Instruction-Guided Lesion Segmentation for Chest X-rays with Automatically Generated Large-Scale Dataset

Anonymous CVPR submission

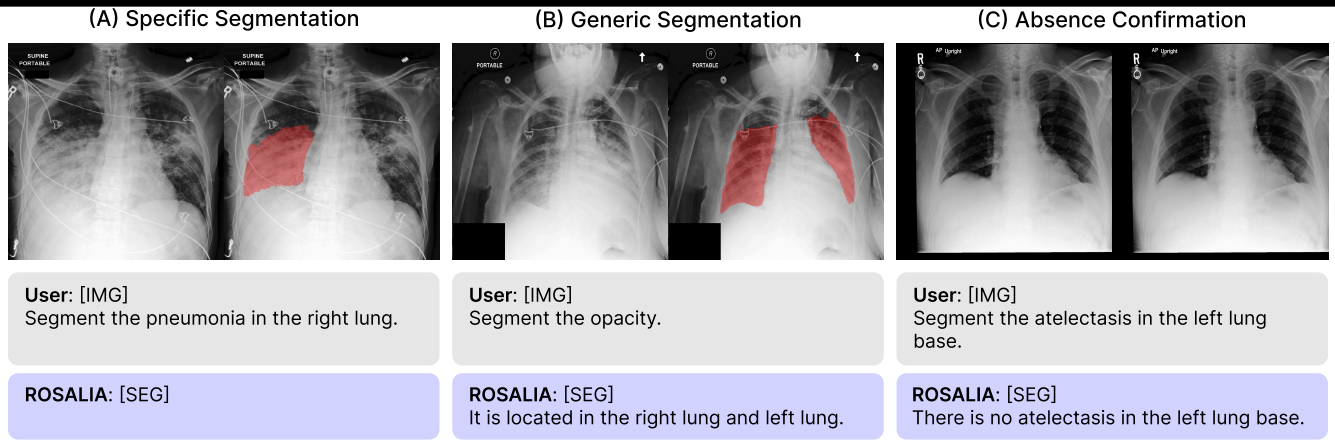

Figure 1. Examples of the instruction-guided CXR lesion segmentation task. Given text instructions for various lesion types and locations of interest, ROSALIA, a VLM trained on our MIMIC-ILS dataset, can: (A) segment lesions in a specified location, (B) segment lesions globally, and (C) detect empty-target cases. As can be seen in (A), ROSALIA correctly ignores the unrequested lesion in the left lung.

Abstract

The applicability of current lesion segmentation models for chest X-rays (CXRs) has been limited both by a small number of target labels and the reliance on long, detailed expert-level text inputs, creating a barrier to practical use. To address these limitations, we introduce a new paradigm: instruction-guided lesion segmentation (ILS), which is designed to segment diverse lesion types based on simple, user-friendly instructions. Under this paradigm, we construct MIMIC-ILS, the first large-scale instruction-answer dataset for CXR lesion segmentation, using our fully automated multimodal pipeline that generates annotations from chest x-ray images and their corresponding reports. MIMIC-ILS contains 1.1M instruction-answer pairs derived from 192K images and 91K unique segmentation masks, covering seven major lesion types. To empirically demonstrate its utility, we introduce ROSALIA, a vision-language model fine-tuned on MIMIC-ILS. ROSALIA can segment diverse lesions and provide textual explanations in response to user instructions. The model achieves high segmentation and textual accuracy in our newly proposed task, highlighting the effectiveness of our pipeline and the value of MIMIC-ILS as a foundational resource for pixel-level CXR lesion grounding.

1. Introduction

Medical imaging is an essential technique in modern medicine, enabling accurate diagnosis and appropriate treatment. Among various imaging modalities, chest X-ray (CXR) is one of the most common examinations due to its high accessibility and rapid acquisition [4]. Radiologists reach a diagnosis by integrating visual evidence from CXRs with their clinical knowledge, and describe these findings in a text format known as a *radiology report*. A key step in this diagnostic process is identifying the precise location and boundary of a *lesion*—an abnormal region with pathological changes [6]. This task is labor-intensive and demands substantial clinical expertise and analytical precision.

To alleviate physicians' workload in localizing pathological regions, there is a growing demand for automated lesion segmentation models in CXRs. Recently, vision–language models (VLMs) equipped with segmentation modules [22, 23, 35] have emerged as a promising solution, as they can interpret diverse user-specific needs expressed through natural language instructions. However, despite the success of such VLMs in general-domain segmentation, their application to CXRs remains limited. Although prior studies [15, 25] have explored CXR lesion segmentation using text prompts, they

Table 1. Existing CXR datasets with spatial annotations for pathologic lesions.

Dataset	# Images	Spatial Annotation				Instruction-Answer Pair
		# Annotations	Type	Multi-Lesion	Method	
VinDr-CXR [33]	15K	9K	Bounding Box	✓	Manual	✗
Padchest-GR [8]	4.6K	7.7K	Bounding Box	✓	Manual	✗
MS-CXR [3]	1K	1.2K	Bounding Box	✓	Manual	✗
TBX-11K [28]	12K	1.2K	Bounding Box	✗	Manual	✗
SIIM-ACR [41]	13K	2.7K	Segmentation Mask	✗	Manual	✗
QaTa-COV19 [9]	121K	9.3K	Segmentation Mask	✗	Semi-Automated	✗
Danilov et al. [7]	1.4K	0.6K	Segmentation Mask	✓	Manual	✗
MIMIC-ILS (Ours)	192K	91K	Segmentation Mask	✓	Fully-Automated	✓

are limited to a single lesion type (*e.g.*, COVID-19) and moreover require long, detailed expert-level medical descriptions based on tailored CXR review (*e.g.*, "Bilateral pulmonary infection, two infected areas, upper right lung and upper left lung.") as input. Such constraints make them impractical not only for physicians who aim to segment diverse lesion types across various anatomical subregions before closely reviewing the image themselves, but especially for non-experts who can hardly interpret CXR images at all.

To address these limitations, we propose a more user-friendly paradigm, namely *instruction-guided lesion segmentation* (ILS). In this paradigm, the model is required to process diverse user instructions, ranging from prompts that specify the lesion type and target location, to requests that look for abnormalities globally. If the requested lesion is not present, the model should reliably report its absence. Additionally, the model should be able to provide textual descriptions regarding a lesion's location or type, even if not explicitly prompted by the user. However, a dataset to support such a versatile task has been unavailable, as constructing a suitable dataset for training and evaluation poses significant challenges—most notably the need for expert-curated mask annotations. Moreover, accurately pairing these masks with precise textual instructions in terms of anatomical locations and specific lesion types remains a highly complex task.

In this work, we introduce the first fully automated pipeline for constructing a large-scale ILS dataset for CXRs. The central challenge is: "How can we derive lesion masks and corresponding instruction-answer text pairs from raw images that contain no explicit annotations?" To address this, we leverage radiology reports as a key source of information for each image. Using paired image–report data, our two-stage pipeline integrates pre-trained vision models and large language models (LLMs) to extract high-confidence anomalous regions and structured textual information. By exploiting the consistency between these heterogeneous modalities, we generate high-quality lesion masks and diverse instruction–answer pairs. Applying our novel framework to MIMIC-CXR [19, 20]—a large, publicly available CXR–report dataset—we constructed **MIMIC-ILS**, a large-scale dataset consisting of 1.1M samples derived from 192K images and 91K lesion masks (Table 1).

Although several datasets [3, 7–9, 28, 33, 41] have tried to introduce spatial annotations in the CXR domain, they are unsuitable for direct use in our ILS paradigm (Table 1). Most provide only coarse bounding-box localization or single lesion type masks that are limited in scale due to reliance on expert annotations. Moreover, they also lack explicit links between mask annotations and textual instructions. MIMIC-ILS bridges these gaps by offering large-scale instruction–answer pairs, each paired with an auto-labeled segmentation mask and a detailed lesion profile. Despite being constructed entirely without human intervention, expert evaluations report a high acceptance rate of over 95% for this dataset.

Leveraging MIMIC-ILS, we train **ROSALIA** (**R**adi**O**logy **S**egmentation **A**ssistant trained on a **L**esion-grounded **I**nstruction-**A**nswer dataset), the first VLM designed for ILS in CXRs. Given user instructions, ROSALIA generates segmentation masks and textual descriptions (Fig. 1), handling a wide range of tasks, such as specific segmentation (*e.g.*, "Segment the pneumonia in the right lung."), generic segmentation (*e.g.*, "Segment the opacity."), and absence confirmation (*e.g.*, "There is no atelectasis in the left lung base."). This flexibility enables ROSALIA to effectively address diverse user needs, delivering tailored outputs for each request.

In summary, our contributions are threefold:

- We introduce a novel automated pipeline that generates lesion masks and corresponding instructions directly from CXRs without any human intervention. Using only image–report pairs, our method produces a large-scale dataset without requiring explicit manual processing.
- Applying our framework to MIMIC-CXR, we construct MIMIC-ILS, the first dataset for instruction-guided lesion segmentation (ILS) in CXRs. The resulting dataset is further validated by medical experts, confirming its high quality and the reliability of the construction process.
- To validate the utility of MIMIC-ILS, we introduce ROSALIA, the first VLM designed for ILS in CXRs. Trained on our million-scale dataset, ROSALIA interprets user instructions across diverse lesion types and locations, producing accurate lesion masks and descriptive outputs. As existing general and medical VLMs significantly struggle with this task, we will publicly release our dataset and model to support advances in fine-grained CXR lesion grounding.

2. Related Work

2.1. Lesion Segmentation and Datasets

Lesion segmentation aims to generate masks corresponding to abnormal regions in medical images. Typically, models are trained on datasets where radiologists have directly annotated lesion masks. For CT and MRI, several studies [16, 17] have utilized public datasets that provide diverse tumor masks [1, 2, 13]. In contrast, such pixel-level annotations are scarce in the CXR domain. While some datasets provide only bounding boxes [8, 33], those that offer segmentation masks usually focus on a single lesion type [9, 28, 41]. Consequently, existing models trained on these datasets are limited in their effective segmentation range for CXRs [42]. Our work directly addresses this gap by constructing a comprehensive, multi-type lesion segmentation dataset for CXRs.

2.2. Language-Guided Image Segmentation

Language-guided image segmentation is the task of segmenting a target specified by text. Early approaches to this task focused on aligning image features with text labels to generate corresponding masks [24, 37, 40, 43]. More recently, advancements in VLMs have enabled researchers to extend their reasoning capabilities to segmentation [22, 23, 35]. These models can generate an appropriate mask based on complex instructions that require real-world knowledge, such as "Segment the object richest in vitamin C in this photo."

Similar research has emerged in the medical domain, but current approaches remain limited. They usually rely on simple prompts including class labels (*e.g.*, "a computerized tomography of a tumor") [5, 27], which cannot handle sentence-level instructions. In the CXR domain specifically, recent VLMs have been trained using free-form text that describes the location and number of lesions [15, 25]. These approaches, however, expect users to have already reviewed the CXR image, thus providing expert-level descriptions as input. In contrast, our model allows users to obtain the lesion mask, its presence or absence, and type information even without having to interpret the CXR images first.

3. Automatic Dataset Construction

This section outlines our approach to automatically constructing a large-scale dataset for training a model that generates both lesion segmentation masks and corresponding textual descriptions in response to user instructions. The main challenges in this process are: (1) generating lesion masks directly from raw CXR images without explicit image annotations, (2) aligning appropriate instruction–answer texts with the obtained masks, and (3) ensuring that the entire pipeline operates in a fully automated, human-free manner. To address these challenges, our framework first extracts textual and spatial information from image–report pairs and generates lesion masks followed by a verification process (Sec. 3.1). Using the verified lesion masks, we then construct diverse instruction–answer pairs (Sec. 3.2).

3.1. Grounded Lesion Mask Generation

To construct our dataset, we use MIMIC-CXR [19, 20], a large collection of CXR images paired with radiology reports. Each report is written by a radiologist and provides visual descriptions of the corresponding CXR image. Based on this dataset, we generate grounded lesion masks through four sequential steps as illustrated in Fig. 2: (1) Report structuring and location mapping; (2) Spatial information extraction; (3) Lesion mask generation; (4) Location verification. The details of each step are provided in Appendix A.

Report Structuring and Location Mapping. The first step employs LLMs to convert radiology reports into a structured form for later steps. Specifically, we instruct an LLM to transform each sentence describing an abnormal finding into a six-element tuple consisting of the following categories: entity, sentence index, presence, certainty, location, and predicted lesion type. The location element is then mapped to one or more anatomical labels to ensure compatibility with the segmentation model used in subsequent processes. For example, if the second sentence in a given radiology report is "The lower lung opacity is pneumonia.", its corresponding output is `(opacity, 2, positive, definitive, [right lung base, left lung base], pneumonia)`. Here, the term "lower lung" in the original report is mapped to "right lung base" and "left lung base".

Spatial Information Extraction. The second step extracts spatial information from CXRs using three distinct models: (1) RadEdit [34], a diffusion-based image editing model; (2) CXAS [36], an anatomy segmentation model; and (3) a pretrained YOLO model for CXR lesion detection [32]. These models are used respectively to generate an anomaly map, anatomy masks, and lesion box masks, which serve as visual cues for lesion mask generation in the subsequent steps.

RadEdit takes an input image $x \in \mathbb{R}^{H \times W}$ containing a lesion and the text prompt "No acute cardiopulmonary process" and outputs an edited image $\hat{x}$ from which the lesion has been removed. We derive $x_{\text{ano}} \in [0,1]^{H \times W}$ as:

$$x_{\text{ano}} = \frac{x - \hat{x}}{I_{\max}},$$

where $I_{\max}$ is the maximum possible pixel intensity (*i.e.*, 255 for an 8-bit image). x_{ano} provides morphological information about hyperintense lesions, which are areas that appear brighter than the normal lung field. From x_{ano}, we define anomaly map $\mathcal{A}$ as:

$$\mathcal{A} = \{(i,j) \mid (x_{\text{ano}})_{i,j} \geq \tau_{\text{ano}}\}, \quad (1)$$

where (i,j) represents a pixel coordinate and τ_{ano} is a threshold for anomaly pixels.

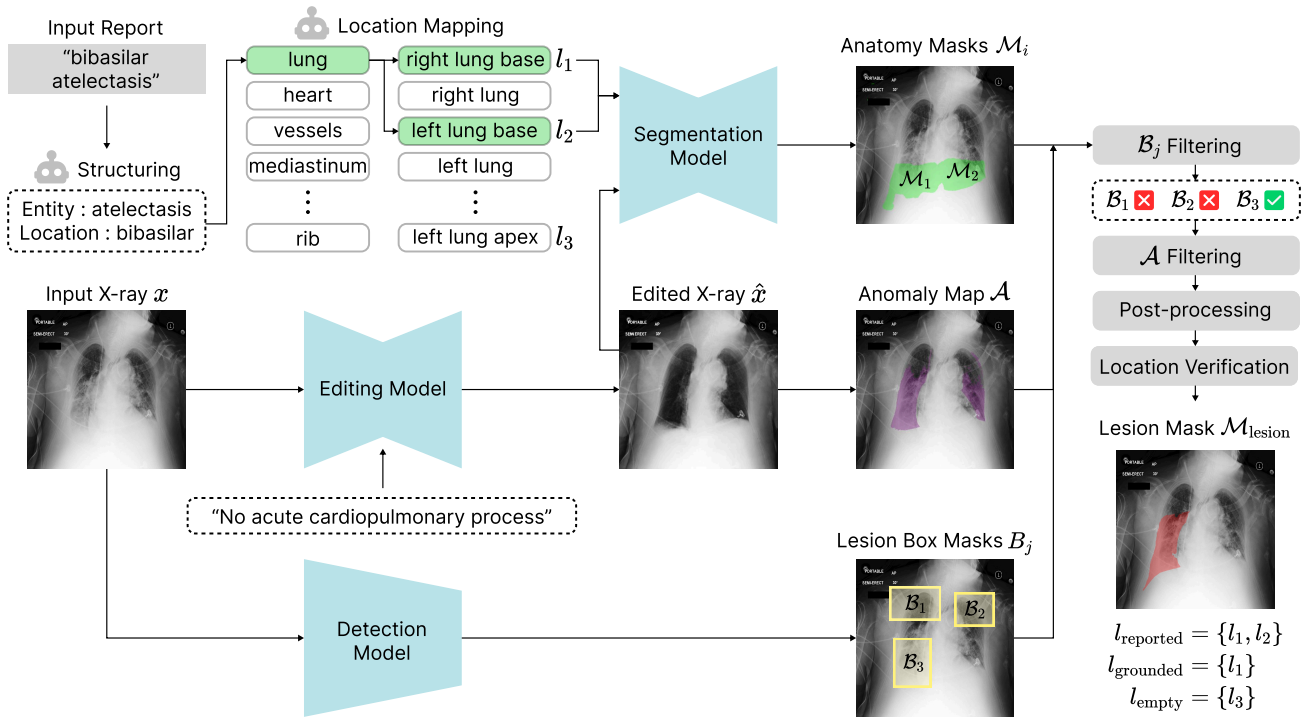

Figure 2. An overview of grounded lesion mask generation. **(Top-left)** Textual information is extracted from the radiology report during the report structuring and location mapping. **(Bottom-left and Center)** Pretrained vision models are also employed to produce spatial information. **(Right)** Finally, a lesion mask is generated by integrating this information. The verification step then confirms the grounded location (l_1), identifies the empty location (l_3) for negative sample generation, and discards the reported-but-ungrounded location (l_2).

CXAS produces anatomy masks corresponding to each location element in the previously derived structured report tuples. We denote these masks as coordinate sets $\{\mathcal{M}_i\}_{i=1}^{n}$, where n is the number of anatomical labels mapped in the previous step. Each $\mathcal{M}_i$ contains the pixel coordinates for a specific anatomy, serving as a spatial approximation of the lesion location mentioned in the radiology report.

In parallel, the pretrained YOLO model is applied to x to detect a diverse range of lesions. It outputs bounding boxes that not only specify the locations of potential lesions but also assign a confidence score to each detection. From these results, we construct a set of lesion box masks, $\{\mathcal{B}_j\}_{j=1}^{m}$, where m denotes the number of detected boxes. Each $\mathcal{B}_j$ represents the pixel coordinates enclosed by a bounding box, accompanied by a confidence score $conf_{\mathcal{B}_j} \in [0, 1]$.

Lesion Mask Generation. With the three visual cues extracted from the previous step, the initial lesion masks can be generated. Here, the anomaly map $\mathcal{A}$ plays a central role, representing a composite signal of all hyperintense lesions. We decompose this signal into individual masks and align them with the specific lesions described in the report. During this process, the anatomy masks $\{\mathcal{M}_i\}_{i=1}^{n}$, lesion box masks $\{\mathcal{B}_j\}_{j=1}^{m}$, and the right and left lung masks (L_r and L_l) are jointly used to select high-quality mask candidates.

The core of this filtering process, outlined in Algorithm 1, selectively retains only appropriate candidates from the initially detected lesion box masks, based on four conditions: (c_1) sufficient overlap with $\{\mathcal{M}_i\}_{i=1}^{n}$; ($c_2$) a high confidence score; (c_3) a high internal signal ratio from $\mathcal{A}$ (*i.e.*, the ratio of the intersection area between the box mask and $\mathcal{A}$ to the area of the box mask); and (c_4) a sufficient size relative to either L_r or L_l. Conditions c_1 and c_2 ensure that the boxes align with the reported locations and are likely to correspond to true lesions. However, the $\mathcal{A}$ can contain false negatives (*i.e.*, coordinates that belong to actual lesion areas but are missing from $\mathcal{A}$), which may result in excessively small or even empty masks. To mitigate this issue, conditions c_3 and c_4 are used to retain only those boxes that contain strong lesion signals and are large enough to allow meaningful segmentation. Once the appropriate lesion box masks are selected, we extract from $\mathcal{A}$ the connected components (*i.e.*, the individual, contiguous 'islands' in 2D space) that intersect with these selected masks. This component then undergoes a post-processing step involving small, noisy mask removal to produce the final, refined lesion mask $\mathcal{M}_{\text{lesion}}$ (see Appendix A.5 for further details).

Location Verification. In the final step, we explicitly verify whether each lesion mask generated by Algorithm 1 has been successfully grounded to the structured report. To assess the grounding status, we define three types of locations: *reported location*, *grounded location* and *empty location*. The *reported location* is a set of anatomical labels extracted from the previous location mapping with LLMs. Based on this set, the *grounded location* is defined as a subset of the *reported*

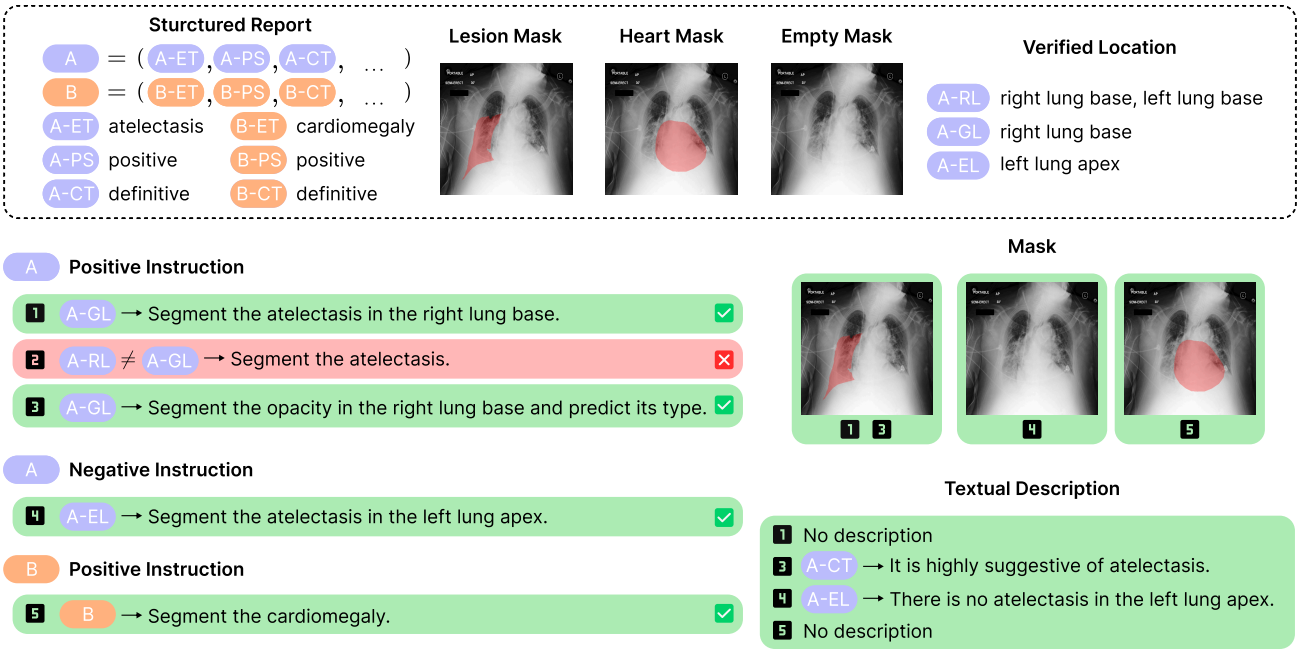

Figure 3. Instruction–answer pair generation process using the example report, "Bibasilar atelectasis. Cardiomegaly." We utilize the elements extracted from the previous lesion mask generation process (see Fig. 2), indicated by the dashed box. Structured tuples (A&B in the top left) are converted to text instructions and mapped to their corresponding ground-truth masks and textual descriptions. Invalid instructions for lesions which lack a corresponding mask are excluded (colored as red), and only valid instructions are retained (colored as green). (ET: entity, PS: presence, CT: certainty, RL: reported location, GL: grounded location, EL: empty location)

Algorithm 1: Lesion Mask Generation

Input: Anomaly map $\mathcal{A}$, anatomy masks $\{\mathcal{M}_i\}_{i=1}^{n}$, lesion box masks $\{\mathcal{B}_j\}_{j=1}^{m}$ with confidences $\{conf_{\mathcal{B}_j}\}_{j=1}^{m}$, right lung mask L_r, left lung mask L_l
Output: Final lesion mask $\mathcal{M}_{\text{lesion}}$

$\mathcal{M}_{\text{lesion}} \leftarrow \emptyset$;
$\mathcal{M}_{\text{union}} \leftarrow \bigcup_{i=1}^{n} \mathcal{M}_i$;
foreach $\mathcal{B}_j \in \{\mathcal{B}_j\}_{j=1}^{m}$ **do**
  $c_1 \leftarrow \dfrac{|\mathcal{B}_j \cap \mathcal{M}_{\text{union}}|}{|\mathcal{B}_j \cup \mathcal{M}_{\text{union}}|} \geq \tau_{\text{anatomy}}$;
  $c_2 \leftarrow conf_{\mathcal{B}_j} \geq \tau_{\text{conf}}$;
  $c_3 \leftarrow \dfrac{|\mathcal{B}_j \cap \mathcal{A}|}{|\mathcal{B}_j|} \geq \tau_{\text{signal}}$;
  $c_4 \leftarrow \left(\dfrac{|\mathcal{B}_j \cap L_r|}{|\mathcal{B}_j \cup L_r|} \geq \tau_{\text{size}}\right) \vee \left(\dfrac{|\mathcal{B}_j \cap L_l|}{|\mathcal{B}_j \cup L_l|} \geq \tau_{\text{size}}\right)$;
  if $c_1 \wedge c_2 \wedge c_3 \wedge c_4$ **then**
    $C \leftarrow \text{FindIntersectingComponent}(\mathcal{B}_j, \mathcal{A})$;
    if C *is not empty* **then**
      $\mathcal{M}_{\text{new}} \leftarrow \text{Refine}(C)$;
      $\mathcal{M}_{\text{lesion}} \leftarrow \mathcal{M}_{\text{lesion}} \cup \mathcal{M}_{\text{new}}$;
return $\mathcal{M}_{\text{lesion}}$

location that spatially overlaps with a generated lesion mask, confirming successful localization of the reported finding. This location is derived from the anatomy masks $\{\mathcal{M}_i\}_{i=1}^{n}$ that intersect with the selected lesion box masks during the lesion mask generation. Finally, we introduce an *empty location*, which refers to a lung region with no reported lesions and is used to generate negative samples.

3.2. Instruction-Answer Pair Generation

With the information extracted from the previous process (*i.e.*, grounded lesion mask generation), we build our dataset for seven major lesion types found in CXRs: cardiomegaly, pneumonia, atelectasis, opacity, consolidation, edema, and effusion. These lesions are not only the most frequently mentioned in radiology reports, but also clinically significant to be common annotation targets in other medical datasets [3, 19, 20, 38]. For each lesion, we construct positive instruction-answer pairs, which include a ground-truth lesion mask. Negative pairs using an empty mask are also generated to enable the model to confirm the absence of lesions. An example of this pair generation process is shown in Fig. 3. Please refer to Appendix B and C for the lesion descriptions and specific dataset generation process.

Instruction Types and Limitations. We consider three types of segmentation instructions (Table 2). A *basic* instruction specifies both the segmentation target and its location. The location can be a broad region (such as left lung or right lung), one of eight more specific zones (apical, upper, mid, and lower zones for each lung), or a combination of these regions. In contrast, a *global* instruction specifies only the segmentation target. A *lesion inference* instruction asks the model to predict the type of lesion represented by an opacity within a given location. The generation of these instructions is inherently constrained by the grounded lesion mask generation. For example, a global instruction becomes invalid if

the generated mask captures only part of the lesion. To address this, our framework dynamically produces only those instruction–answer pairs that are valid given the grounding information available for each image.

Table 2. Templates for each question type. Each type includes answer templates for both positive and negative cases, with the negative answers positioned in the last row of each cell.

Type	Role	Template
Basic	Instruction	Segment the [Target] in the [Location].
	Answer	[SEG] [SEG] There is no [Target] in the [Location].
Global	Instruction	Segment the [Target].
	Answer	[SEG] It is located in the [Location]. [SEG] There is no [Target].
Lesion Inference	Instruction	Segment the opacity in the [Location] and predict its type.
	Answer	[SEG] It is highly suggestive of [Lesion]. [SEG] It possibly reflects [Lesion]. [SEG] There is no opacity in the [Location].

Instruction Generation. The instruction generation process begins by creating a basic instruction for each grounded lesion. Next, we determine whether a global instruction can be generated. The global instruction is created only when the *grounded location* and the *reported location* are identical. Separately, we generate lesion inference instructions by transforming the basic instructions for *pneumonia, atelectasis*, and *edema*, replacing these specific lesion types with *opacity*. This transformation is motivated by the fact that these findings are all specific types of "opacity," a more fundamental visual concept in medical imaging. Negative samples are generated by (1) selecting lesion types that are not mentioned or explicitly negated in the radiology report; or (2) utilizing *empty locations* to substitute the original location in the basic instruction of a positive sample.

Answer Generation. Each answer consists of a lesion mask and a textual description. The answer lesion masks for positive pairs are determined differently depending on whether they are organ-level or localized abnormalities. For cardiomegaly, we utilize a heart mask as its corresponding lesion mask since this condition is defined by the state of a specific organ [11]. In contrast, localized abnormalities (*e.g.*, pneumonia or effusion) can appear in variable locations, so for these findings, we use the lesion masks generated in Sec. 3.1. For negative pairs, an empty mask is used. As for the textual description, it is also provided for both positive and negative samples. Specifically, the answer template for lesion inference incorporates a certainty level.

4. MIMIC-ILS Dataset

Our final dataset, MIMIC-ILS, consists of 1.1M instruction-answer pairs (135K positive and 930K negative samples) derived from 192K MIMIC-CXR images. This final image set is obtained by first filtering out low-quality images (*e.g.*, images with extreme contrast issues), and then excluding any images for which no instruction-answer pairs are generated through our pipeline in Sec. 3. The positive samples are generated from 91K unique lesion masks, where each mask can be associated with multiple instruction–answer samples. The resulting dataset covers seven distinct lesion types, and the overall statistics are presented in Fig. 4. Following the official MIMIC-CXR split, the dataset is divided into 1M training samples, 8.2K validation samples, and 12K test samples. Details on quality control and a distribution of MIMIC-ILS are presented in Appendix D and E.

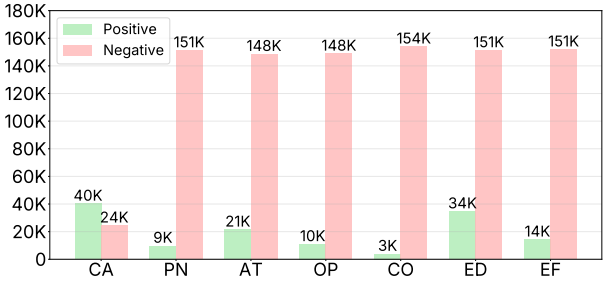

Figure 4. Distribution of MIMIC-ILS dataset. The y-axis indicates the number of samples, and the x-axis represents the lesion type. (CA: cardiomegaly, PN: pneumonia, AT: atelectasis, OP: opacity, CO: consolidation, ED: edema, EF: effusion)

Human Evaluation. To assess the quality of MIMIC-ILS, an expert review was conducted by four radiation oncologists specializing in lesion contouring on medical images. For the test set samples, clinicians classified each case as either acceptable or unacceptable based on mask quality. Positive cases were reviewed by all experts, while negatives were split among them. Any sample judged unacceptable by at least one expert was excluded from the final test set, and the results are summarized in Table 3. Among the 10.5K mask samples initially reviewed, 96.3% were rated as acceptable and finally included in the test set. More details on the expert profiles and quality assessment are provided in Appendix E.

Table 3. Acceptance rate and number of evaluated samples for the human evaluation. Each sample corresponds to a unique combination of lesion mask, target, and location.

Expert	Total		Positive		Negative	
	Rate (%)	# Samples	Rate (%)	# Samples	Rate (%)	# Samples
Expert A	96.1	4,055	95.6	1,841	96.5	2,214
Expert B	97.2	3,968	96.0	1,841	98.3	2,217
Expert C	98.7	4,002	99.8	1,841	97.8	2,161
Expert D	97.6	4,039	96.9	1,841	98.2	2,198
Overall	96.3	10,541	90.1	1,841	97.7	8,700

5. Model Training

Using MIMIC-ILS, we train our ILS model, ROSALIA. The model adopts the architecture of LISA [22], which demonstrated strong zero-shot language-guided segmentation performance in the general domain. As illustrated in Fig. 5, the

architecture integrates a VLM backbone with the Segment Anything Model (SAM) [21]. The VLM processes both the image and the input instruction to produce a special token, [SEG], along with its textual description. This [SEG] token embedding is then passed to SAM together with the input image for mask prediction. Within SAM, the frozen image encoder extracts embeddings from the image, and the mask decoder integrates these embeddings with the hidden embedding of [SEG] token to generate the final mask.

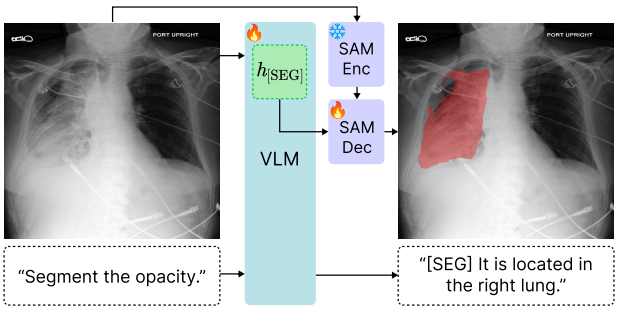

Figure 5. Overview of ROSALIA. The architecture integrates a VLM with the SAM. The VLM takes a CXR image and a segmentation instruction as input, generating both a textual description and a special [SEG] token. The hidden embedding of this [SEG] token is then passed to SAM's decoder to produce the final mask.

The overall loss function $\mathcal{L}$ consists of two components: (1) a language loss and (2) a mask loss. It is formulated as:

$$\mathcal{L} = \lambda_{\text{txt}}\mathcal{L}_{\text{txt}} + \mathcal{L}_{\text{mask}},$$
$$\mathcal{L}_{\text{mask}} = \lambda_{\text{bce}}\mathcal{L}_{\text{bce}} + \lambda_{\text{dice}}\mathcal{L}_{\text{dice}}.$$

$\mathcal{L}_{\text{txt}}$ denotes the autoregressive cross-entropy loss for the answer text, and $\mathcal{L}_{\text{mask}}$ represents the segmentation loss computed between the ground-truth mask and the predicted foreground probability map, which combines the binary cross-entropy loss $\mathcal{L}_{\text{bce}}$ and the DICE [30] loss $\mathcal{L}_{\text{dice}}$. The λ_{txt}, λ_{bce}, and λ_{dice} are coefficients for each loss term.

6. Experiments

6.1. Implementation Details

Training Details. ROSALIA is built on the LISA-7B architecture and is fine-tuned from its original checkpoint [22]. Following LISA, we adopted LLaVA [26] as the VLM backbone and employed the largest version of SAM (SAM-H). LoRA [14] fine-tuning was applied to the VLM with a rank of 128 and an alpha of 256, while the mask decoder was fully fine-tuned. The epochs and the initial learning rate were set to 15 and 0.0003, respectively, using the AdamW optimizer [29]. The total batch size was 256, and the ratio of positive to negative samples was maintained at 1:1 in each mini-batch. The loss coefficients λ_{txt}, λ_{bce}, and λ_{dice} were set to 0.5, 5, and 1, respectively, and the DICE loss was computed only for positive samples. Further model training details are described in the Appendix F.

Baseline Models. Since we present the first dataset for ILS in CXRs, no existing models have been directly trained on our proposed task. Nonetheless, we evaluated several models from both the general domain (LISA [22], Text4Seg [23], PixelLM [35]) and the medical domain (BiomedParse [42], RecLMIS [15], IMIS-Net [5]), which can take an image and text as input to produce a segmentation output.

Evaluation Metrics. We used three metrics to evaluate model performance. For positive samples, we used Intersection-over-Union (IoU)–based measures: gIoU and cIoU [22]. gIoU is the average IoU across samples, while cIoU is the ratio of total intersection to total union across the dataset. For negative cases, we used empty-target accuracy (N-Acc.), the proportion of samples correctly predicted to have no masks [39].

6.2. Main Results

Table 4 presents the results of the baselines and our proposed model on the MIMIC-ILS test set. While existing VLM-based segmentation models from both the general and medical domains struggle with the ILS task, ROSALIA achieves notably high performance. In particular, not only do these baselines yield low IoU scores on positive cases, but they also frequently fail on empty-target cases, where the N-Acc. rate is nearly zero in most instances. These results highlight the need for a dedicated dataset to effectively address the ILS task in CXRs. Furthermore, the strong results of ROSALIA on the physician-verified test set demonstrate that the training set of MIMIC-ILS serves as a high-quality resource—even without manual expert filtering.

Table 4. Segmentation results (%) on the MIMIC-ILS test set. "N-Acc." denotes the accuracy of correctly predicting empty targets. ¶ indicates medical domain baselines. The best and second-best results are marked in **bold** and underline, respectively.

Model	gIoU	cIoU	N-Acc.
LISA-7B [22]	8.3	12.8	0.7
LISA-13B [22]	8.9	12.2	0.0
Text4Seg [23]	6.1	10.3	20.6
PixelLM-7B [35]	9.2	11.8	0.0
PixelLM-13B [35]	12.8	15.4	0.0
BiomedParse¶ [42]	<u>23.8</u>	18.5	0.6
RecLMIS¶ [15]	21.9	<u>18.6</u>	0.0
IMIS-Net¶ [5]	9.8	11.8	<u>21.6</u>
ROSALIA (Ours)	**71.2**	**75.6**	**91.8**

Table 5 presents the performance of ROSALIA across different lesion types. The overall gIoU exceeds 0.7, indicating that more than 80% of the regions overlap between the predicted and ground-truth masks when the two are of similar size. Even for the lesion type with the lowest gIoU, the score remains above 0.55, suggesting over 70% regional overlap under similar mask sizes between the ground truth and predictions.

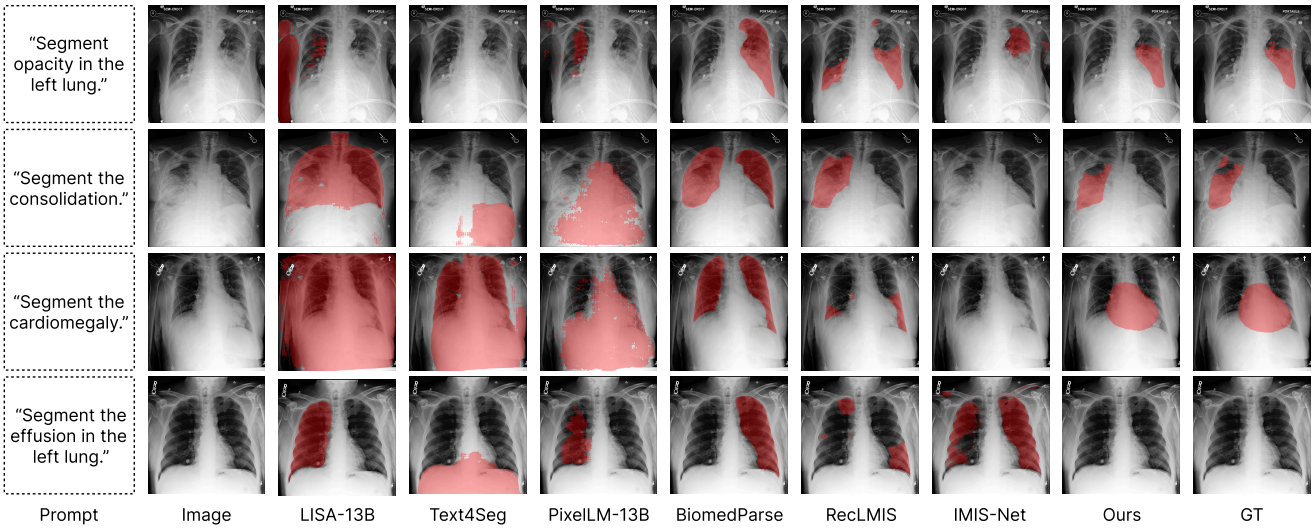

Figure 6. Visualized inference results of ROSALIA and baseline models. The first three rows show results for positive cases, while the last row presents results for negative cases with an empty target mask. Additional examples are demonstrated in Appendix G.

Table 5. Segmentation performance (%) of ROSALIA for each lesion type.

Lesion	gIoU	cIoU	N-Acc.
Cardiomegaly	89.0	89.0	85.8
Pneumonia	57.2	60.4	97.1
Atelectasis	60.2	58.7	91.7
Opacity	60.5	64.2	85.0
Consolidation	61.9	65.6	91.2
Edema	64.8	66.6	92.2
Effusion	60.3	59.6	90.4
Total	**71.2**	**75.6**	**91.8**

We also evaluate the accuracy of text responses across different question types, as shown in Table 6. A response is considered correct only when both the template and all variables for each question type (denoted by square brackets in Table 2) exactly match the structured ground-truth information. Despite this strict criterion, ROSALIA achieves high accuracy across most question types (see Appendix G for text accuracy of each lesion type).

Table 6. Text response accuracy (%) of ROSALIA.

Type	Overall	Basic	Global	Lesion Inf.
Positive	90.7	95.4	93.7	75.1
Negative	95.3	96.9	82.3	90.6
Total	**94.4**	**96.8**	**88.8**	**84.8**

6.3. Qualitative Results

Fig. 6 presents qualitative examples from each model for the ILS task. The baseline models largely fail, either producing entirely incorrect masks or segmenting the whole anatomical regions (*e.g.*, the left or right lung). In contrast, ROSALIA accurately segments only the lesion specified in the instruction within the designated region and correctly identifies empty-target cases. Additionally, Fig. 7 demonstrates the outputs produced from diverse instructions applied to the same input image. Although multiple lesions coexist in the image, ROSALIA accurately interprets each instruction and generates results tailored to the user's specific request. This highlights the model's ability to handle diverse lesion types and locations of interest.

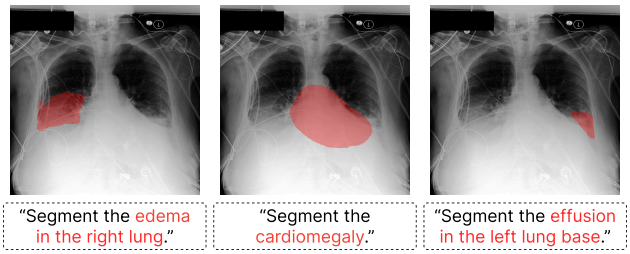

Figure 7. Examples of outputs from different instructions applied to the same image. Among the multiple lesions present, ROSALIA can selectively segment only the lesion and location of interest.

7. Conclusion

In this study, we introduce MIMIC-ILS, the first dataset for instruction-guided lesion segmentation in CXRs, along with ROSALIA, a VLM developed for this new paradigm. Our automated pipeline enables the construction of this million-scale dataset, and expert evaluations show a remarkably high acceptance rate, confirming the quality and reliability of our fully human-free data generation process. Trained on MIMIC-ILS, ROSALIA demonstrates a comprehensive ability to generate accurate lesion segmentations and textual responses across diverse user instructions. These findings indicate that MIMIC-ILS and ROSALIA offer a strong foundation for advancing research on fine-grained lesion grounding in the CXR domain.

References

[1] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022. 3

[2] Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. The liver tumor segmentation benchmark (lits). *Medical image analysis*, 84:102680, 2023. 3

[3] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pages 1–21. Springer, 2022. 2, 5

[4] Joshua Broder. Imaging the chest: the chest radiograph. *Diagnostic imaging for the emergency physician*, page 185, 2011. 1

[5] 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 Computer Vision and Pattern Recognition Conference*, pages 20841–20851, 2025. 3, 7

[6] Sherri de Coronado, Margaret W Haber, Nicholas Sioutos, Mark S Tuttle, and Lawrence W Wright. Nci thesaurus: using science-based terminology to integrate cancer research results. In *MEDINFO 2004*, pages 33–37. IOS Press, 2004. 1

[7] Viacheslav V Danilov, Alex Proutski, Alex Karpovsky, Alexander Kirpich, Diana Litmanovich, Dato Nefaridze, Oleg Talalov, Semyon Semyonov, Vladimir Koniukhovskii, Vladimir Shvartc, et al. Indirect supervision applied to covid-19 and pneumonia classification. *Informatics in Medicine Unlocked*, 28:100835, 2022. 2

[8] Daniel Coelho de Castro, Aurelia Bustos, Shruthi Bannur, Stephanie L Hyland, Kenza Bouzid, Maria Teodora Wetscherek, Maria Dolores Sánchez-Valverde, Lara Jaques-Pérez, Lourdes Pérez-Rodríguez, Kenji Takeda, et al. Padchest-gr: A bilingual chest x-ray dataset for grounded radiology report generation. *NEJM AI*, 2(7):AIdbp2401120, 2025. 2, 3

[9] Aysen Degerli, Serkan Kiranyaz, Muhammad EH Chowdhury, and Moncef Gabbouj. Osegnet: Operational segmentation network for covid-19 detection using chest x-ray images. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 2306–2310. IEEE, 2022. 2, 3

[10] Nicolas Gaggion, Candelaria Mosquera, Martina Aineseder, Lucas Mansilla, Diego Milone, and Enzo Ferrante. Chexmask database: a large-scale dataset of anatomical segmentation masks for chest x-ray images. 1

[11] Nicolas Gaggion, Lucas Mansilla, Candelaria Mosquera, Diego H. Milone, and Enzo Ferrante. Improving anatomical plausibility in medical image segmentation via hybrid graph neural networks: applications to chest x-ray analysis. *IEEE Transactions on Medical Imaging*, 2022. 6

[12] Nicolás Gaggion, Candelaria Mosquera, Lucas Mansilla, Julia Mariel Saidman, Martina Aineseder, Diego H Milone, and Enzo Ferrante. Chexmask: a large-scale dataset of anatomical segmentation masks for multi-center chest x-ray images. *Scientific Data*, 11(1):511, 2024. 1

[13] Nicholas Heller, Sean McSweeney, Matthew Thomas Peterson, Sarah Peterson, Jack Rickman, Bethany Stai, Resha Tejpaul, Makinna Oestreich, Paul Blake, Joel Rosenberg, et al. An international challenge to use artificial intelligence to define the state-of-the-art in kidney and kidney tumor segmentation in ct imaging., 2020. 3

[14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 7

[15] 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*, 2024. 1, 3, 7

[16] Yankai Jiang, Zhongzhen Huang, Rongzhao Zhang, Xiaofan Zhang, and Shaoting Zhang. Zept: Zero-shot pan-tumor segmentation via query-disentangling and self-prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11386–11397, 2024. 3

[17] Yankai Jiang, Wenhui Lei, Xiaofan Zhang, and Shaoting Zhang. Unleashing the potential of vision-language pre-training for 3d zero-shot lesion segmentation via mask-attribute alignment. *arXiv preprint arXiv:2410.15744*, 2024. 3

[18] Alistair Johnson, Matt Lungren, Yifan Peng, Zhiyong Lu, Roger Mark, Seth Berkowitz, and Steven Horng. Mimic-cxr-jpg-chest radiographs with structured labels. *PhysioNet*, 101: 215–220, 2019. 1

[19] A Johnson, T Pollard, R Mark, S Berkowitz, and S Horng. Mimic-cxr database (version 2.1. 0). physionet. rrid: Scr_007345, 2024. 2, 3, 5, 1

[20] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019. 2, 3, 5, 1

[21] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 7

[22] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024. 1, 3, 6, 7

[23] Mengcheng Lan, Chaofeng Chen, Yue Zhou, Jiaxing Xu, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Text4seg: Reimagining image segmentation as text generation. *arXiv preprint arXiv:2410.09855*, 2024. 1, 3, 7

[24] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022. 3

[25] 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. 1, 3

[26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 7

[27] Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. Clip-driven universal model for organ segmentation and tumor detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 21152–21164, 2023. 3

[28] Yun Liu, Yu-Huan Wu, Yunfeng Ban, Huifang Wang, and Ming-Ming Cheng. Rethinking computer-aided tuberculosis diagnosis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2646–2655, 2020. 2, 3

[29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7

[30] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. 7

[31] Duc Nguyen, DungNB, Ha Q. Nguyen, Julia Elliott, Nguyen-ThanhNhan, and Phil Culliton. Vinbigdata chest x-ray abnormalities detection. https://kaggle.com/competitions/vinbigdata-chest-xray-abnormalities-detection, 2020. Kaggle. 1

[32] Dung Nguyen, Minh Khoi Ho, Huy Ta, Thanh Tam Nguyen, Qi Chen, Kumar Rav, Quy Duong Dang, Satwik Ramchandre, Son Lam Phung, Zhibin Liao, Minh-Son To, Johan Verjans, Phi Le Nguyen, and Vu Minh Hieu Phan. Localizing before answering: A benchmark for grounded medical visual question answering. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2025)*, pages 7670–7676, 2025. 3

[33] Ha Q Nguyen, Khanh Lam, Linh T Le, Hieu H Pham, Dat Q Tran, Dung B Nguyen, Dung D Le, Chi M Pham, Hang TT Tong, Diep H Dinh, et al. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. *Scientific Data*, 9(1): 429, 2022. 2, 3, 1

[34] Fernando Pérez-García, Sam Bond-Taylor, Pedro P Sanchez, Boris van Breugel, Daniel C Castro, Harshita Sharma, Valentina Salvatelli, Maria TA Wetscherek, Hannah Richardson, Matthew P Lungren, et al. Radedit: stress-testing biomedical vision models via diffusion image editing. In *European Conference on Computer Vision*, pages 358–376. Springer, 2024. 3, 1

[35] Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26374–26383, 2024. 1, 3, 7

[36] 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. 3, 1

[37] Gyungin Shin, Weidi Xie, and Samuel Albanie. Reco: Retrieve and co-segment for zero-shot transfer. *Advances in Neural Information Processing Systems*, 35:33754–33767, 2022. 3

[38] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017. 5

[39] Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. Gsva: Generalized segmentation via multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3858–3869, 2024. 7

[40] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18134–18144, 2022. 3

[41] Anna Zawacki, Carol Wu, George Shih, Julia Elliott, Mikhail Fomitchev, Mohannad Hussain, ParasLakhani, Phil Culliton, and Shunxing Bao. Siim-acr pneumothorax segmentation. https://kaggle.com/competitions/siim-acr-pneumothorax-segmentation, 2019. Kaggle. 2, 3

[42] 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. 3, 7

[43] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European conference on computer vision*, pages 696–712. Springer, 2022. 3