Appendix

<Table of Contents>

A. Grounded Lesion Mask Generation

A.1. Report Pre-Processing

Textual information for the CXR images from MIMIC-CXR was extracted from their corresponding radiology reports. From these raw reports, we extract the findings, impression, and last paragraph sections following the official MIMIC report pre-processing code. We then adhere to a hierarchical fallback logic to select a single representative text section for each study: the impression section is used if the findings section is missing, and the last paragraph is used if the impression is also absent. Studies that lack all three of these sections are excluded.

A.2. Large Language Models and Prompts

To extract information from the pre-processed report section, our pipeline employs two distinct large language models (LLMs). The initial report structuring step utilizes Mistral-Small-3.1-24B-Instruct-2503. Using the prompt shown in Figure 9, we extract lesion information from the report as six-element tuples. For the subsequent location mapping step, we employ medgemma-27b-text-it, which is specialized in the medical domain. Using the prompt shown in Figure 10, we normalize the lesion's location in a two-step process. In compliance with the PhysioNet credentialed data use agreement for MIMIC-CXR, both models were run on our local GPU setup.

A.3. Vision Models and Characteristics

Pretrained HybridGNet. We utilized a HybridGNet model, pretrained on the CheXMask dataset [11–13], to segment the right lung, left lung, and heart. It demonstrates robust segmentation performance for these three organs, even in challenging CXRs from patients with severe conditions characterized by dense opacities.The resulting masks serve multiple, distinct roles in our pipeline. The heart mask is used directly as the ground-truth lesion mask for cardiomegaly. The right and left lung masks serve two purposes: they are used as the (L_r, L_l) inputs in Algorithm 1, and they are also merged with the heart mask to define the editing region for RadEdit. Details regarding the use of this model were omitted from the main text for brevity.

RadEdit. This diffusion-based image editing model takes a chest X-ray image and a text prompt as input [38]. To transform the input into a normal-appearing image, we used the standard prompt on which RadEdit was trained: "No acute cardiopulmonary process". Additionally, it requires a mask specifying the editing region. For this, we used the merged masks from the pretrained HybridGNet described above. Notably, we used the original MIMIC-CXR dataset [21, 22], which contains DICOM files, rather than the MIMIC-CXR-JPG version [20]. This is because RadEdit was trained on the original MIMIC-CXR, and we observed that inputting the histogram-equalized MIMIC-CXR-JPG images significantly degraded the quality of the edited image.

CXAS. Designed for anatomy segmentation in CXRs, this model is capable of segmenting 159 anatomical region classes [41]. Specifically, in our research, we input the opacity-removed images (the output of RadEdit) into CXAS to segment the anatomy. This is because CXAS tends to produce lower-quality anatomy masks for patients with significant opacities.

Pretrained YOLO. For lesion detection, we employed a YOLO model, specifically utilizing the checkpoint from the submitted solution in the VinBigData Chest X-ray Abnormalities Detection competition [35, 37]. Although this model can detect various types of lesions (aortic enlargement, atelectasis, calcification, consolidation, ILD, infiltration, lung opacity, nodule/mass, other lesion, pleural effusion, pleural thickening, pneumothorax, pulmonary fibrosis), we filtered its outputs to retain only those findings considered hyperintense lesions. We therefore excluded aortic enlargement, other lesion, and pneumothorax from the detection categories.

A.4. Thresholds for Lesion Mask Generation

The thresholds in Equation 1 and Algorithm 1 were carefully calibrated to ensure maximum mask quality. The final values were determined through an iterative process involving multiple quality checks by a physician, who identified the settings that maximized the yield of high-quality masks. The finalized thresholds are summarized in Table 7. With the exception of edema, the threshold settings are identical for all other lesion types. We set the threshold values for edema lower than for other lesion types because it tends to spread widely throughout the lungs.

Table 7. Threshold values used to generate the lesion masks. 'General' lesions refer to all lesions other than edema.

Threshold	General	Edema
τ_{ano}	0.10	0.01
τ_{anatomy}	0.25	0.25
τ_{conf}	0.20	0.01
τ_{signal}	0.20	0.20
τ_{size}	0.10	0.10

A.5. Lesion Mask Post-Processing

To further enhance the quality of the final lesion masks, additional post-processing steps were applied. Sequential erosion and dilation operations are used to remove small, scattered noise. Although omitted from the main text for brevity, this

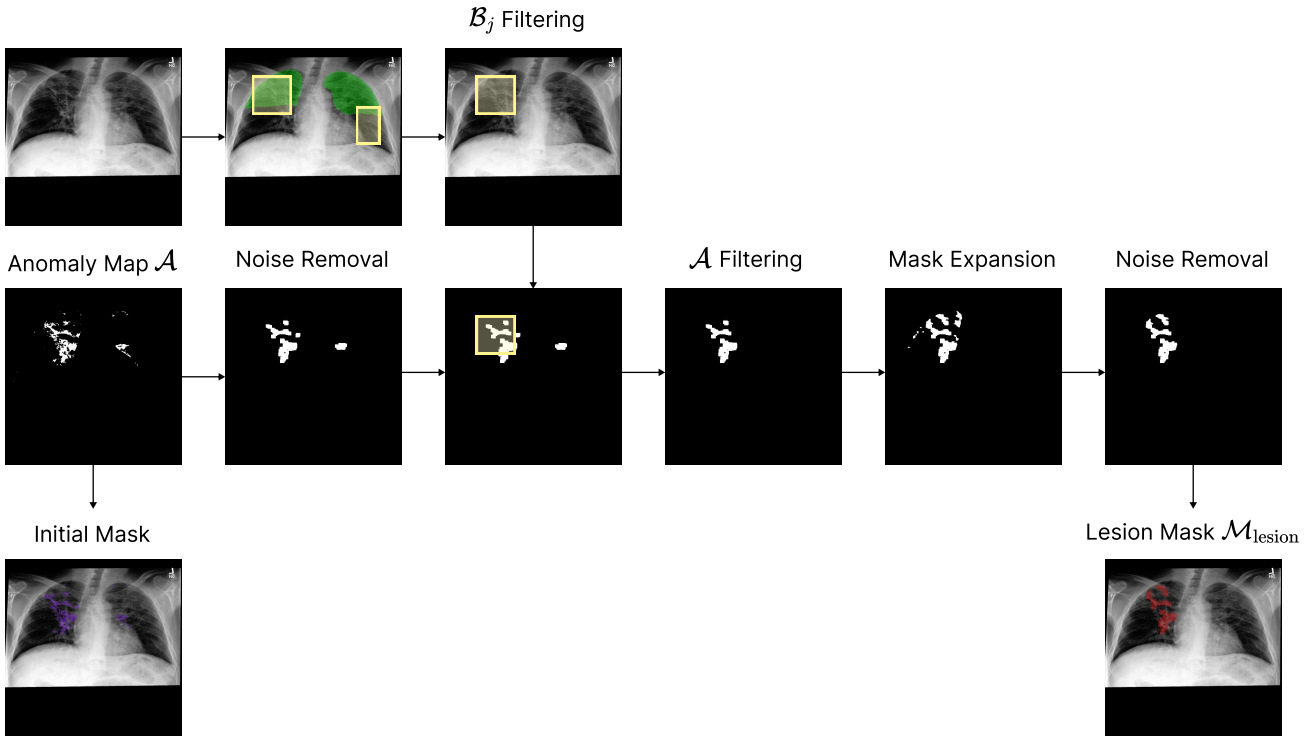


Figure 8. An example of detailed lesion mask generation where the report mentions "Areas of streaky opacity are again seen in the upper lobes." but the mask is grounded only to the right upper lobe. In the top row, the yellow lesion box mask in the left lung is discarded due to insufficient overlap with the green upper lung mask. The remaining box mask in the right lung is then used to filter the anomaly map. Intermediate post-processing steps, including noise removal and mask expansion, are applied to enhance the final mask quality.

noise removal step is also performed prior to filtering the anomaly map with the lesion box mask. We also expanded the lesion masks to include adjacent pixels with similar intensity values for more complete segmentation. Furthermore, specifically for effusions at the lung base, we incorporated the lower portion of the lung masks from the pretrained HybridGNet to ensure clean coverage extending to the costophrenic angle. The detailed process is illustrated in Figure 8.

A.6. Empty Location

We extract an "empty location" during the location verification step. An empty location is defined as a lung region where a specific lesion is not present. We designated a lung region as an empty location if it did not overlap at all with the anatomy masks corresponding to the reported location. To identify locations that are truly free of any reported lesions, we compute this not only for the seven major lesions but for all lesions mentioned in the report, and utilize this information in a subsequent data generation step.

A.7. Sample Discarding and Pipeline Recall

In our pipeline, samples that did not meet the strict cross-model consistency criteria were excluded, and the resulting recall rate is reported in Table 8. While the recall can be flexibly increased by relaxing these criteria, our primary goal is to generate high-confidence samples rather than to maximize recall. Since training the LISA on a higher-recall dataset led to degraded performance on the MIMIC-ILS test set (gIoU: 54.3%, cIoU: 61.4%, N-Acc: 95.8%), we opted for a high-threshold setting.

Table 8. Recall comparison for different threshold values.

Lesion	**atel.**	**pneu.**	**effu.**	**opac.**	**edem.**	**cons.**	**Avg.**
Recall (default τ)	13.4	28.5	16.0	21.6	58.0	31.9	28.3
Recall (half τ)	28.7	49.8	32.6	40.3	82.1	50.6	47.3

B. Lesion Types

We construct our dataset around seven major lesion types commonly observed in CXRs, identified through discussions with board-certified physicians: opacity, consolidation, pneumonia, atelectasis, edema, cardiomegaly, and effusion. These disease categories are widely utilized in many CXR-related studies [10, 17, 43]. First, we include *opacity* and *consolidation*, which are high-level, comprehensive terms referring to hyperintense lesions. These broad categories can be mapped to specific lung lesion types, including *pneumonia*, *atelectasis*, and *edema*. We also include two major non-lung disease categories: *cardiomegaly*, the enlargement of the heart, and *(pleural) effusion*, the accumulation of fluid in the pleural space.

Prompt Template for Report Structuring

Given a chest X-ray report, extract lesion information in a structured format.

Information definitions & labeling rules

Entity
• Indicates a clinical entity (disease/finding) in the radiograph.
• Only separate location when it's a general anatomical descriptor that can apply to multiple entity types.
• Examples: "pneumothorax", "consolidation", "pleural effusion", "pleural thickening", "bronchovascular markings"

Sentence Index
• Indicates the index of the sentence in the report section.
• Examples: "1", "2", "3", "4", "5", "6", "7", "8"

Presence
• Indicates whether a clinical entity is present or absent in the radiograph.
• Positive:
  • The entity (disease/finding) is present.
  • Examples: "Pleural effusion has improved", "Consolidation is stable", "No change in pneumothorax"
• Negative:
  • The entity (disease/finding) is explicitly stated to be absent or resolved.
  • Examples: "No pleural effusion", "Pneumothorax has been resolved"
  • Extract only when there is a specific disease/finding name being explicitly stated as absent (e.g., "No pneumothorax", "No pleural effusion")
  • Do not extract general descriptive terms like "clear", "unremarkable", "normal", "within normal limits"

Certainty
• Indicates the level of certainty expressed regarding the presence or absence of a clinical entity.
• Definitive:
  • The statement conveys a clear and confident assertion about the presence or absence of an entity.
  • Examples: "No pneumothorax", "Definite consolidation"
• Tentative:
  • The statement conveys uncertainty, possibility, or a lack of definitive conclusion.
  • Examples: "Possible pneumonia", "Suggests effusion", "Cannot exclude pneumothorax"

Location
• Position of findings. If multiple locations, include all in the tuple.
• If an anatomical location is embedded in the entity phrase (e.g., "hilar adenopathy", "retrocardiac opacity"), extract the location into the Location field and remove it from the Entity field. For example, "hilar adenopathy" → Entity: "adenopathy", Location: "hilar".
• Do not duplicate anatomical location terms across both Entity and Location. The Entity must be free of location descriptors.
• Examples: "right lower lobe", "bilateral", "left upper and middle lobe"

Predicted Lesion Type
• When the current entity is a non-specific finding like opacity or consolidation, and the sentence suggests what specific disease/condition it represents
• The entity that is inferred from non-specific findings like opacity or consolidation
• Examples: "opacity reflects pneumonia"

Extraction rules
1. Extract tuples of (Entity, Sentence Index, Presence, Certainty, Location, Predicted Lesion Type) for disease/findings only.
2. Do **not** extract **medical devices** (e.g., "endotracheal tube", "central line", "pacemaker") as entities. Only extract **diseases or findings**.
3. Each entity should be assigned exactly one status (Positive or Negative) and one certainty (Definitive or Tentative).
4. If there is multiple locations for an entity, include all in the Location field.
5. If there is no Location or Predicted Lesion Type, set them to None.
6. If there is no disease/finding in the report section, return None.

Figure 9. A prompt template for report structuring.

Prompt Template for Location Mapping Step1

Given an anatomical location term and its associated entity from a chest X-ray report, map it to the most appropriate category from the predefined list. If the anatomical location is 'no specific location information', use only the entity to determine the appropriate category. If the term doesn't clearly correspond to any category, return "none".

Predefined Categories
- thoracic spine (e.g., vertebrae t1)
- cervical spine (e.g., vertebrae c1)
- lumbar spine (e.g., vertebrae l1)
- clavicle (e.g., left clavicle)
- scapula (e.g., right scapula)
- rib (e.g., right posterior 7th rib)
- sternum (e.g., lower sternum)
- diaphragm (e.g., left diaphragm)
- mediastinum (e.g., esophagus, cardiomediastinum, upper mediastinum, anterior mediastinum)
- abdomen (e.g., stomach, small bowel, duodenum, liver, pancreas, left kidney)
- heart (e.g., left heart atrium, heart myocardium)
- breast (e.g., right breast)
- trachea (e.g., tracheal bifurcation)
- vessels (e.g., ascending aorta, aortic arch, pulmonary artery, inferior vena cava, pulmonary vessels, vasculature)
- lung (e.g., left lung, left, right, bilateral, hilar, costophrenic angle, lingular)
- pleura (e.g., right pleural, left pleural, pleural)
- lateral view location (e.g., middle mediastinum, retrocardiac space, retrosternal space)

Mapping Rules
1. Match the input term to the most anatomically appropriate category based on standard chest X-ray interpretation.
2. Consider synonyms and commonly used anatomical variants (e.g., "cardiac" → heart, "pulmonary" → lung, "vasculature" → vessels).
3. If the input term is simply "left", "right", or "bilateral" without further specification, assume it refers to lung and map to "lung".
4. Any anatomical term that refers to locations visible primarily or exclusively in lateral view X-rays MUST ALWAYS be mapped to "lateral view location" category.
5. If the term is unrelated to chest anatomy or clearly doesn't fit any listed category, return None.

Output Format
Return only the mapped category name or None.

Prompt Template for Location Mapping Step2 (Lung)

Given an anatomical location term and its associated entity from a chest X-ray report, map it to the most appropriate category from the predefined list. If the anatomical location is 'no specific location information', use only the entity to determine the appropriate category. If the term doesn't clearly correspond to any category, return "none".

Predefined Categories
right upper zone lung, right mid zone lung, right lung base, right apical zone lung, left upper zone lung, left mid zone lung, left lung base, left apical zone lung, lung lower lobe left, lung upper lobe left, lung lower lobe right, lung middle lobe right, lung upper lobe right, right lung, left lung

Mapping Rules
1. Match the input term to the most anatomically appropriate category based on standard chest X-ray interpretation.
2. If the term overlaps multiple categories, choose multiple categories.
3. Avoid selecting overlapping categories (e.g., choose "pleural" instead of "right pleural, left pleural, pleural").
4. If there is no relevant category, return None.

Output Format
Return only the mapped category name or None.

Figure 10. A prompt template for location mapping. Step 2 illustrates the scenario following the ‘lung’ mapping from Step 1.

C. Instruction-Answer Pair Generation

C.1. Positive Instruction

Basic Instruction. The instructions for positive samples are generated directly from the grounded lesion mask generation results. For instance, if a definitive finding of pneumonia has a grounded location of right lung base and left lung base, we generate a basic instruction: "Segment the pneumonia in the right lung base and left lung base.". However, if the finding's certainty is tentative, indicating the lesion's presence is not definitive, we substitute it with the more general term opacity to create the basic instruction. For example, the previous instruction becomes: "Segment the opacity in the right lung base and left lung base.". As listed in Table 9, the target location can be specified as a single area—either a broad region or a specific lung zone—or as a combination of these areas.

Table 9. List of valid target locations for basic instructions. Locations are categorized into broad regions and specific lung zones.

Lung Region	Location Name
Broad Regions	right lung left lung
Lung Zones (Right)	right apical zone lung right upper zone lung right mid zone lung right lung base
Lung Zones (Left)	left apical zone lung left upper zone lung left mid zone lung left lung base

Global Instruction. A global instruction is used to segment all instances of a lesion across the entire lung, without specifying a location. To create a valid global instruction, the generated lesion must cover all lesions cited in the report; this ensures the mask can serve as a complete ground truth. Therefore, we only generate global instructions when the *grounded location*, where masks were actually generated, and the *reported location*, the complete area mentioned in the report, are identical. If the lesion type is cardiomegaly, we always generate this instruction type. This is because cardiomegaly represents a condition of the heart itself, rather than a lesion that can appear in variable locations.

Lesion Inference Instruction. We generate lesion inference instructions to enable the model to infer the specific lesion type from an opacity at a given location. These instructions are generated for findings regardless of their original certainty level, as the certainty is instead reflected in the ground-truth text description. We selected pneumonia, atelectasis, and edema as the target lesion types for this task. This choice reflects clinical reporting practices, where radiologists often describe these specific findings using an inferential process. In contrast, other major lesion types are typically stated directly. For example, a report rarely states, "There is an opacity in the left lung. It is highly suggestive of effusion."; instead, the finding is stated directly as "Left lung effusion."

C.2. Negative Instruction

Negative instructions are generated in two main scenarios. First, we generate instructions for lesions that are either never mentioned or negated (*e.g.*, "no pneumonia") in the report. For these findings, we create a negative instruction, which can be either a basic type by randomly assigning a lung region (*e.g.*, "Segment the pneumonia in the left lung."), or a global type (*e.g.*, "Segment the pneumonia"). The second method involves pairing a target lesion with a randomly selected empty location. For example, if right lung apex and left lung apex are empty locations, we can generate the instruction, "Segment the atelectasis in the right lung apex." To prevent an excessive number of negative samples, our logic restricts the generation to a maximum of one negative instruction per lesion type for each study.

C.3. Clinical Utility of MIMIC-ILS

Diversity of Instructions. The functional scope of the dataset is driven primarily by the diversity of disease–anatomy combinations rather than by the number of instruction templates. Linguistic diversity can be readily addressed by paraphrasing existing instructions in MIMIC-ILS. To demonstrate this feasibility, we used Qwen3-Next-80B-A3B-Instruct to paraphrase each original instruction into nine variants reflecting three user personas (medical experts, laypersons, and AI developers). LISA trained on this enriched dataset still demonstrate strong performance (gIoU: 67.3%, cIoU: 73.1%, N-Acc: 96.5%) on the paraphrased test set.

Usability for Laypersons. Users without medical expertise cannot be expected to visually identify lesions in a scan to provide basic instructions. However, because MIMIC-ILS incorporates negative cases for absence confirmation, a model trained on this dataset naturally overcomes this limitation. By iteratively querying the model across various anatomical locations, the system can autonomously verify the presence of a lesion—outputting a precise segmentation mask if it exists, or confirming its absence otherwise. Similarly, global instructions (e.g., "Segment the opacity") offer an intuitive way for users to make broad inquiries. Coupled with the model's robust handling of negative cases, these capabilities ensure that users can effectively obtain screening results without ever needing to visually inspect the image themselves.

D. Quality Control

D.1. Chest X-rays

We exclusively utilized Posteroanterior (PA) and Anteroposterior (AP) view images from the MIMIC-CXR dataset. However, even within these designated views, the dataset contains noisy samples, including mislabeled lateral views, non-chest X-rays, or images with severe anatomical truncation. To ensure data quality, we leveraged metadata from CXReasonBench [26, 27]. This dataset was meticulously constructed from frontal view images within the MIMIC-CXR dataset that had verified high image quality. Specifically, we utilized its pre-extracted information such as the count of extractable CXAS anatomy masks and indicators of full chest visibility to identify and exclude these problematic images beforehand.

D.2. Lung and Heart Masks

Lung and heart masks are a critical component for the construction of MIMIC-ILS. However, both models can produce erroneous results: the pretrained HybridGNet occasionally generates abnormal masks, and CXAS (even when applied to RadEdit-processed images) also generates suboptimal masks. To address this, we cross-referenced the masks from both models and excluded cases with significant discrepancies, interpreting this as a failure in either the HybridGNet or CXAS segmentation. Specifically, we determined that large differences in the outermost x-coordinates of the lung masks or the lowermost y-coordinates of the heart masks would cause problems for subsequent grounded lesion mask generation, and thus excluded these studies.

D.3. Cardiomegaly

To generate reliable negative samples for cardiomegaly, we measured the cardiothoracic ratio (CTR) using the right lung, left lung, and heart masks generated by the pretrained HybridGNet. We then filtered these samples, exclusively including those with a CTR of 0.45 or less in our final negative dataset. This 0.45 threshold was calibrated by a physician who analyzed the distribution of CXRs across different CTR intervals to establish a clinically sound cutoff.

E. MIMIC-ILS Dataset

E.1. Details for Dataset Splits

The data splits and distribution by instruction type for MIMIC-ILS are presented in Tables 10, 11, and 12. Note that the counts for the test set reflect the final numbers after excluding cases that were rejected during the quality assessment process.

E.2. Quality Assessment

A rigorous quality assessment was conducted on the test split by four physicians. All positive samples were reviewed by all four physicians, while the negative samples were divided among them for evaluation. The reviewers were provided with an CXR image, the lesion type, and the mapped anatomical location text generated from our information-grounding process, along with the corresponding ground-truth radiology report. They were then asked to mark each pair as either "Acceptable" or "Not Acceptable" on a review sheet. The evaluation process was conducted independently for each expert, ensuring that no reviewer could access the others' evaluation results.

E.3. Details on Expert Evaluation

The expert evaluations were conducted by four physicians, all experienced radiation oncologists with extensive training in lesion contouring. Their professional backgrounds are as follows: Experts A and B are board-certified physicians with 9 and 7 years of clinical experience, respectively, while Experts C and D are resident doctors, each with 6 years of clinical experience. Also, the lesion-level acceptance rate in human evaluation are shown in Table 13.

E.4. Data Generation Examples

With our proposed data generation pipeline, we can produce high-quality lesion masks and their corresponding instruction–answer pairs from grounded information. Figure 11 illustrates representative examples across various lesion types and anatomical locations, including both positive and negative cases, along with their corresponding structured information.

Table 10. Number of generated instruction-answer pairs per lesion and template type in MIMIC-ILS train split.

Lesion	# IAs	Basic		Global		Lesion Inference	
		pos	neg	pos	neg	pos	neg
cardiomegaly	63,153	0	0	39,108	24,045	0	0
pneumonia	158,059	4,542	145,317	511	3,147	4,542	0
atelectasis	166,935	8,846	128,943	3,331	16,969	8,846	0
opacity	156,807	9,113	73,619	1,532	274	0	72,269
consolidation	154,955	3,428	144,489	379	6,659	0	0
edema	182,233	14,150	145,251	5,390	3,292	14,150	0
effusion	162,998	10,244	114,375	3,713	34,666	0	0
Total	1,045,140	50,323	751,994	53,964	89,052	27,538	72,269

Table 11. Number of generated instruction-answer pairs per lesion and template type in MIMIC-ILS validation split.

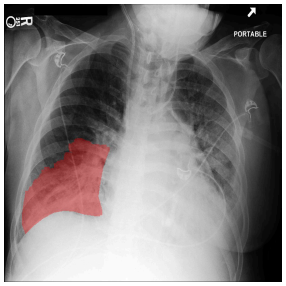
Lesion	# IAs	Basic		Global		Lesion Inference	
		pos	neg	pos	neg	pos	neg
cardiomegaly	539	0	0	332	207	0	0
pneumonia	1,211	28	1,130	2	23	28	0
atelectasis	1,316	75	986	33	147	75	0
opacity	1,225	75	581	13	0	0	556
consolidation	1,213	30	1,137	3	43	0	0
edema	1,469	129	1,124	59	28	129	0
effusion	1,273	416	885	22	287	0	0
Total	8,246	416	5,843	464	735	232	556

Table 12. Number of generated instruction-answer pairs per lesion and template type in MIMIC-ILS test split.

Lesion	# IAs	Basic		Global		Lesion Inference	
		pos	neg	pos	neg	pos	neg
cardiomegaly	965	0	0	803	162	0	0
pneumonia	1,767	60	1,596	8	43	60	0
atelectasis	1,842	110	1,466	45	111	110	0
opacity	1,753	174	779	26	5	0	769
consolidation	1,756	69	1,612	10	65	0	0
edema	2,274	283	1,551	103	54	283	0
effusion	1,878	156	1,312	54	356	0	0
Total	12,235	852	8,316	1,049	796	453	769

Table 13. Acceptance rate (%) by lesion type for each expert in the human evaluation.

Lesion	Expert A			Expert B			Expert C			Expert D		
	Pos	Neg	Total	Pos	Neg	Total	Pos	Neg	Total	Pos	Neg	Total
Cardiomegaly	97.7	97.1	97.7	99.2	100.0	99.2	100.0	100.0	100.0	99.4	100.0	99.4
Pneumonia	90.0	98.5	97.3	97.1	99.7	99.4	98.6	99.5	99.4	98.6	98.8	98.7
Atelectasis	97.2	98.8	98.4	79.9	99.5	94.2	100.0	99.0	99.3	99.3	97.3	97.8
Opacity	92.6	92.3	92.4	96.5	96.5	95.1	99.5	92.7	96.0	96.0	96.6	96.3
Consolidation	97.2	95.7	95.9	100.0	97.2	97.6	100.0	98.3	98.5	100.0	99.0	99.2
Edema	95.8	95.1	95.4	97.9	100.0	99.0	100.0	97.3	98.5	89.8	98.2	94.4
Effusion	89.8	96.6	94.1	89.3	97.3	94.3	99.5	97.6	98.3	95.4	98.8	97.6

section_id: sxxxxxxxx_findings

dicom_id: xxxxxxxx-xxxxxxxx-xxxxxxxx-xxxxxxxx-xxxxxxxx

report: "... Bibasilar pulmonary opacities are increasing from the prior examination done yesterday and are likely related to increasing atelectasis."

target: atelectasis

certainty: Tentative

grounded_location: right lung base

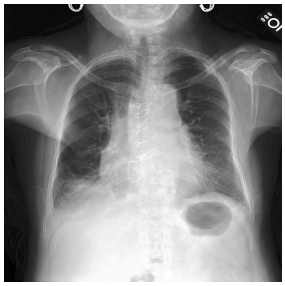
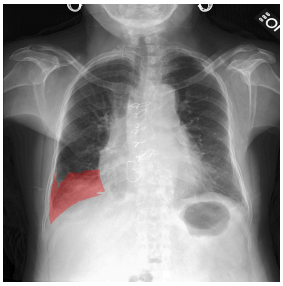
reported_location: left lung base, right lung base

instruction: ["Segment the opacity in the right lung base.",
"Segment the opacity in the right lung base and predict its type."]

answer: ["[SEG]", "[SEG] It possibly reflects atelectasis."]

seg: true

seg_mask_path: sxxxxxxxx_findings/atelectasis_lesion_mask.png

section_id: sxxxxxxxx_findings

dicom_id: xxxxxxxx-xxxxxxxx-xxxxxxxx-xxxxxxxx-xxxxxxxx

report: "As compared to the previous radiograph, there is unchanged evidence of a small right pleural effusion. ..."

target: effusion

certainty: Definitive

grounded_location: right lung base

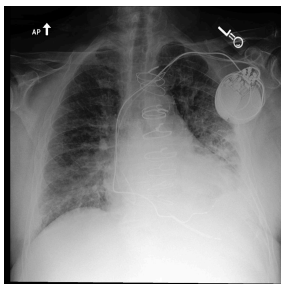
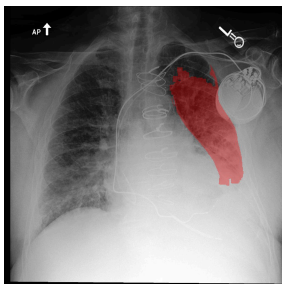
reported_location: right lung base

instruction: ["Segment the effusion.",
"Segment the effusion in the right lung base."]

answer: ["[SEG] It is located in the right lung base.", "[SEG]"]

seg: true

seg_mask_path: sxxxxxxxx_findings/effusion_lesion_mask.png

section_id: sxxxxxxxx_impression

dicom_id: xxxxxxxx-xxxxxxxx-xxxxxxxx-xxxxxxxx-xxxxxxxx

report: "Moderately severe interstitial pulmonary edema has worsened accompanied by new or increased small left pleural effusion. ..."

target: edema

certainty: Definitive

grounded_location: left lung

reported_location: right lung, left lung

instruction: ["Segment the edema in the left lung.",
"Segment the opacity in the left lung and predict its type."]

answer: ["[SEG]", "[SEG] It is highly suggestive of edema."]

seg: true

seg_mask_path: sxxxxxxxx_findings/edema_lesion_mask.png


section_id: sxxxxxxxx_findings

dicom_id: xxxxxxxx-xxxxxxxx-xxxxxxxx-xxxxxxxx-xxxxxxxx

report: "... There is no effusion, or overt signs of CHF. ..."

target: effusion, opacity, atelectasis

certainty: none

grounded_location: none

reported_location: none

instruction: ["Segment the effusion in the right lung.",
"Segment the opacity in the right lung base.",
"Segment the atelectasis in the left mid zone lung."]

answer: ["[SEG] There is no effusion in the right lung.",
"[SEG] There is no opacity in the right lung base.",
"[SEG] There is no atelectasis in the left mid zone lung."]

seg: false

seg_mask_path: none

Figure 11. Examples of final generated samples in our MIMIC-ILS dataset.

F. Model Training Details

In our experiments, training the LISA-7B–based model took approximately two and a half days on two NVIDIA H100 GPUs for 15 epochs. Using the DeepSpeed package [39], we trained the model with the DeepSpeed Stage-2 configuration and a WarmupDecayLR scheduler, with 100 warmup steps and a minimum and maximum learning rate of 0 and 0.0003, respectively. For inference on the test set, which contains 12K examples, segmentation alone takes about 20 minutes, whereas segmentation with text outputs requires approximately 1.5 hours. During training, each input image had a 50% chance of being processed with histogram equalization.

G. Additional Experimental Results

G.1. Lesion-Wise Text Accuracy

The text-response accuracy of ROSALIA for each lesion type is summarized in Table 14. Across most lesion and question types, the model consistently achieves high accuracy, similar to the segmentation performance reported in Table 5. For lesion-inference questions, CXR alone typically cannot provide a definitive diagnosis and often requires additional examinations (e.g., blood tests or cultures). As a result, radiologists generally provide only a differential diagnosis based on visual findings. Given this inherent uncertainty in CXR interpretation, improvements in accuracy for lesion-inference questions are naturally limited. Nevertheless, we evaluated how well the trained model on our dataset can perform on this question type and leave further advancements in this direction to future work.

Table 14. Text response accuracy (%) of ROSALIA across different question and lesion types.

Lesion	Overall	Basic	Global	Lesion Inf.
Cardiomegaly	96.0	-	96.0	-
Pneumonia	96.3	99.2	72.6	36.7
Atelectasis	92.9	96.9	69.2	69.1
Opacity	91.5	92.2	93.6	90.5
Consolidation	97.3	97.6	92.0	-
Edema	93.4	96.2	74.5	85.5
Effusion	94.5	96.9	85.9	-
Total	**94.4**	**96.8**	**88.8**	**84.8**

G.2. Additional Qualitative Examples

We present additional qualitative examples comparing ROSALIA with baseline models in Figure 12. Unlike the baselines, ROSALIA produces accurate segmentation outputs tailored to diverse user instructions. In addition, examples that include both text responses and segmentation outputs are shown in Figure 13. In these cases as well, ROSALIA provides highly factual text responses alongside precise segmentation results.

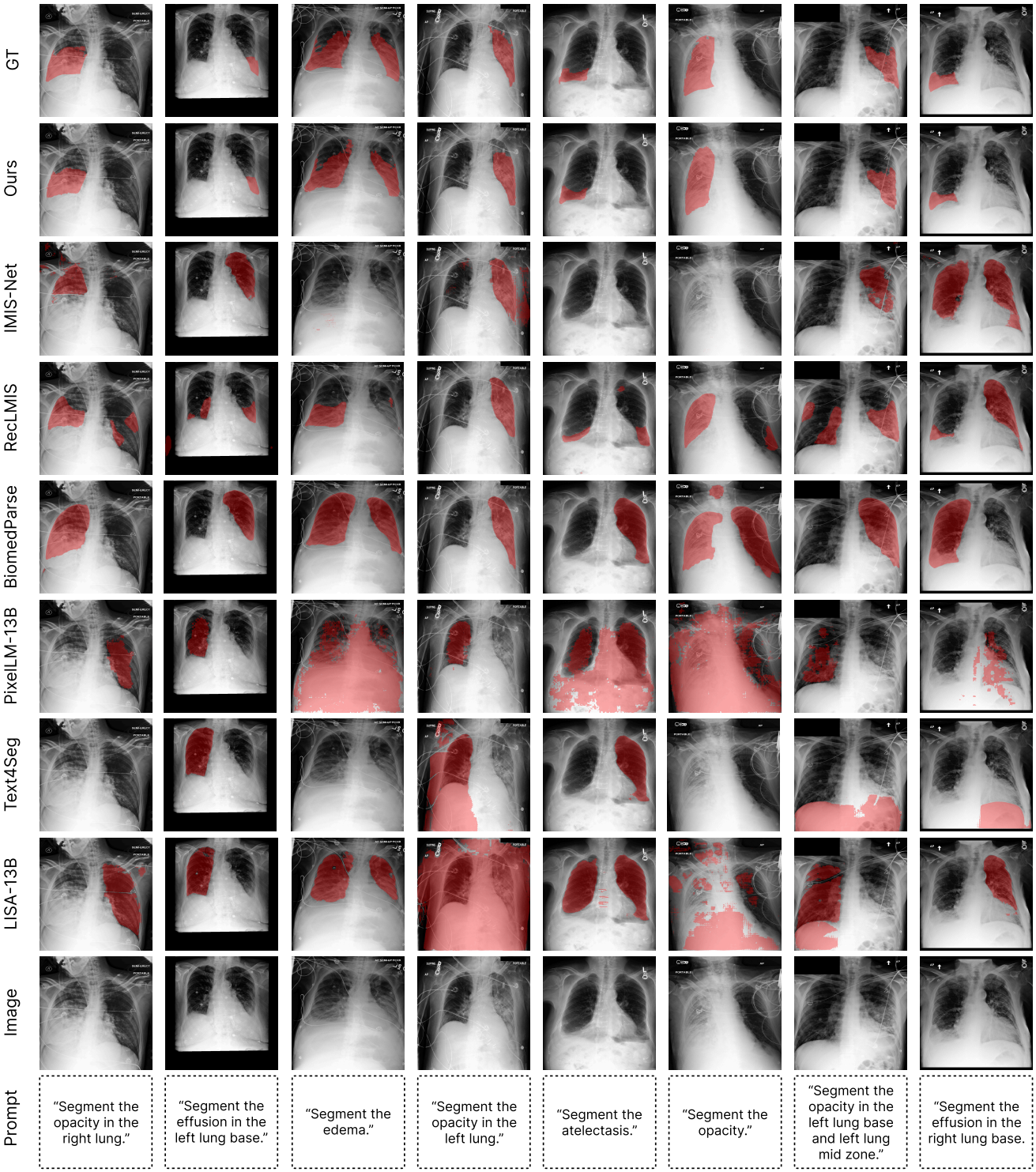


Figure 12. Qualitative comparison of segmentation results between ROSALIA and baseline models.

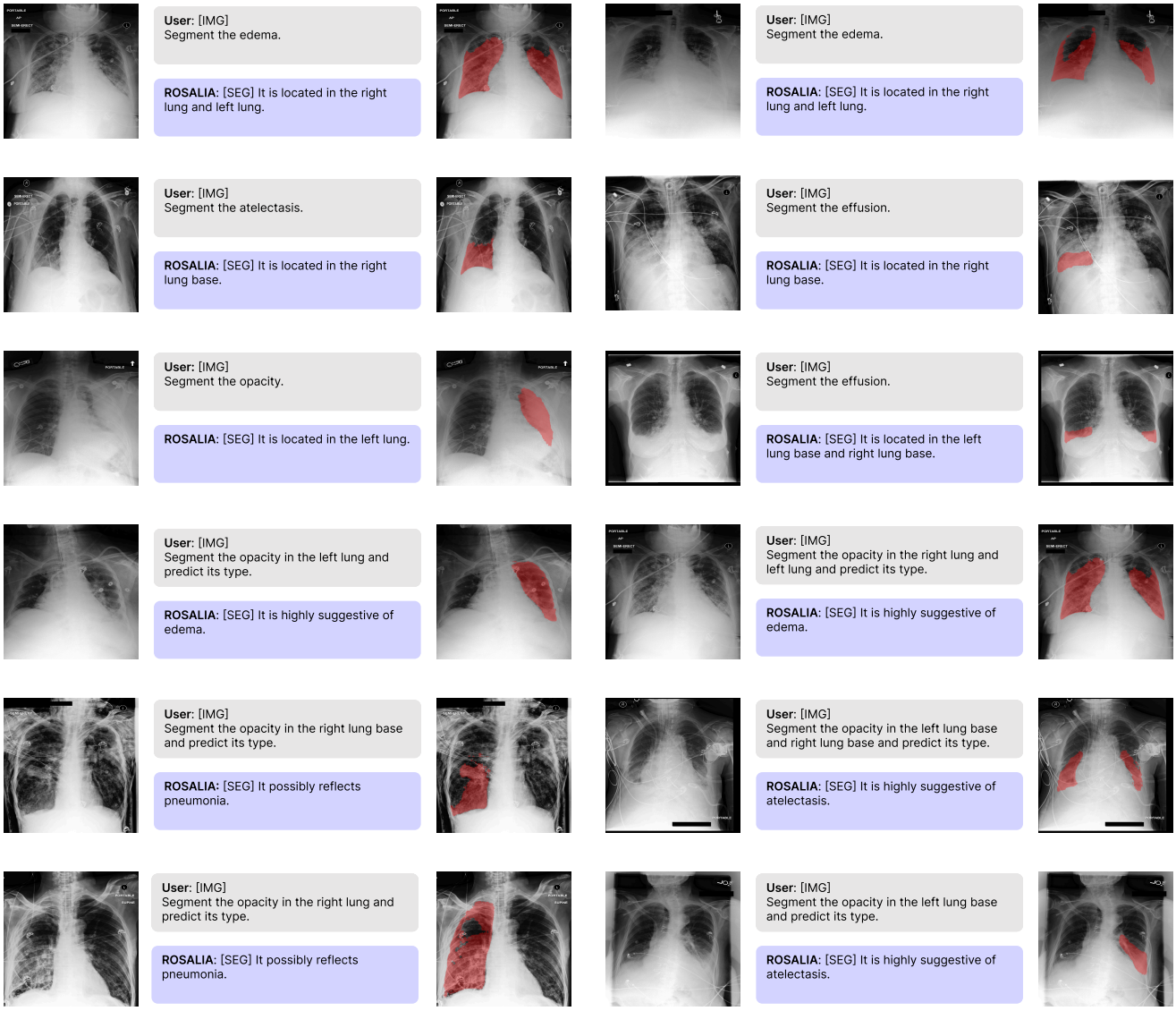


Figure 13. Examples of textual responses generated by ROSALIA. All generated text responses correctly match the ground-truth answers, and both the segmentation and textual outputs in this figure are rated as good examples by medical experts.