinsky et al., 2025), while the presence label and the lesion boxes are derived directly from MIMIC-ILS.

The RL reward combines a localization term with two auxiliary terms,

$$R_{\text{init}} = R_{\text{box}} + R_{\text{format}} + R_{\text{rep}}, \qquad R_{\text{box}} = \text{IoU}(b, b^{\text{GT}}). \tag{6}$$

For a negative case, R_{box} is instead binary and depends only on whether the model emits $\mathcal{B} = \emptyset$. The other two terms act on the form of the output rather than its content. R_{format} checks that the generated text follows the expected structure, so that the grounded boxes can be parsed from it, and R_{rep} penalizes emitting the same box more than once.

4.4.2 Segmentation Revision Policy

SFT examples for π_{rev} are revision trajectories constructed by simulating the process the policy will later perform. We perturb the ground-truth box b^{GT} and feed it to F_{seg}, which yields an imperfect mask $\tilde{M}$. Starting from $\tilde{M}$, we place clicks following the simulated-user protocol standard in interactive segmentation (Xu et al., 2016; Sofiiuk et al., 2022; Kirillov et al., 2023), which selects at each step the point that brings the resulting mask closest to M^{GT}. One trajectory is thus

$$(M_0, a_1, M_1, a_2, \ldots), \qquad M_0 = \tilde{M}, \tag{7}$$

where a point prompt (s_t, p_t) turns M_{t-1} into M_t, and the trajectory ends either at the step limit or when the policy issues the termination action. Each M_t enters the example as the observation o_{t+1}, rendered by overlaying the mask and the points issued so far on I, exactly as at inference time.

The RL reward pairs the gain in IoU over the trajectory the model itself produces with the same format term as before,

$$R_{\text{rev}} = R_{\text{gain}} + R_{\text{format}}, \qquad R_{\text{gain}} = \text{IoU}(M_\tau, M^{\text{GT}}) - \text{IoU}(M_0, M^{\text{GT}}), \tag{8}$$

where τ is the step at which the policy terminates and $M_0 = \tilde{M}$ is the mask the trajectory starts from. Rewarding the net gain lets the model stop as soon as further revision would no longer help.

5 Experiments

5.1 Implementation Details

F_{seg} is built on SAM3, fine-tuned with the image encoder unfrozen. Both policies are built on MedGemma-1.5-4B-it (Sellergren et al., 2026) and trained on eight A100-80GB GPUs with LoRA (Hu et al., 2021) (r=128, α=256) on all linear layers and the vision tower unfrozen, at an effective batch size of 256 for π_{init} and 64 for π_{rev} during supervised fine-tuning. RL then uses GRPO (Shao et al., 2024) with 8 generations per prompt at temperature 1.0, and the rollouts of π_{rev} run with F_{seg} in the loop for at most four clicks per episode. Before that, π_{rev} is warm-started from π_{init} at its SFT checkpoint, so that it inherits grounding ability without also inheriting a habit of always committing to a box. The remaining hyperparameters are given in the Appendix A.4.

5.2 Experimental Setup

Evaluation set. As described in Sec. 4.2, evaluation is carried out on part of the CheXpercept subset of MIMIC-ILS, the cases whose masks physicians judged most reliable. This portion is disjoint from all training data, so that every evaluation image is unseen. Even within this subset, human error may remain, so we apply an additional filter. We obtain further evidence on whether a lesion is present and where it lies from two more sources, Chest ImaGenome (Wu et al., 2021b;a) and an LLM parse of the radiology report accompanying the X-ray, and rebuild the evaluation set from the cases on which all of them agree. The resulting set queries the same studies at three levels of location specificity: chest, naming the finding alone ("Segment the edema."; 382 positives, 444 negatives); lung, adding the side ("... in the left lung."; 429, 1,050); and zone, naming the regions directly ("... in the left mid zone and left lung base."; 330, 3,286). Appendices A.2 and B give the agreement criteria and the per-finding counts.

We additionally evaluate on CheXlocalize (Saporta et al., 2022), an external dataset whose images come from CheXpert (Irvin et al., 2019) and whose masks are drawn from scratch by radiologists. We use the 775 positive and 1,643 negative cases over the six findings that CheXlocalize shares with the set of lesions XFlow can segment, and query them at the chest level only (e.g., "Segment the consolidation"). CheXlocalize also provides segmentations drawn by a second, independent pair of radiologists, which we score through the same pipeline as the models and report as a human benchmark.

Table 1: Comparison with prior models. On CheXpercept the three question levels are averaged with equal weight, and on CheXlocalize every case is queried at the chest level. The radiologists behind the human benchmark were told which finding to draw, so F1 is not defined for them. The best score among the models is shown in **bold** and the second best is underlined.

Model	CheXpercept			CheXlocalize		
	gIoU	cIoU	F1	gIoU	cIoU	F1
Human benchmark	—	—	—	*0.314*	*0.339*	—
LISA-7B	0.014	0.024	0.401	0.082	0.119	0.491
PixelLM-7B	0.031	0.032	0.416	0.125	0.125	0.485
Text4Seg	0.035	0.038	0.402	0.119	0.132	0.475
IMIS-Net	0.080	0.087	0.413	0.005	0.007	0.423
IBISAgent-7B†	0.135	0.131	0.416	0.081	0.082	0.484
BiomedParse	0.257	0.254	0.413	0.105	0.108	0.483
RecLMIS	0.302	0.305	0.419	0.138	0.143	0.488
ROSALIA	0.621	0.688	0.945	0.233	0.289	**0.634**
XFlow (ours)	**0.703**	**0.731**	**0.961**	**0.283**	**0.331**	0.632

Baselines and metrics. We compare our model against two groups of models. The first consists of general-domain models, LISA (Lai et al., 2024), PixelLM (Ren et al., 2024), and Text4Seg (Lan et al., 2025), which take a free-form instruction and return a mask. The second covers models developed for the medical domain, BiomedParse (Zhao et al., 2024), RecLMIS (Huang et al., 2024), IMIS-Net (Cheng et al., 2025) and IBISAgent (Jiang et al., 2026)†. We also evaluate fine-tuned ROSALIA (Choi et al., 2026c) on the same MIMIC-ILS cases (D_{SFT} and D_{RL}). We report gIoU, the mean per-sample IoU, and cIoU, the ratio of total intersection to total union. F1 measures how well a model separates positive from negative cases, treating a returned mask as positive and an empty output as negative.

5.3 Main Results

Comparison with prior models. XFlow records the highest gIoU and cIoU of all models we compare against, on both CheXpercept and CheXlocalize. The margin is largest over the baselines outside the ROSALIA family, but it also holds over ROSALIA, the only prior model trained for this task. For a fair comparison, we retrain ROSALIA on exactly the cases XFlow uses, so that both models receive the same lesion supervision and the gain cannot be attributed to additional annotations. XFlow raises gIoU on CheXpercept from 0.621 to 0.703 and F1 from 0.945 to 0.961. The same pattern holds on CheXlocalize, where gIoU rises from 0.233 to 0.283 and cIoU from 0.289 to 0.331, approaching the human benchmark at 0.314 and 0.339, while F1 remains on par (0.632 vs. 0.634). Since these masks were drawn independently by radiologists, the gain carries over to an annotation protocol our training data never saw.

Effect of each stage and its training phase. We examine how much SFT and RL contribute to each stage (Table 2). Each addition improves or maintains gIoU and consistently improves cIoU on both benchmarks. Most of the gain comes from Stage 1. SFT alone reaches a gIoU of 0.668 on CheXpercept, above ROSALIA's 0.621, though its F1 of 0.932 falls short of ROSALIA's 0.945. RL raises gIoU to 0.693 and F1 to 0.961, suggesting that the outcome-based reward sharpens the presence decision and box placement that SFT only imitates. Stage 2 adds a further 0.010 gIoU on CheXpercept and 0.003 on CheXlocalize, as it only makes local corrections and can neither recover a lesion Stage 1 missed nor change a presence decision.

5.4 Qualitative Results

Figure 3 shows how the predicted masks differ across models. Apart from the ROSALIA family, the baselines do not follow the instruction in any meaningful sense and have little ability to delineate a lesion, returning regions that bear no relation to the finding that was asked for. ROSALIA responds

†IBISAgent is an intermediate training checkpoint, rather than the final model, has been publicly released, and the model cannot be reliably retrained, as the released code does not cover the full training pipeline in sufficient detail and its data construction relies on manual annotations by physicians.

Table 2: Contribution of each stage and training. Every component is added on top of the one before it. Stage 2 refines boxes Stage 1 has already emitted and never changes a presence decision, so it inherits Stage 1's F1. ROSALIA is shown for reference and columns are as in Table 1. The best score is shown in **bold** and the second best is underlined.

Model	Stage 1		Stage 2		CheXpercept			CheXlocalize		
	SFT	RL	SFT	RL	gIoU	cIoU	F1	gIoU	cIoU	F1
ROSALIA					0.621	0.688	0.945	0.233	0.289	**0.634**
XFlow	✓				0.668	0.709	0.932	0.268	0.325	0.631
	✓	✓			0.693	0.718	**0.961**	0.280	0.327	0.632
	✓	✓	✓		0.699	0.726	**0.961**	0.280	0.329	0.632
	✓	✓	✓	✓	**0.703**	**0.731**	**0.961**	**0.283**	**0.331**	0.632

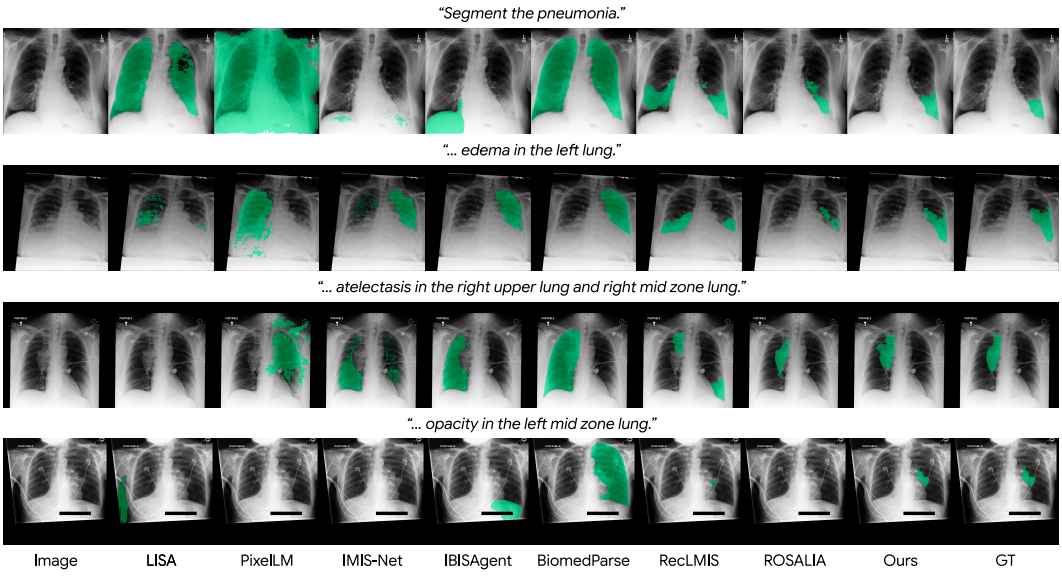


Figure 3: Qualitative comparison of XFlow against baseline models.

to the instruction but its masks are imprecise. It misses small lesions entirely, and where it does respond, the predicted boundary fails to track the lesion contour and is accompanied by scattered spurious fragments. XFlow, by contrast, recovers even small lesions lying close to the heart, and its boundaries follow the ground truth closely with far fewer stray regions. Additional qualitative examples are provided in Appendix B.2.

6 Further Analyses

6.1 Mask Noise

Physicians tend to delineate a finding as a single smooth, continuous region unless it appears in clearly separate locations. XFlow's masks are visibly closer to this than ROSALIA's, yet IoU cannot capture the difference. We therefore count connected components relative to the GT mask for the same case,

$$\Delta C = C(M^{\text{pred}}) - C(M^{\text{GT}}), \tag{9}$$

where $C(\cdot)$ is the number of connected components in a mask. ΔC thus measures how many more components a model produces than the GT. A positive value means the prediction breaks the finding into excess fragments, and a negative value means it yields fewer components than the GT. Additionally, we count specks, which we define as components smaller than 0.1% of the image area. As shown in Table 3, ROSALIA produces 5.63 more components than the GT on average, whereas

Table 3: Mask fragmentation analysis over all 1,141 positive cases in the CheXpercept evaluation set. ΔC and specks are defined in Section 6.1. *Outside GT* counts the specks that do not overlap the GT mask, and *Mask area* is the fraction of the predicted mask area occupied by specks.

Model	ΔC	**Specks**		
		count ↓	outside GT ↓	mask area % ↓
GT	—	*0.01*	—	*0.01*
XFlow (ours)	+1.71	1.69	1.06	0.14
ROSALIA	+5.63	5.57	3.43	2.60

Table 4: Expert preference study. Three medical experts ranked three blinded masks on 95 balanced positive cases, giving 285 completed rankings. Log-strengths come from two tie-aware extensions of the Bradley–Terry model, Davidson and Rao–Kupper, which differ in how they treat a tie but agree here. Values are centred to sum to zero, with 95% bootstrap intervals over cases, and a difference of d means the stronger of the two masks is preferred e^d times as often when the reader does not tie.

Model	**Log-strength ↑**		Mean rank ↓
	Davidson	Rao–Kupper	
GT	0.326 [0.17, 0.48]	0.294 [0.16, 0.44]	1.79
XFlow (ours)	0.185 [0.01, 0.38]	0.170 [0.01, 0.33]	1.88
ROSALIA	−0.511 [−0.72, −0.33]	−0.464 [−0.64, −0.30]	2.33

XFlow produces 1.71. Most of this excess consists of specks, as ROSALIA leaves 5.57 specks per case against XFlow's 1.69. Of these, 3.43 and 1.06, respectively, touch no part of the GT. These specks account for 2.6% of ROSALIA's mask area but only 0.14% of XFlow's. XFlow therefore produces fewer small mask fragments that do not belong to the lesion.

6.2 Expert Preference Study

We additionally ran a reader study in which medical experts ranked the mask quality of the ground truth, XFlow, and ROSALIA, which we include as the previous state of the art. For every case the reader saw the three masks in a counterbalanced random order, without any indication of their source, and ranked them by how well each delineated the queried finding, with ties allowed when two masks looked comparable. Three physicians each reviewed the same 95 cases, drawn so that every finding and question level is equally represented. Cardiomegaly contributes five cases at the chest level, and each of the remaining six findings contributes five cases at each of the three levels (e.g., "Segment the atelectasis", "... in the right lung", and "... in the right lung base").

Table 4 reports the result. We fit two tie-aware extensions of the Bradley–Terry model (Bradley & Terry, 1952), Davidson (Davidson, 1970) and Rao–Kupper (Rao & Kupper, 1967), which differ in how they treat a tie, and report the log-strength each assigns to the three masks. XFlow sits close to the ground truth, a difference of 0.141 under Davidson, so that a reader who separates the two prefers the ground truth 1.15 times as often, and their intervals overlap over most of their range. ROSALIA falls well short of both XFlow is preferred over it 2.01 times as often, a difference of 0.696, and the ground truth by more still, with ROSALIA's interval overlapping neither of theirs. More details on the human study are provided in Appendix C.

7 Conclusion

We introduced XFlow, a workflow model for CXR lesion segmentation. XFlow mirrors the way a radiologist reads an image, separating coarse localization from fine-grained delineation and carrying them out in sequence rather than mapping an image to a mask in one step. XFlow substantially outperforms prior models and surpasses ROSALIA, the previous state of the art, even when the two are trained on the same lesion annotations. In future work, the intermediate steps and masks XFlow produces could serve as a foundation for other CXR tasks.

AI USE STATEMENT

In this work, we used generative AI tools to implement the methods described in this paper and to edit the manuscript for grammar and readability. The research idea, the workflow formulation, and the experimental design are our own; AI tools were used only to turn those ideas into code, never to originate or refine them. The resulting code was tested continuously throughout development and training, and its behaviour was verified against the intended design by the authors at every stage. All AI-assisted edits to the text were reviewed by the authors, and no claim, number, or citation in this paper was produced by a generative AI tool without verification. We have not used generative AI tools to generate synthetic data sets, to formulate mathematical claims, to propose or refine hypotheses, to design our methodology or experiments, or to interpret results. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All data used in this work are drawn from publicly available sources, namely MIMIC-ILS, Chest ImaGenome, and CheXlocalize. Full training details, including data construction and hyperparameters, are provided in Appendix A, and evaluation details, including the construction of the evaluation set, are provided in Appendix B. We will publicly release the code and trained model weights.

REFERENCES

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39:324, 1952. URL https://api.semanticscholar.org/CorpusID:125209808.

Joshua Broder. Imaging the chest: the chest radiograph. *Diagnostic imaging for the emergency physician*, pp. 185, 2011.

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.

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, Nalee Kim, Jeong Yun Jang, Hyunju Shin, Hyunki Park, Sang Hoon Seo, and Edward Choi. Chexpercept: A benchmark for evaluating expert-level lesion perception in chest x-rays. *arXiv preprint arXiv:2606.21020*, 2026a.

Geon Choi, Hangyul Yoon, Hyunju Shin, Hyunki Park, Sang Hoon Seo, Eunho Yang, and Edward Choi. MIMIC-CXR-Ext-ILS: Lesion Segmentation Masks and Instruction-Answer Pairs for Chest X-rays. *PhysioNet*, March 2026b. doi: 10.13026/8ejy-4t06. URL https://doi.org/10.13026/8ejy-4t06. Version 1.0.0.

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, 2026c.

Matias Cosarinsky, Nicolas Gaggion, Rodrigo Echeveste, and Enzo Ferrante. Chexmask-u: Quantifying uncertainty in landmark-based anatomical segmentation for x-ray images. *arXiv preprint arXiv:2512.10715*, 2025.

Roger R. Davidson. On extending the bradley-terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association*, 65:317–328, 1970. URL https://api.semanticscholar.org/CorpusID:121759206.

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

Xiaoshuang Huang, Hongxiang Li, Meng Cao, Long Chen, Chenyu You, and Dong An. Cross-modal conditioned reconstruction for language-guided medical image segmentation. *IEEE Transactions on Medical Imaging*, 44(4):1821–1835, 2024.

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, 2025.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 590–597, 2019.

Yankai Jiang, Qiaoru Li, Binlu Xu, Haoran Sun, Chao Ding, Junting Dong, Yuxiang Cai, Xuhong Zhang, and Jianwei Yin. Ibisagent: Reinforcing pixel-level visual reasoning in mllms for universal biomedical object referring and segmentation. *arXiv preprint arXiv:2601.03054*, 2026.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

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.

Mengcheng Lan, Chaofeng Chen, Yue Zhou, Jiaxing Xu, Yiping Ke, Xinjiang Wang, Litong Feng, and Wei Zhang. Text4seg: Reimagining image segmentation as text generation. In *International Conference on Learning Representations*, volume 2025, pp. 1634–1661, 2025.

Zihan Li, Yunxiang Li, Qingde Li, Puyang Wang, Dazhou Guo, Le Lu, Dakai Jin, You Zhang, and Qingqi Hong. Lvit: language meets vision transformer in medical image segmentation. *IEEE transactions on medical imaging*, 43(1):96–107, 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*, 2025.

Jun Ma, Zongxin Yang, Sumin Kim, Bihui Chen, Mohammed Baharoon, Adibvafa Fallahpour, Reza Asakereh, Hongwei Lyu, and Bo Wang. Medsam2: Segment anything in 3d medical images and videos. *arXiv preprint arXiv:2504.03600*, 2025.

P. V. Rao and L. L. Kupper. Ties in paired-comparison experiments: A generalization of the bradley-terry model. *Journal of the American Statistical Association*, 62(317):194–204, 1967. doi: 10.1080/01621459.1967.10482901. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1967.10482901.

Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26364–26373. IEEE, 2024.

Adriel Saporta, Xiaotong Gui, Ashwin Agrawal, Anuj Pareek, Steven QH Truong, Chanh DT Nguyen, Van-Doan Ngo, Jayne Seekins, Francis G Blankenberg, Andrew Y Ng, et al. Benchmarking saliency methods for chest x-ray interpretation. *Nature Machine Intelligence*, 4(10): 867–878, 2022.

Constantin Seibold, Alexander Jaus, Matthias A Fink, Moon Kim, Simon Reiß, Ken Herrmann, Jens Kleesiek, and Rainer Stiefelhagen. Accurate fine-grained segmentation of human anatomy in radiographs via volumetric pseudo-labeling. *arXiv preprint arXiv:2306.03934*, 2023.

Andrew Sellergren, Chufan Gao, Fereshteh Mahvar, Timo Kohlberger, Fayaz Jamil, Madeleine Traverse, Alberto Tono, Bashir Sadjad, Lin Yang, Charles Lau, et al. Medgemma 1.5 technical report. *arXiv preprint arXiv:2604.05081*, 2026.

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.

Konstantin Sofiiuk, Ilya A Petrov, and Anton Konushin. Reviving iterative training with mask guidance for interactive segmentation. In *2022 IEEE international conference on image processing (ICIP)*, pp. 3141–3145. IEEE, 2022.

Joy Wu, Nkechinyere Agu, Ismini Lourentzou, Arjun Sharma, Joseph Paguio, Jasper Seth Yao, Edward Christopher Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, Leo Anthony Celi, Tanveer Syeda-Mahmood, and Mehdi Moradi. Chest ImaGenome Dataset. *PhysioNet*, July 2021a. doi: 10.13026/wv01-y230. URL https://doi.org/10.13026/wv01-y230. Version 1.0.0.

Joy T Wu, Nkechinyere N Agu, Ismini Lourentzou, Arjun Sharma, Joseph A Paguio, Jasper S Yao, Edward C Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, et al. Chest imagenome dataset for clinical reasoning. *arXiv preprint arXiv:2108.00316*, 2021b.

Ning Xu, Brian Price, Scott Cohen, Jimei Yang, and Thomas S Huang. Deep interactive object selection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 373–381, 2016.

Theodore Zhao, Yu Gu, Jianwei Yang, Naoto Usuyama, Ho Hin Lee, Tristan Naumann, Jianfeng Gao, Angela Crabtree, Jacob Abel, Christine Moung-Wen, et al. Biomedparse: a biomedical foundation model for image parsing of everything everywhere all at once. *arXiv preprint arXiv:2405.12971*, 2024.

Muzhi Zhu, Yuzhuo Tian, Hao Chen, Chunluan Zhou, Qingpei Guo, Yang Liu, Ming Yang, and Chunhua Shen. Segagent: Exploring pixel understanding capabilities in mllms by imitating human annotator trajectories. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3686–3696. IEEE, 2025.

A TRAINING

A.1 DATA CURATION

Every split is built from two sources. MIMIC-ILS (Choi et al., 2026c;b) supplies the images, the lesion masks and the report-derived labels, and CheXpercept (Choi et al., 2026a) supplies the subset of those cases whose masks medical experts reviewed and judged reliable. MIMIC-ILS is large enough to train on, while CheXpercept is small and trustworthy enough to reinforce against and to score on. We therefore hold all 2,100 CheXpercept radiographs out of the supervised split and divide them in half, one half for RL and one for evaluation. What follows describes the splits themselves. Each stage builds its own prompts, boxes and masks from the studies its split contains, so one study gives rise to many samples, and a question asked at three levels of location specificity gives rise to three.

Table 5: How the splits divide the two sources. CheXpercept is a curated subset of the MIMIC-ILS training split, so holding it out leaves the supervised split disjoint from the other two by construction. Counts are radiographs; MIMIC-ILS carries one frontal radiograph per study, so they are also study counts.

Split	Drawn from	Radiographs	Selection
D_{SFT}	MIMIC-ILS train $\setminus$ CheXpercept	185,488	everything the review did not cover
	MIMIC-ILS val	1,508	MIMIC-ILS's own split, unchanged
D_{RL}	MIMIC-ILS train $\cap$ CheXpercept	1,296	70% of the agreed set, by patient
D_{eval}	MIMIC-ILS train $\cap$ CheXpercept	560	the remaining 30%
MIMIC-ILS train, before the exclusion		*187,588*	
CheXpercept, all of it inside MIMIC-ILS train		*2,100*	
of which the three sources agree on		*1,856*	

D_{SFT}. The MIMIC-ILS training split with the CheXpercept radiographs removed, together with MIMIC-ILS's own validation split. CheXpercept was drawn entirely from the training split, all 2,100 of its radiographs sitting there and none in validation or test, so removing them leaves 185,488 of the original 187,588 and leaves the 1,508 validation radiographs untouched.

D_{RL} and D_{eval}. Both come from CheXpercept, and both are filtered by the three-source agreement of Section A.2 before being divided, which leaves 1,856 of the 2,100 radiographs.

The division is by patient rather than by radiograph. Those 1,856 radiographs come from 1,526 patients, and 226 of them contribute more than one, covering 556 radiographs in all, so a split on radiograph identity would put the same chest on both sides. Patients are assigned greedily to whichever side is furthest below its quota on the scarcest stratum that patient carries, because matching the overall target ratios is not enough on its own: the evaluation's power to test per-side discrimination rests on the 85 unilateral (radiograph, finding) pairs and on the per-finding positives, both of which a ratio-matched split can strand almost entirely on one side. The result is 1,296 radiographs for D_{RL} and 560 for D_{eval}, with no patient in both.

A.2 THREE-SOURCE AGREEMENT

CheXpercept's labels can be wrong, and an evaluation set inherits whatever it gets wrong, so we check each of them against two further sources and keep only what all three support. All three record both whether a finding is present and where it is, which is what makes the check possible; they differ in how they record it.

- **CheXpercept** gives the list of zones a finding occupies.
- **The radiology report** gives free text, which we parse with MedGemma into a present or absent call per finding and region.
- **Chest ImaGenome** gives a yes or no per finding and region in its silver scene graphs.

Each is then reduced to the same verdict, present or absent for one (radiograph, finding, region). A triple is admitted only when the three verdicts are identical, and takes that verdict as its label; anything else, including a missing source, is dropped. 1,856 of the 2,100 radiographs keep at least one admitted triple, and those are the radiographs D_{RL} and D_{eval} are drawn from.

A.3 Building the Training Examples

Every example is automatically generated from what a split already carries, the radiograph, the lesion mask and the report-derived label. Table 6 reports how many examples this yields for each stage and phase.

Table 6: Examples built from each split. The unit differs by stage: π_{init} is trained one instruction at a time, while one π_{rev} example is a whole revision trajectory of up to four turns.

Stage	Split	Unit	Examples
π_{init} SFT	D_{SFT}	instruction	1,031,506 train / 8,246 val
π_{init} RL	D_{RL}	instruction	13,488 (2,605 positive, 10,883 negative)
π_{rev} SFT	D_{SFT}	trajectory	469,011 train / 4,423 val
π_{rev} RL	D_{RL}	starting state	58,320

Initial segmentation. A training example for this stage runs from a chest X-ray and a user question to the bounding boxes the model is to emit. Questions are generated by filling templates with the (finding, region) labels that accompany each image in D_{SFT}, which keeps the phrasing varied while the underlying query stays grounded in the annotation. The target answer begins with the lungs. We obtain their boxes by running CXAS and CheXmask-U to extract lung masks and taking the tightest box around each. Which lungs appear follows the question, so a question about the thorax as a whole grounds both lungs and a question restricted to one side grounds only that one. The lesion boxes follow immediately, derived from the MIMIC-ILS mask for the queried finding.

Every element of the answer is wrapped in markers such as `<think>` and `<|box_start|>` so that it can be parsed from the generated text. Figure 4 gives the full example. RL needs only the prompt side of this, since the model generates its own answer and is scored on the outcome rather than on matching a target.

Segmentation revision. An SFT example for this stage is a conversation that begins with the X-ray, the initial mask overlaid on it, and the same user question as in the first stage. The trajectory then alternates between the two sides of the interaction. The model issues a point prompt, that prompt is run through F_{seg}, and the resulting mask is overlaid on the X-ray to form the next observation, which comes back as the following turn. Figure 5 gives the full example. Here too RL needs only the prompt side, with the trajectory produced by the model itself.

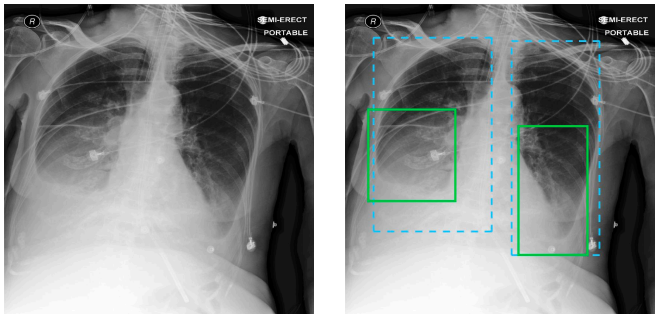
Since no model is run while the data is built, the initial mask has to be synthesized. We perturb the ground-truth lesion box at random to obtain an imperfect box and pass it with the image through F_{seg}, which yields a mask that stands in for what the first stage would supply. The gap between this mask and M^{GT} then defines the false-negative region the model has missed and the false-positive region it has over-segmented, and we pick the click that best repairs it following the standard oracle of interactive segmentation (Sofiiuk et al., 2022).

The oracle takes the distance transform of each region and clicks the deepest point of whichever is larger, a positive click if the missed region dominates and a negative one otherwise, since the deepest point leaves the widest margin for error. Prompting F_{seg} with that click gives an improved mask, from which the next click is computed in turn, and the alternating masks and clicks form the trajectory. The distance transform labels every pixel of a region with its distance to the nearest pixel outside it, so its maximum is the point furthest from the region's boundary.

Algorithm 1 states the procedure as it is run for one ground-truth component, the components of a finding being revised independently. Two guards keep a trajectory usable and are omitted from the pseudocode for readability. A click of a given polarity may not land within 25 pixels of an earlier

click of the same polarity, and a trajectory that clicked without improving its mask is discarded rather than demonstrated.

Initial Segmentation



[system]
You are a precise and expert medical image localization agent. Your mission is to localize a target lesion in a chest X-ray image by predicting tight bounding box(es) around each region of the target. You will be given an image and an instruction identifying the target.

IMPORTANT — LOCATION SCOPE:
• If the instruction NAMES a specific anatomical location (e.g. "in the left lung base", "in the right lower zone"), you must output boxes ONLY for the target region(s) that fall within that named location. Do NOT include lesions in other anatomical regions even if they are visible.
• If the instruction does NOT name any location, output boxes for ALL regions of the target visible in the image, no matter where they are.

Your response must strictly follow this structure: <|object_ref_start|>{target}<|object_ref_end|><|box_start|>(x1, y1),(x2, y2)<|box_end|> — where (x1, y1) is the top-left corner and (x2, y2) the bottom-right corner. All coordinates must be normalized to a 0.0-1.0 scale relative to the image dimensions and formatted to four decimal places, for example: <|box_start|>(0.3496, 0.2930),(0.7891, 0.6406)<|box_end|>. When the target consists of multiple disjoint regions, emit one <|box_start|>...<|box_end|> block per region under a single <|object_ref_start|>{target}<|object_ref_end|>, sorted by area in descending order. If the target is NOT present in the image, emit the same skeleton with an empty box: <|object_ref_start|>{target}<|object_ref_end|><|box_start|><|box_end|> Think before you answer. Inside <think>...</think>, ground the anatomy: emit the box for each lung the question concerns, or both lungs if the question names no location. After the tags, answer the question that was asked: the box(es) for the target within the named location, and one sentence about that location only.

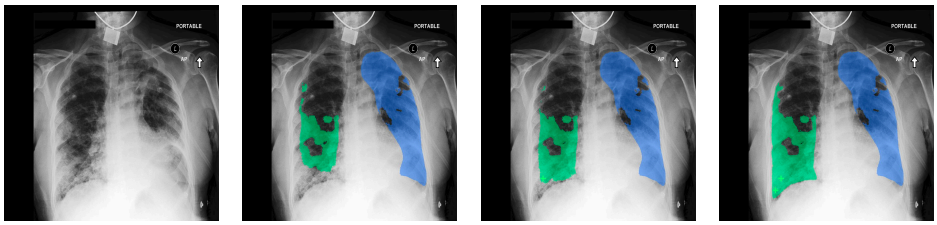
[user]
<image>Where can I find the atelectasis in this image?

[target]
<think><|object_ref_start|>right lung<|object_ref_end|><|box_start|>(0.0918, 0.1094),(0.4736, 0.7344)<|box_end|><|object_ref_start|>left lung<|object_ref_end|><|box_start|>(0.5371, 0.1191),(0.8213, 0.8105)<|box_end|></think>
<|object_ref_start|>atelectasis<|object_ref_end|><|box_start|>(0.0723, 0.3379),(0.3594, 0.6387)<|box_end|><|box_start|>(0.5566, 0.3906),(0.7852, 0.8125)<|box_end|>

It is located in the left lung base and right lung base.

Figure 4: Example of SFT data for initial segmentation. The image shown at the top is what the `<image>` placeholder stands for; the box overlay is drawn here for illustration and is not part of the input.

Segmentation Revision



[system]
You are a precise and expert medical image localization agent refining an existing segmentation of a chest X-ray.
You are given the original image and an overlay of the current segmentation.
OVERLAY COLORS:
• GREEN — the region you are refining. This is the only region you may change.
• BLUE — other lesion regions in the same image. They are being handled separately. Ignore them; never click on them.

IMPORTANT — WHAT TO DO:
• Compare the GREEN region against the lesion you can see in the X-ray.
• If GREEN covers area that is NOT lesion, emit a NEGATIVE click there.
• If the lesion extends beyond GREEN, emit a POSITIVE click on the part that was missed.
• If GREEN already matches the lesion, emit <stop/> and change nothing.
Stay on the GREEN region: click on it or just outside its edge, never across the image.

DO NOT REPEAT YOURSELF:
• Your earlier clicks are drawn on the overlay — a green + where you clicked positive, a red x where you clicked negative.
• Never place the same kind of click at a spot you already used. It was already applied; repeating it changes almost nothing.
• Each turn, look at what your last click DID to the GREEN region, and move to a different part of the boundary that is still wrong.
Coordinates are normalized to [0, 1] as (x, y).
Answer with exactly one action and nothing else.

[user]
<image><image>Outline the edema.

[target]
<click type="positive">(0.2881, 0.8024)</click>

[user]
<image>

[target]
<click type="positive">(0.2619, 0.8571)</click>

[user]
<image>

[target]
<stop/>

Figure 5: Example of SFT data for segmentation revision. The images shown at the top are what the `<image>` placeholders stand for.

Algorithm 1 Construction of one revision trajectory.

Require: image I, ground-truth component G, segmenter F_{seg}, click budget $T = 4$, stop threshold $\tau = 0.70$

$b \leftarrow \textsc{Perturb}(\textsc{Box}(G))$ ▷ stands in for a first-stage box
$M \leftarrow F_{\text{seg}}(I, b)$; $P \leftarrow \emptyset$; $\mathcal{T} \leftarrow \emptyset$
for $t = 0$ **to** T **do**
  if $t = T$ **or** $\text{IoU}(M, G) \geq \tau$ **then**
    append $(M, \texttt{<stop/>})$ to $\mathcal{T}$; **break**
  end if
  $D_{\text{miss}} \leftarrow \textsc{DistanceTransform}(G \setminus M)$ ▷ what the mask has missed
  $D_{\text{over}} \leftarrow \textsc{DistanceTransform}(M \setminus G)$ ▷ what it has over-segmented
  if $\max D_{\text{miss}} \geq \max D_{\text{over}}$ **then**
    $(s, p) \leftarrow (\text{positive}, \arg\max D_{\text{miss}})$
  else
    $(s, p) \leftarrow (\text{negative}, \arg\max D_{\text{over}})$
  end if
  append $\big(M, (s, p)\big)$ to $\mathcal{T}$ ▷ one turn: the observation, then the action taken on it
  $P \leftarrow P \cup \{(s, p)\}$
  grow b to contain the positive clicks in P
  $M \leftarrow F_{\text{seg}}(I, b, P, M)$
end for
return $\mathcal{T}$

A.4 Training Details

Table 7 lists the hyperparameters used to train the three components of XFlow, the segmenter F_{seg} and the two policies π_{init} and π_{rev}.

Table 7: Hyperparameters. Each policy is trained in two phases, and the RL phase starts from the SFT checkpoint of the same policy. A dash marks a setting that does not apply to that column. Effective batch size is per-device batch × gradient accumulation × 8 GPUs, counted in samples for SFT and in generated completions for GRPO.

	F_{seg}	π_{init}		π_{rev}	
		SFT	RL	SFT	RL
Initialized from	SAM3	MedGemma	π_{init} SFT	π_{init} SFT	π_{rev} SFT
Adapted weights	all	LoRA + ViT	LoRA	LoRA + ViT	LoRA
LoRA r / α	—	128 / 256	128 / 256	128 / 256	128 / 256
Optimizer	AdamW, `bf16`				
Learning rate	1×10^{-4}	1×10^{-5}	1×10^{-5}	1×10^{-5}	1×10^{-6}
Encoder rate	0.6×	0.3×	—	0.3×	—
Layer decay	0.9	—	—	—	—
Weight decay	0.1	0	0	0	0
Gradient clip	0.1	1.0	1.0	1.0	1.0
Effective batch	16	256	32	64	64
Epochs	5	15	10	10	10
Generations / prompt	—	—	8	—	8
Temperature / top-p	—	—	1.0 / 1.0	—	1.0 / 0.95
KL coefficient β	—	—	0.1	—	0.02

F_{seg} is trained at 1024×1024 with the SAM 2 loss, $20 \cdot \mathcal{L}_{\text{focal}} + \mathcal{L}_{\text{dice}} + \mathcal{L}_{\text{IoU}}$, taking three mask hypotheses per prompt and supervising only the one with the lowest loss. The encoder is unfrozen throughout at $0.6\times$ the decoder rate with a layer-wise decay of 0.9 along the trunk, which follows MedSAM2 (Ma et al., 2025).

Both policies apply LoRA to every linear layer, but the two phases differ in what else moves. During SFT the vision tower is unfrozen at $0.3\times$ the base rate, since the boxes and clicks the policies emit

are image coordinates and the tower has never seen a chest radiograph. During RL the adapters are restricted to the language layers and the tower is left alone.

B EVALUATION

B.1 EVALUATION SET CONSTRUCTION

The evaluation set is built from D_{eval}, the 560 radiographs held out in Section A.3. Every triple the three sources agreed on becomes one question, asked at whichever levels of location specificity apply to it, so one radiograph contributes several rows and the same finding appears at more than one level. Table 8 gives the distribution.

Table 8: Distribution of the evaluation set. A positive row carries a mask and is scored by IoU against it; a negative row is answered correctly only with an empty mask. Cardiomegaly appears at the chest level alone, being a mediastinal finding for which the side- and zone-level questions do not arise.

Finding	Chest		Lung		Zone	
	positive	negative	positive	negative	positive	negative
Atelectasis	61	77	65	204	60	644
Cardiomegaly	57	25	—	—	—	—
Consolidation	49	102	62	254	57	772
Edema	58	103	110	253	40	760
Effusion	60	23	61	53	61	220
Opacity	60	14	89	38	72	134
Pneumonia	37	100	42	248	40	756
Total	382	444	429	1,050	330	3,286

Negatives outnumber positives further at each level because the question grows more specific while the finding does not move: a consolidation confined to the right lung base is absent from the left lung, and absent from every other zone of the right one. The zone level therefore holds 68.7% of all the negatives against the chest level's 9.3%, which is why the three levels are averaged with equal weight rather than pooled. Pooling would report a model on how well it declines the zone-level questions and would barely register what it gives up when a question names no location at all.

B.2 ADDITIONAL QUALITATIVE EXAMPLES

Figure 6 shows further comparisons of the predicted masks across models, following the layout of Figure 3.

C HUMAN STUDY

C.1 LABELING PROCESS

For each case we prepared a panel showing the X-ray alongside the ground-truth mask and the predictions of ROSALIA and XFlow, each overlaid on the image, and gave it to the readers together with an annotation sheet. Cases were drawn at random from the positives on which a ground-truth mask exists and both models returned a mask. The three masks appear as A, B and C in a random order, with no indication of their source, and the reader ranks them. The review panel and the annotation sheet are shown in Figure 7.

C.2 EXPERT PROFILES

The three readers are board-certified radiation oncologists with 8, 4, and 4 years of clinical experience, respectively.

C.3 PREFERENCE MODELING

Each case gives one ranking of the three masks with ties allowed, which we read as three pairwise outcomes, each either a strict preference or a tie. Writing $\pi_i > 0$ for the strength of mask i, the Bradley–Terry model (Bradley & Terry, 1952) gives

$$P(i \succ j) = \frac{\pi_i}{\pi_i + \pi_j}, \tag{10}$$

and has no outcome left for a tie. Scoring a tie as half a win for each side would make it usable, but that discards exactly what the tie reports, that the reader could not separate the two, and enters one observation as two. We therefore fit two models that give the tie a probability of its own, and report both.

Davidson. A tie is treated as a third outcome alongside the two wins, so that each comparison resolves into one of three possibilities,

$$P(i \succ j) = \frac{\pi_i}{D}, \qquad P(i \sim j) = \frac{\nu\sqrt{\pi_i \pi_j}}{D}, \qquad D = \pi_i + \pi_j + \nu\sqrt{\pi_i \pi_j}, \tag{11}$$

where $\nu \geq 0$ governs how often ties occur, its reciprocal serving as an index of how finely the reader discriminates (Davidson, 1970). The tie term is proportional to the geometric mean of the two strengths, so for a given ratio π_i/π_j a tie is most likely when the pair is evenly matched, which is the behaviour a reader shows. Setting $\nu = 0$ recovers Bradley–Terry.

Rao–Kupper. A tie instead arises when neither mask is far enough ahead to be declared the winner, with a threshold $\theta \geq 1$ setting how large a gap the reader needs before committing (Rao & Kupper, 1967),

$$P(i \succ j) = \frac{\pi_i}{\pi_i + \theta\pi_j}, \qquad P(i \sim j) = \frac{\pi_i \pi_j (\theta^2 - 1)}{(\pi_i + \theta\pi_j)(\pi_j + \theta\pi_i)}, \tag{12}$$

which again reduces to Bradley–Terry at $\theta = 1$. Unlike Davidson's, this tie probability is not symmetric in the two strengths, so the two models place ties differently across the pairs.

Fitting and reporting. Both are fit by maximum likelihood over all 855 pairwise outcomes at once, with ν and θ estimated jointly with the strengths rather than set in advance. Against an observed tie rate of 9.9% this gives $\nu = 0.232$ and $\theta = 1.241$. We report $\beta_i = \log \pi_i$, centred so that $\sum_i \beta_i = 0$, so a difference $\beta_i - \beta_j = d$ means that among the comparisons a reader did not tie, mask i was preferred e^d times as often as mask j. Intervals come from $2{,}000$ bootstrap resamples over *cases* rather than over comparisons, because the three comparisons a case contributes come from one ranking and are not independent.

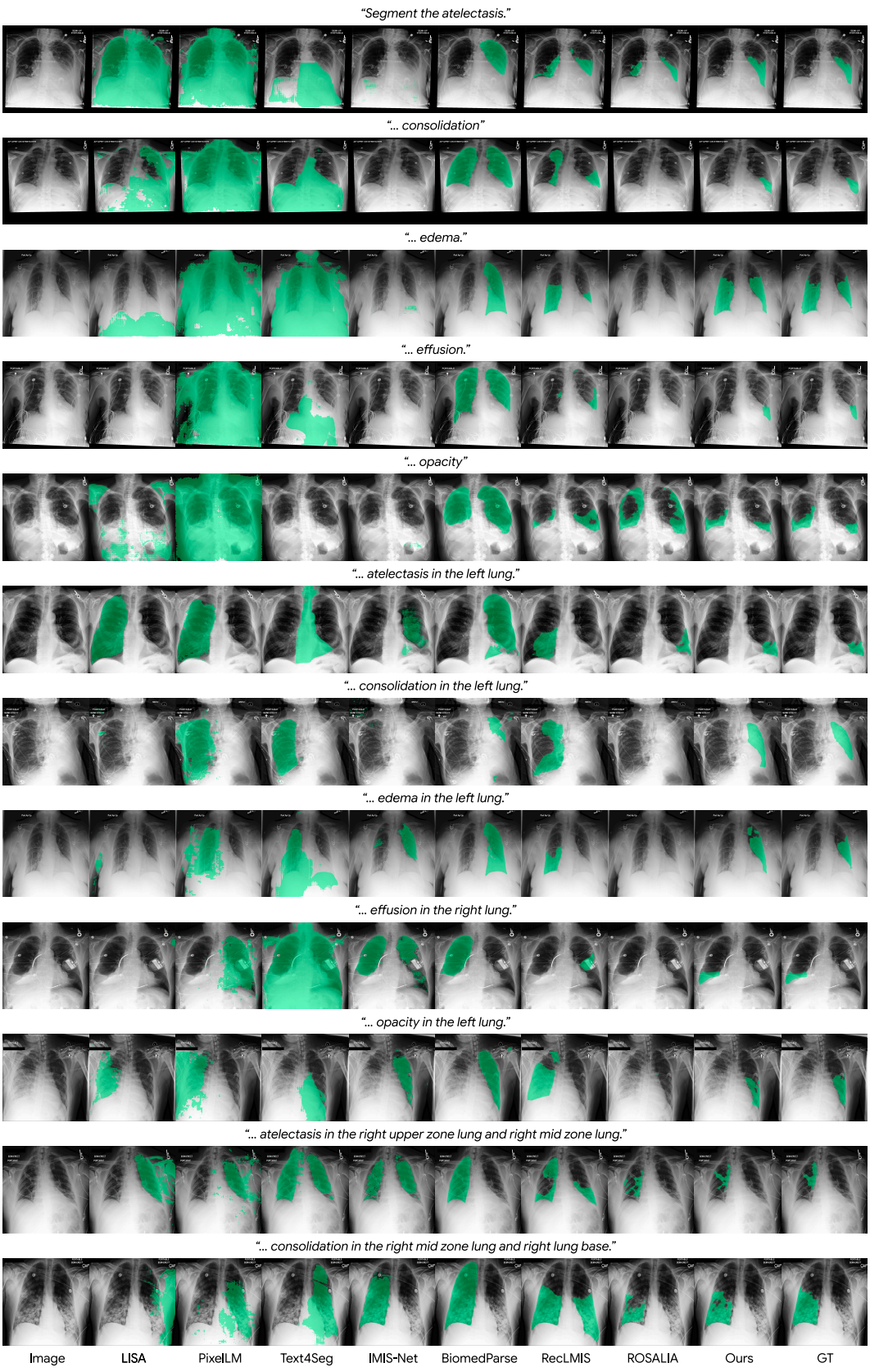
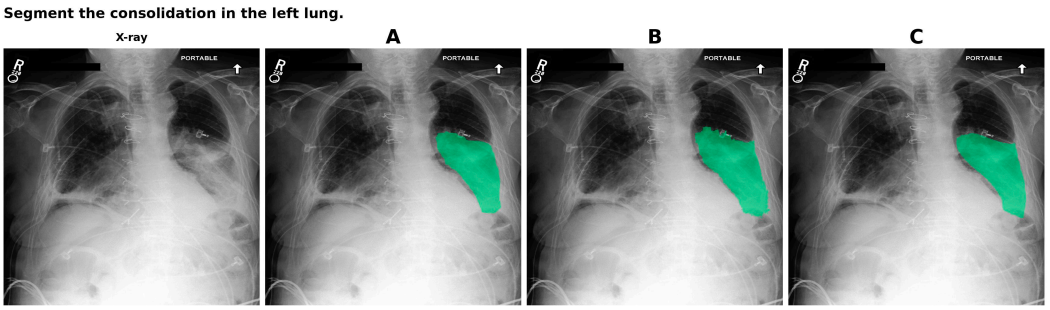


Figure 6: Additional qualitative comparisons across models. The layout follows Figure 3.



reader_sheets .XLSX

File Edit View Insert Format Data Tools Gemini Help

Menus 100% $ % .0 .00 123 Defaul... 11 B I

A1 fx Seq

	A	B	C	D	E	F	G	H
1	Seq	Image	Question	Rank A	Rank B	Rank C	Check	Notes
2	1	001.png	Segment the consolidation in the left lung.	2	3	1		
3	2	002.png	Segment the effusion in the right lung base.	2	1	3		
4	3	003.png	Segment the effusion in the right lung.	3	1	2		
5	4	004.png	Segment the edema in the left mid zone lung and left lung base.	3	1	2		
6	5	005.png	Segment the opacity.	2	1	3		
7	6	006.png	Segment the edema.	3	1	2		
8	7	007.png	Segment the pneumonia.	1	2	3		
9	8	008.png	Segment the pneumonia in the right mid zone lung and right lung base.	3	1	2		
10	9	009.png	Segment the edema in the right mid zone lung and right lung base.	3	2	1		
11	10	010.png	Segment the atelectasis in the right lung.	2	1	3		
12	11	011.png	Segment the edema.	3	1	2		
13	12	012.png	Segment the edema in the left lung.	2	3	1		
14	13	013.png	Segment the consolidation.	3	1	2		
15	14	014.png	Segment the atelectasis.	3	1	2		
16	15	015.png	Segment the opacity.	3	1	2		
17	16	016.png	Segment the consolidation in the left lung.	2	1	3		
18	17	017.png	Segment the atelectasis in the left lung base.	1	1	3		
19	18	018.png	Segment the opacity in the right mid zone lung and right lung base.	1	3	2		
20	19	019.png	Segment the effusion in the left lung base.	3	1	1		

Figure 7: Review panel and annotation sheet.