AnatomiX, an Anatomy-Aware Grounded Multimodal Large Language Model for Chest X-Ray Interpretation

Supplementary Material

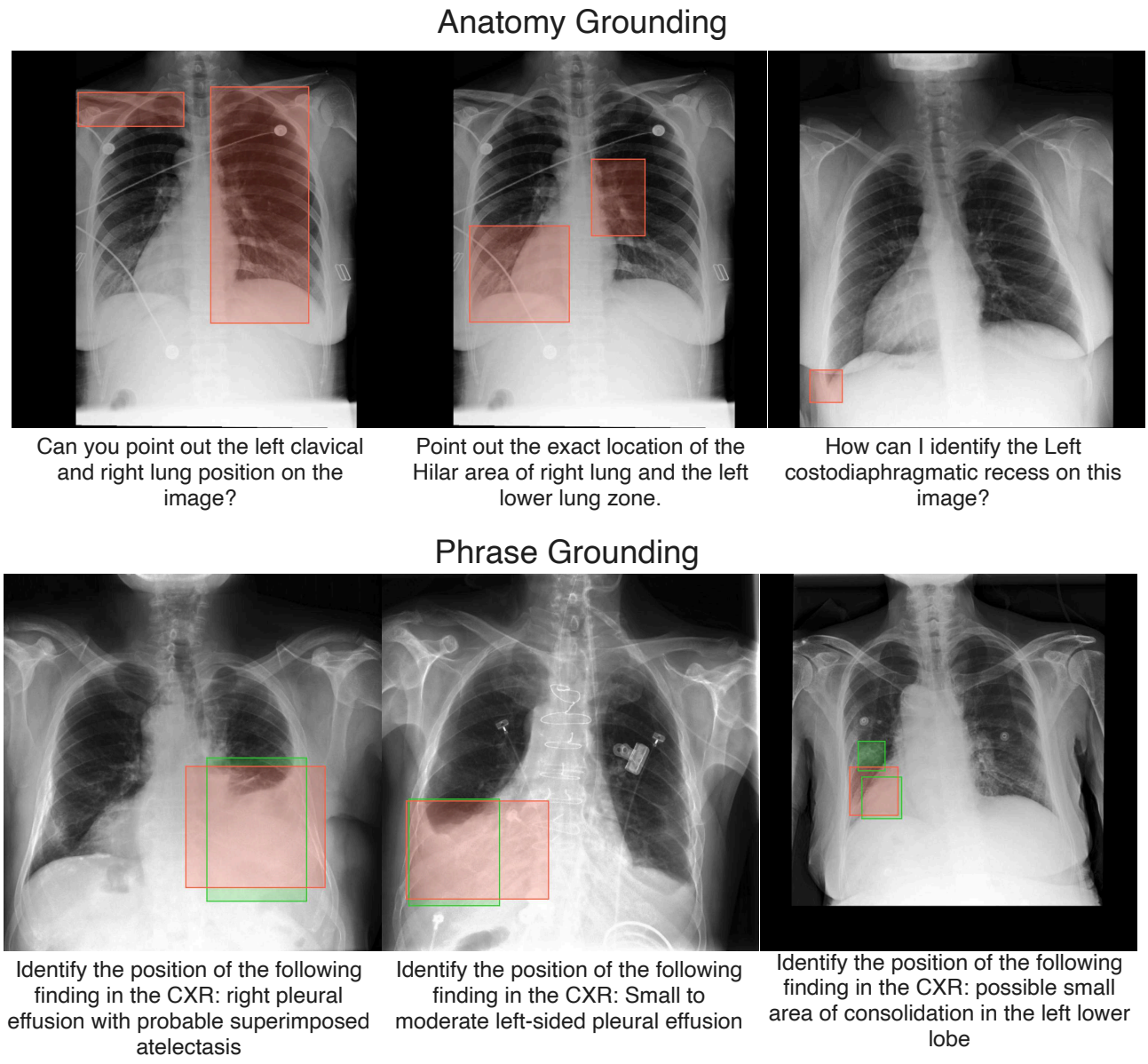

Figure S1. Additional samples for AnatomiX's output for anatomy and phrase grounding on flipped images (left ↔ right) with radiographic markers removed. Bottom: Green boxes show the ground truth, while red show the model prediction.

Multimodal Prompt

User:
You are a professional radiologist. I will provide you with context containing likely features about different parts of the chest X-rays.

Image:
<image_start> <image> ... <image> <image_end>

Likely findings:
<emb><obj_0></emb> <box>(0,387),(670,1024)</box> Abdominal cavity shows enteric tube.
<emb><obj_1></emb> <box>(300,118),(394,207)</box> Aortic arch structure is healthy.
...
<emb><obj_N></emb> <box>(398,391),(517,518)</box> Right cardiophrenic sulcus is healthy.

Task: [TASK]

Model Response: [MODEL RESPONSE]

Figure S2. Multimodal prompt template used in $\mathcal{LM}$. Colored tags (<emb> and <box>) denote special tokens corresponding to anatomical object embeddings and bounding boxes, respectively. Each <obj_i> token represents the embedding of the i^{th} anatomical object, while <image> indicates image patch embeddings.

S1. Self-Similarity Loss Matrix

The Contrastive Alignment stage of the Anatomy Perception Module (APM) utilizes the Self-Similarity matrix S_{self} to model fine grained semantic relations among anatomical descriptions. In this stage, a pretrained sentence encoder $\mathcal{S}$ provides indirect supervision by embedding textual inputs into a continuous semantic space that captures linguistic and clinical similarities. Given a set of input sentences S_t, each sentence is encoded through $\mathcal{S}$ to obtain text embeddings $\mathrm{S}_E \in \mathbb{R}^{N\times 768}$, where each row corresponds to the representation of one sentence in a 768-dimensional embedding space. To ensure consistency and comparability across representations, the embeddings are first normalized using ℓ_2 normalization:

$$\bar{\mathrm{S}}_E = \frac{\mathrm{S}_E}{|\mathrm{S}_E|_2} \tag{S1}$$

This normalization projects the embeddings onto a unit hypersphere, ensuring that they encode directional (semantic) differences rather than magnitude based variations. The normalized embeddings are then used to compute the Self-Similarity matrix:

$$\mathrm{S}_{self} = \mathrm{Softmax}(\bar{\mathrm{S}}_E) \cdot \mathrm{Softmax}(\bar{\mathrm{S}}_E)^{\mathrm{T}} \tag{S2}$$

where $\mathrm{S}_{self} \in \mathbb{R}^{N\times N}$ encodes pairwise similarity scores between all sentences in S_t. The softmax operation (applied row wise) ensures these similarities are smooth and probabilistically interpretable.

Next, to align anatomical and textual semantics, we compute a projected similarity matrix $\hat{\mathrm{K}}_A$ between the projected anatomical features $\hat{\mathrm{O}}_A$ and projected text embeddings $\hat{\mathrm{S}}_A$:

$$\hat{\mathrm{K}}_A = \mathrm{Softmax}\left(\frac{\hat{\mathrm{O}}_A \hat{\mathrm{S}}_A^{\mathrm{T}}}{\tau}\right) \tag{S3}$$

where $\hat{\mathrm{K}}_A \in \mathbb{R}^{N\times N}$, and τ is the temperature coefficient (set to 0.01) that controls the sharpness of the similarity distribution.

The final contrastive alignment loss is defined as the averaged KL-divergence between the anatomical–textual similarity matrix $\hat{\mathrm{K}}_A$ and the self-similarity matrix S_{self}, computed in both row-wise and column-wise directions to enforce mutual consistency:

$$\mathcal{L}_{CL} = \frac{1}{2} KL\left(\hat{\mathrm{K}}_A, \mathrm{S}_{self}\right) + \frac{1}{2} KL\left(\hat{\mathrm{K}}_A^{\mathrm{T}}, \mathrm{S}_{self}^{\mathrm{T}}\right) \tag{S4}$$

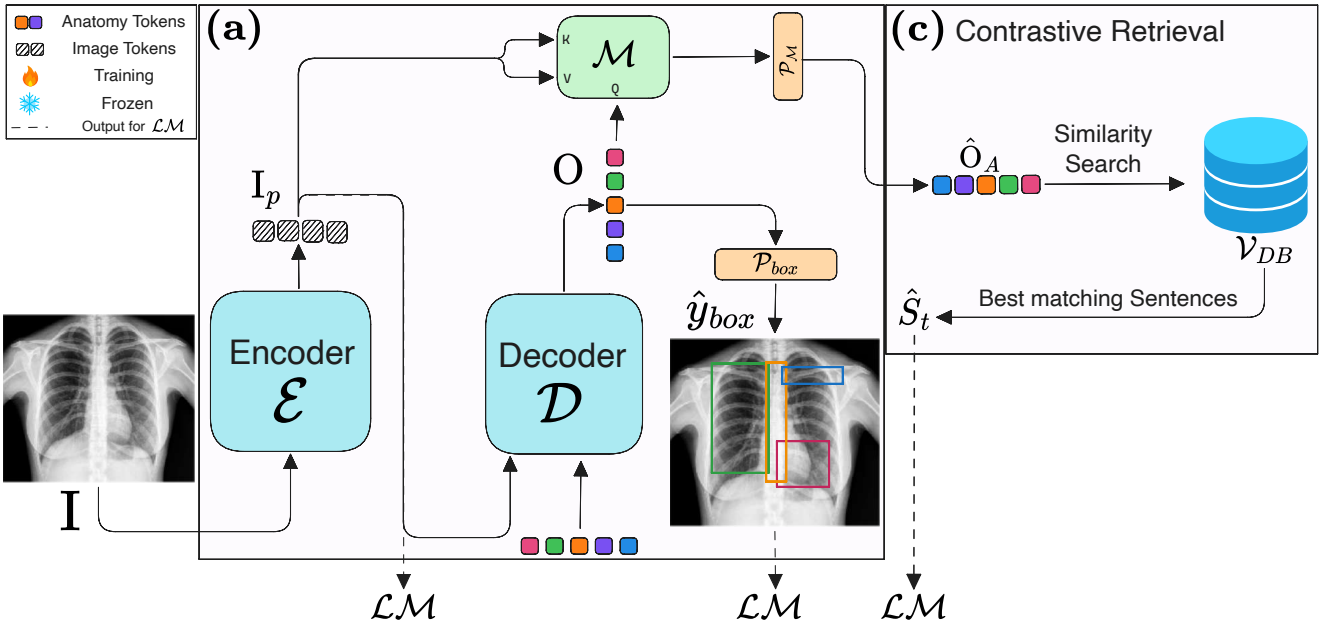

Figure S3. APM architecture during inference, where the Contrastive Alignment's components are replaced with vector database for the Contrastive Retrieval. See Fig. 2 for training architecture.

This formulation encourages $\hat{\mathrm{O}}_A$ and $\hat{\mathrm{S}}_A$ to maintain pairwise relationships that reflect the semantic structure captured in S_{self}. As a result, the APM preserves semantic coherence and clinical consistency across related sentences, capturing overlapping anatomical features rather than enforcing strict one-to-one alignments.

S2. Vector Database

During APM inference, we replace the Contrastive Alignment with the Contrastive Retrieval (see Fig. S3) to identify the semantically most similar sentences to the anatomical object tokens $\hat{\mathrm{O}}_A$. These retrieved sentences represent the most probable observations for each anatomical object and thus provide important contextual information for downstream descriptive tasks in $\mathcal{LM}$, as discussed in ablations.

The vector database, denoted as $\mathcal{V}_{DB}$, stores all unique sentences associated with each anatomical object from the validation set of the Chest-ImaGenome dataset. For every object in an image, we construct a concise descriptive sentence using the corresponding phrases and attributes given in the original dataset (example sentence: "Right lower lung shows pleural effusion and atelectasis"). To build $\mathcal{V}_{DB}$, we first compile the set of unique sentences for each anatomical object. Each sentence is then encoded using the sentence encoder $\mathcal{S}$ and the trained projection head $\mathcal{P}_S$, producing s-dimensional embeddings (see Fig. 2 for $\mathcal{P}_S$). These embeddings, along with their corresponding sentences, are stored as key–value pairs in $\mathcal{V}_{DB}$, with a distinct sub-database allocated to each anatomical object. Consequently, $\mathcal{V}_{DB}$ comprises N independent sub-databases. The full distribution for the size of each anaomical object database is shown in Fig. S4.

The compact nature of both the embeddings and the sentences ensures that $\mathcal{V}_{DB}$ remains lightweight, enabling efficient retrieval at inference time. During inference, a similarity search is performed between each anatomical object token $\hat{\mathrm{O}}_A^i$ and the sentence embeddings within the corresponding sub-database of $\mathcal{V}_{DB}$, thereby retrieving the most relevant descriptive sentences for each object. These sentences are then passed to $\mathcal{LM}$ using a multimodal prompt template shown in Fig. S2, where they provide important contextual information about each anatomy.

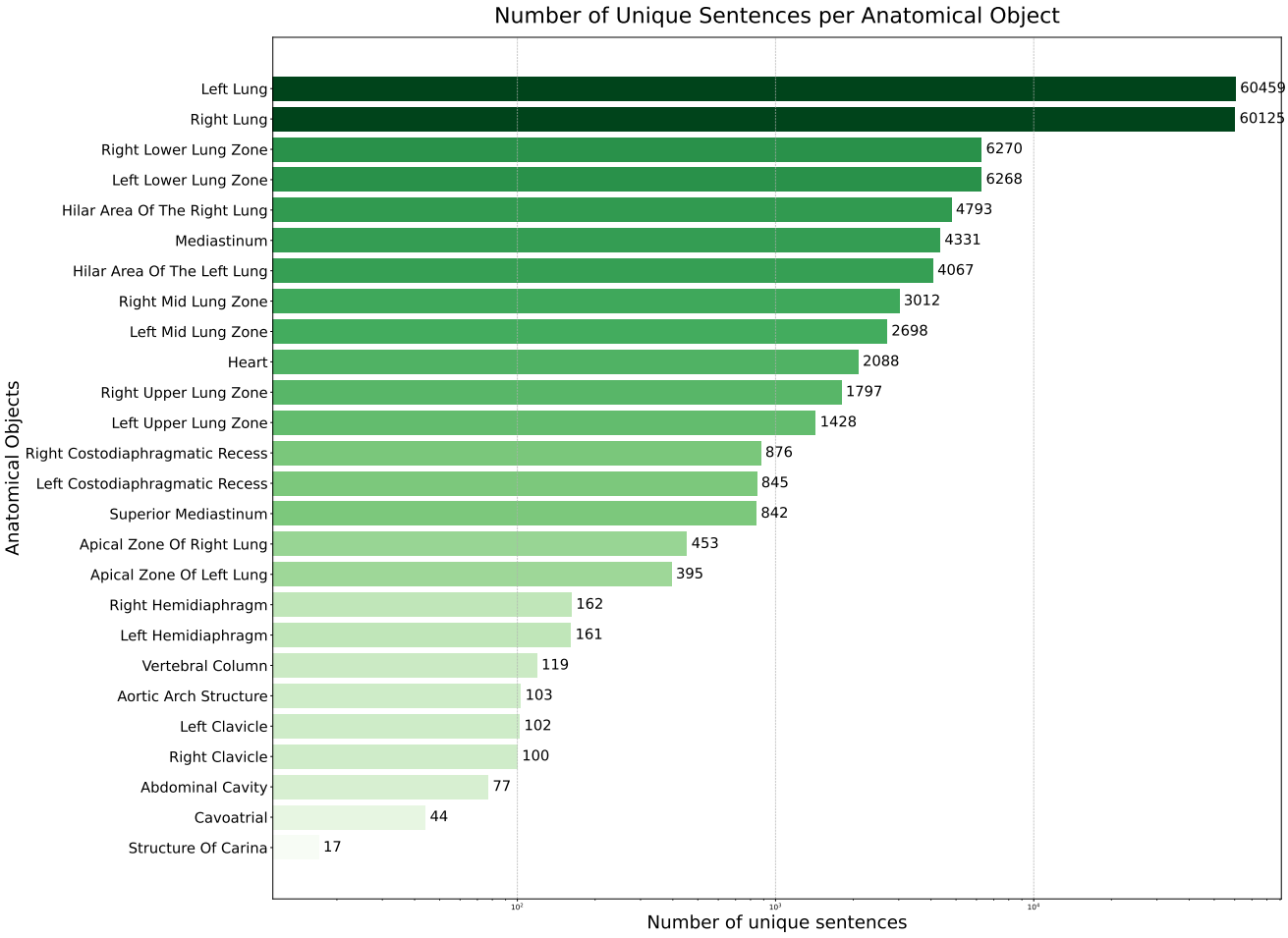

Figure S4. Size of the vector database. Each bar shows the number of unique sentences associated with a specific anatomical object. Anatomical objects with fewer than 10 unique sentences are omitted for clarity.

S3. Datasets

S3.1. Anatomy Grounding

For instruction tuning, we construct the Anatomy Grounding dataset using bounding box annotations for 36 distinct anatomical structures. We design 20 question and answer templates, as illustrated in Fig. S5. The question templates query the location of a specific anatomical structure, while the answer templates include both the anatomical name and its location (i.e., bounding box coordinates). During dataset construction, we randomly sample from the question and answer templates to increase data diversity. The final dataset contains the same number of samples as Chest-ImaGenome, maintains a uniform distribution of anatomical structures, and follows the official data split of Chest-ImaGenome.

S3.2. VinDr-Instruct

The VinDr-Instruct dataset is constructed from the VinDr-CXR dataset and comprises question–answer pairs for abnormality detection, phrase grounding, and grounded diagnosis tasks. We adopt the question templates proposed in CheXagent, as illustrated in Fig. S6. The answers are formatted as single- or multi-word responses that contain only the essential information, without full sentence structures. In this work, we follow the original train–test split of the VinDr-CXR dataset.

Templates for VinDr-Instruct

Questions:

- Where is the {anatomy} located in this Chest X-ray?
- Can you point out the {anatomy}'s position on the image?
- What's the location of the {anatomy} in the X-ray?
- Identify where the {anatomy} is on this Chest X-ray, please.
- Where exactly is the {anatomy} found on this image?
- Could you specify where to find the {anatomy} on this X-ray?
- Highlight the {anatomy}'s area on the image.
- Show me the {anatomy}'s location on this CXR.
- Where should I look to find the {anatomy} in this image?
- Can you locate the {anatomy} on this X-ray for me?
- Please point to the {anatomy} on this Chest X-ray.
- Indicate the position of the {anatomy} on this image.
- Describe the location of the {anatomy} on the X-ray.
- Where on this image is the {anatomy} located?
- Point out the exact location of the {anatomy} in the Chest X-ray.
- How can I identify the {anatomy} on this image?
- Where is the {anatomy} situated in this CXR?
- Can you highlight the {anatomy} on this image?
- Indicate where the {anatomy} is found on this X-ray.
- Describe where to find the {anatomy} on this Chest X-ray.

Answers:

- The ¡ref¿{anatomy}¡/ref¿ is located at the coordinates ¡box¿{boxes}¡/box¿ on the image.
- You'll find the ¡ref¿{anatomy}¡/ref¿ at ¡box¿{boxes}¡/box¿ in the X-ray.
- The ¡ref¿{anatomy}¡/ref¿ can be seen at ¡box¿{boxes}¡/box¿ on the Chest X-ray.
- The location of the ¡ref¿{anatomy}¡/ref¿ is at ¡box¿{boxes}¡/box¿ on the image.
- For the ¡ref¿{anatomy}¡/ref¿, the coordinates are ¡box¿{boxes}¡/box¿ on the X-ray.
- The ¡ref¿{anatomy}¡/ref¿ is situated at ¡box¿{boxes}¡/box¿ in the image.
- On the Chest X-ray, the ¡ref¿{anatomy}¡/ref¿ is located at ¡box¿{boxes}¡/box¿
- The ¡ref¿{anatomy}¡/ref¿ appears at the coordinates ¡box¿{boxes}¡/box¿ on the image.
- In the X-ray, the ¡ref¿{anatomy}¡/ref¿ is identifiable at ¡box¿{boxes}¡/box¿.
- The location for the ¡ref¿{anatomy}¡/ref¿ is marked at ¡box¿{boxes}¡/box¿ on the Chest X-ray.
- The ¡ref¿{anatomy}¡/ref¿ is positioned at ¡box¿{boxes}¡/box¿ on the image.
- The area occupied by the ¡ref¿{anatomy}¡/ref¿ is at ¡box¿{boxes}¡/box¿ in the X-ray.
- On the image, you can find the ¡ref¿{anatomy}¡/ref¿ at ¡box¿{boxes}¡/box¿.
- The ¡ref¿{anatomy}¡/ref¿'s location is at ¡box¿{boxes}¡/box¿ on the Chest X-ray.
- In terms of coordinates, the ¡ref¿{anatomy}¡/ref¿ is found at ¡box¿{boxes}¡/box¿ on the image.
- Regarding the ¡ref¿{anatomy}¡/ref¿, it is located at ¡box¿{boxes}¡/box¿ on the X-ray.
- The ¡ref¿{anatomy}¡/ref¿ specifically is at ¡box¿{boxes}¡/box¿ on the Chest X-ray.
- Concerning the ¡ref¿{anatomy}¡/ref¿, you will find it at ¡box¿{boxes}¡/box¿ in the image.
- The ¡ref¿{anatomy}¡/ref¿ is at ¡box¿{boxes}¡/box¿ on the X-ray.
- For identifying the ¡ref¿{anatomy}¡/ref¿, look at ¡box¿{boxes}¡/box¿ on the Chest X-ray.

Figure S5. Instruction QA templates for Anatomy Grounding dataset. {anatomy} and {boxes} represents the Anatomy name and location (coordinates), respectively.

Templates for VinDr-Instruct

Abnormality Detection:

- Detect {disease} in the given image.
- Locate areas in the chest X-ray where {disease} are present, using bounding box coordinates
- Perform abnormality detection (in the bounding box format) for the given image.
- Find the locations of {disease} in the bounding box format for the given image.
- Locate {disease} for the given image.
- Examine the chest X-ray and mark the regions affected by {disease} with bounding boxes.
- Detect the following in the image: {disease}.
- Examine the image for regions affected by {disease}, and indicate their positions with bounding boxes.
- Perform detection for {disease}.

Phrase Grounding:

- Detect {disease} in the given image.
- Locate areas in the chest X-ray where {disease} is present, using bounding box coordinates.
- Localize {disease} in the bounding box format for the given image.
- Find the locations of {disease} in the bounding box format for the given image.
- Locate {disease} for the given image.
- Examine the chest X-ray and mark the regions affected by {disease} with bounding boxes.
- Detect the following in the image: {disease}.
- Examine the image for regions affected by {disease}, and indicate their positions with bounding boxes.
- Perform detection for {disease}.

Grounded Diagnosis:

- Please give the corresponding diagnosis for the following region(s): {boxes}
- Provide a diagnosis based on the content of the following region(s): {boxes}

Figure S6. Instruction templates used for generating VinDr-Instruct dataset. {disease} and {boxes} represents the input abnormality name and box coordinates, respectively.

Table S1. Number of training, validation, and test samples for the nine datasets used in $\mathcal{LM}$ training and validation. The Source column indicates the original public dataset used directly or as the basis for dataset creation.

Dataset	Source	Test	Train	Val
MIMIC-VQA	MIMIC-CXR-VQA	5,497	101,963	4,926
RaDialog Instruct	RaDialog Instruct	799	6,274	822
SLAKE	SLAKE	298	1,175	285
Anatomy Grounding	Chest-ImaGenome	3,403	237,938	1,959
MS-CXR	MS-CXR	528	2,445	507
VinDr-Instruct	VinDr-CXR	6,166	38,122	4,099
PadChest-Grounding	PadChest-Grounding	1,121	3,920	558
MIMIC-CXR Classification	MIMIC-CXR	2,957	182,425	1,666
MIMIC Report Gen	MIMIC-CXR	1,722	135,049	1,078
Total	**–**	**22,491**	**709,311**	**15,900**

S4. Radiology Tasks

AnatomiX is trained and evaluated on nine CXR-related tasks, spanning four categories: image understanding, grounding, report generation, and visual question answering (VQA). Each of these tasks is focused on specific aspect of the CXR interpretation and uses different dataset(s).

Image Understanding: This category includes multi-label image classification across 14 classes using the MIMIC-CXR dataset, as well as CXR abnormality detection leveraging the VinDr-Instruct dataset. Fig. S8 shows sample input-output samples for classification and abnormality detection tasks along with the output of our model.

Grounding: We include four challenging grounding tasks in this work, namely: Phrase Grounding, Grounded Diagnosis, Grounded Captioning, and Anatomy Grounding. In Phrase Grounding, the model identifies the spatial location of a given phrase within an input image, utilizing the MS-CXR, PadChest-Gr, and VinDr-Instruct datasets. Grounded Diagnosis and Grounded Captioning require the model to infer a diagnosis and generate a textual description for a specified image region, respectively; we use VinDr-Instruct and MS-CXR for Grounded Diagnosis and MS-CXR for Grounded Captioning. Finally, Anatomy Grounding uses the Anatomy-Grounding dataset to localize anatomical structures based on user-provided textual prompts. Fig. 5 shows sample input and output pairs for these tasks, along with the output of our model.

Report Generation: This task involves generating the full report, including both the findings and impression sections using MIMIC-CXR dataset. A sample image-report is shown in Fig. S7 along with the output of AnatomiX.

Visual Question Answering: The VQA category consists of open-ended and closed-ended question answering tasks, derived from a combination of the MIMIC-VQA, SLAKE, and Radialog-Instruct datasets.

S5. Ablations

This section contains the detailed results for the ablations conducted for APM and $\mathcal{LM}$.

Table S2. Ablation results for grounding tasks. Grounded Diagnosis (GD) and Grounded Captioning (GC) results are given as: GD / GC.

Model	**NLG (GD/GC)**			**Clinical (GD/GC)**		**Phrase Gr.**		**Anatomy Gr.**	
	BERTScore	ROUGE	METEOR	RadGraph-F1	CheXbert-14-F1	IoU	mAP	IoU	mAP
AnatomiX-I_p	0.10 / 0.06	0.11 / 0.04	0.07 / 0.04	0.08 / 0.05	0.25 / 0.21	0.10	0.03	0.04	0.01
AnatomiX-$\hat{\mathrm{O}}_A$	0.42 / 0.17	0.38 / 0.12	0.26 / 0.06	0.35 / 0.08	0.42 / 0.23	0.24	0.16	0.36	0.27
AnatomiX-$\hat{S}_t$	0.19 / 0.25	0.17 / 0.23	0.16 / 0.18	0.23 / 0.22	0.28 / 0.24	0.11	0.05	0.06	0.02
AnatomiX-$\hat{y}_{box}$	0.31 / 0.23	0.34 / 0.21	0.25 / 0.13	0.37 / 0.25	0.40 / 0.39	0.17	0.12	0.46	0.37
AnatomiX-$\hat{S}_t$-$\hat{y}_{box}$	0.49 / 0.45	0.52 / 0.36	0.34 / 0.26	0.51 / 0.34	0.49 / 0.61	0.26	0.17	0.47	0.35
AnatomiX-$\hat{O}_A$-$\hat{y}_{box}$	0.52 / 0.40	0.58 / 0.33	0.37 / 0.28	0.62 / 0.37	0.50 / 0.67	0.36	0.24	0.61	0.53
AnatomiX-$\hat{O}_A$-$\hat{S}_t$	0.56 / 0.48	0.54 / 0.45	0.36 / 0.31	0.62 / 0.48	0.51 / 0.66	0.42	0.31	0.58	0.49
AnatomiX	**0.63 / 0.65**	**0.60 / 0.56**	**0.42 / 0.48**	**0.58 / 0.50**	**0.54 / 0.78**	**0.46**	**0.35**	**0.73**	**0.66**

Table S3. Ablations results for report generation task grouped by NLG and Clinical metrics.

Model	**NLG Metrics**			**Clinical Metrics**	
	ROUGE	BERTScore	METEOR	RadGraph	CheXbert-14 F1
AnatomiX-I_p	0.15	0.18	0.09	0.15	0.24
AnatomiX-$\hat{\mathrm{O}}_A$	0.14	0.18	0.09	0.13	0.22
AnatomiX-$\hat{S}_t$	0.32	0.27	0.13	0.19	0.30
AnatomiX-$\hat{y}_{box}$	0.13	0.15	0.10	0.11	0.21
AnatomiX-$\hat{S}_t$-$\hat{y}_{box}$	0.27	0.24	0.10	0.15	0.23
AnatomiX-$\hat{O}_A$-$\hat{y}_{box}$	0.15	0.21	0.11	0.14	0.22
AnatomiX-$\hat{O}_A$-$\hat{S}_t$	0.46	0.33	0.19	0.23	0.39
AnatomiX	**0.53**	**0.38**	**0.21**	**0.26**	**0.42**

S5.1. LLM

The naming convention for $\mathcal{LM}$'s ablations is as follows: (1) AnatomiX-I_p uses only image embeddings I_p; (2) AnatomiX-$\hat{\mathrm{O}}_A$ augments I_p with anatomical tokens $\hat{\mathrm{O}}_A$; (3) AnatomiX-$\hat{S}_t$ combines retrieved sentences $\hat{S}_t$ with I_p; (4) AnatomiX-$\hat{y}_{box}$ integrates predicted bounding boxes $\hat{y}_{box}$ with I_p; (5) AnatomiX-$\hat{S}_t$-$\hat{y}_{box}$ uses $\hat{S}_t$, $\hat{y}_{box}$, and I_p; (6) AnatomiX-$\hat{O}_A$-$\hat{y}_{box}$ combines $\hat{O}_A$, $\hat{y}_{box}$, and I_p; and (7) AnatomiX-$\hat{O}_A$-$\hat{S}_t$ incorporates $\hat{O}_A$, $\hat{S}_t$, and I_p.

Results on grounding tasks (Table S2) indicate substantial performance degradation when anatomical tokens, bounding boxes, and retrieved sentences are removed. Incorporating predicted boxes $\hat{y}_{box}$ markedly improves anatomy grounding and further benefits other tasks when combined with $\hat{S}_t$ and $\hat{O}_A$. Anatomical tokens $\hat{O}_A$ particularly enhance phrase grounding and anatomy grounding, but contribute less to grounded captioning, which requires detailed descriptions; in this setting, retrieved sentences $\hat{S}_t$ provide clear gains. Consistently, report generation results (Table S3) show that adding $\hat{S}_t$ yields the largest improvement, underscoring its importance for descriptive generation. Similar patterns are observed in VQA and image understanding (Tables S4 and S5), where $\hat{O}_A$ supports spatial reasoning, while $\hat{S}_t$ primarily benefits linguistically intensive tasks. Overall, combining all components achieves the best performance across tasks, particularly demonstrating strong anatomical understanding.

Table S4. Performance on image classification and abnormality detection tasks.

Model	Classification	Detection	
	CheXbert-14 F1	IoU	mAP
AnatomiX-$\hat{O}_A$	0.77	0.23	0.14
AnatomiX-$\hat{S}_t$	0.81	0.11	0.05
AnatomiX-$\hat{y}_{box}$	0.78	0.18	0.10
AnatomiX-$\hat{S}_t$-$\hat{y}_{box}$	0.83	0.21	0.13
AnatomiX-$\hat{O}_A$-$\hat{y}_{box}$	0.79	0.28	0.19
AnatomiX-$\hat{O}_A$-$\hat{S}_t$	<u>0.84</u>	<u>0.29</u>	<u>0.19</u>
AnatomiX-I_p	0.75	0.08	0.06
AnatomiX	**0.85**	**0.31**	**0.20**

Table S5. Open and close ended VQA task performance.

Model	Open-Ended VQA		Close-Ended VQA	
	BERTScore	CheXbert-14 F1	BERTScore	CheXbert-14 F1
AnatomiX-$\hat{O}_A$	0.70	0.74	0.73	0.91
AnatomiX-$\hat{S}_t$	0.81	0.84	0.81	0.94
AnatomiX-$\hat{y}_{box}$	0.67	0.72	0.76	0.90
AnatomiX-$\hat{S}_t$-$\hat{y}_{box}$	0.78	0.79	0.83	0.95
AnatomiX-$\hat{O}_A$-$\hat{y}_{box}$	0.73	0.75	0.79	0.85
AnatomiX-$\hat{O}_A$-$\hat{S}_t$	<u>0.82</u>	<u>0.85</u>	<u>0.86</u>	**0.95**
AnatomiX-I_p	0.68	0.72	0.72	0.87
AnatomiX	**0.86**	**0.86**	**0.89**	<u>0.95</u>

S5.2. APM

In addition to the ablations conducted for $\mathcal{LM}$, we study the contribution of different architectural components in APM. First, we replace the image encoder $\mathcal{E}$ with a pretrained DINOv3 model, denoted as AnatomiX-Dino. Second, to evaluate the role of the feature extraction module $\mathcal{M}$, we remove it and perform contrastive alignment directly on the decoder output $\mathcal{D}$, resulting in AnatomiX-wo-$\mathcal{M}$. Third, in APM-CLIP, we replace the proposed contrastive self-similarity loss with the standard CLIP loss. We evaluate all variants on object detection (bounding box prediction y_{box}) and sentence retrieval. For detection, we report Intersection over Union (IoU). For retrieval, we compare the retrieved sentence $\hat{S}_t$ with ground-truth text using CheXbert-14-F1, RadGraph-F1, and METEOR.

As shown in Table S6, DINOv3 achieves competitive IoU but underperforms on retrieval metrics, indicating weaker cross-modal alignment despite strong visual representations. Removing $\mathcal{M}$ results in a consistent performance drop across tasks, suggesting that decoupling bounding box prediction from textual feature alignment facilitates more effective learning and improves both anatomical localization and sentence retrieval. Finally, replacing the proposed soft self-similarity loss with the standard CLIP loss degrades performance, highlighting the importance of the tailored contrastive objective in APM. Furthermore, anatomy-wise breakdown of retrieval results in given in Tables S9-S12 (d-g).

Table S6. Ablation experiments for APM discussed in S5.

	y_{box} Metrics	$\hat{S}_t$ Metrics		
Model	IoU	CheXbert-14 F1	RadGraph F1	METEOR
APM-wo-$\mathcal{M}$	0.781	0.627	<u>0.689</u>	<u>0.710</u>
APM-Dino	<u>0.792</u>	0.611	0.685	0.704
APM-CLIP	0.775	**0.640**	0.679	0.692
APM	**0.812**	<u>0.634</u>	**0.709**	**0.727**

S6. Additional Results

Table S7 contains the per dataset results for phrase grounding tasks, while Table S8 provides the per dataset breakdown for grounded diagnosis task. Furthermore, Fig. S7 shows a sample reporet generated using AnatomiX, and Fig. S8 shows sample input-output for VQA and image understanding tasks.

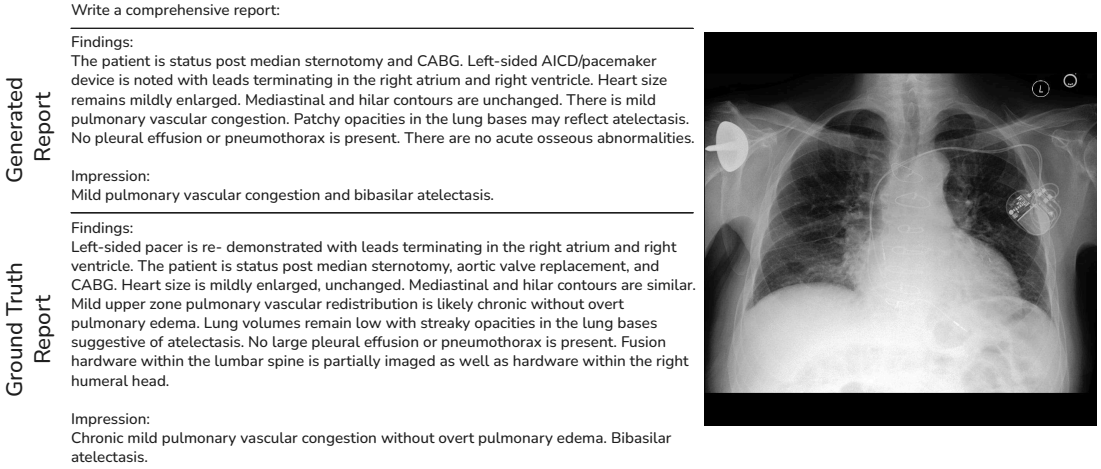

Figure S7. Sample report generation with AnatomiX.

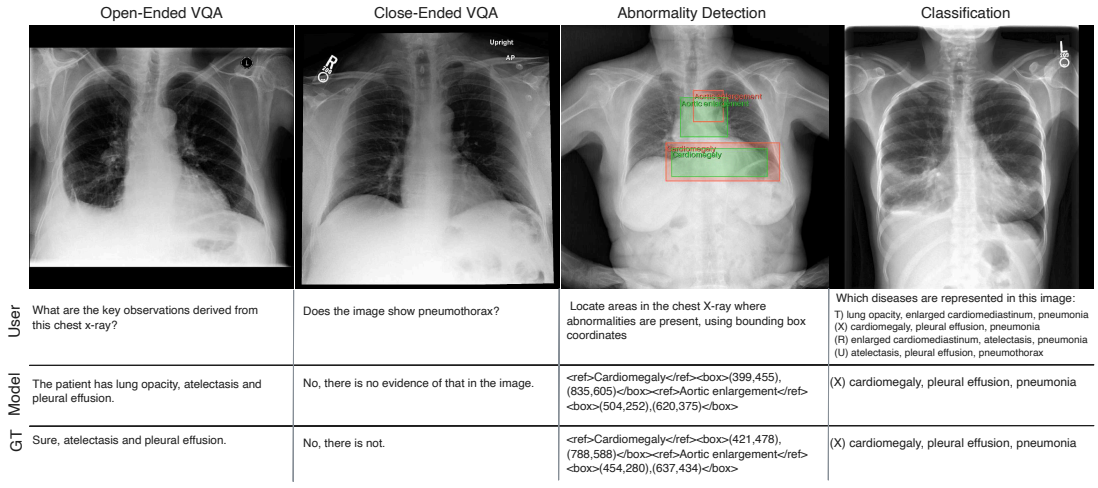

Figure S8. Example input-output-ground truth for image understanding and visual question answering tasks. Box colors: green represents the ground truth while red box shows the model's output.

Table S7. Per dataset performance on Phrase Grounding task.

Dataset	IoU	mAP
MS-CXR	0.532	0.388
PadChest-Gr	0.444	0.337
VinDr-Inst	0.230	0.180

Table S8. Grounded diagnosis performance on different datasets.

Dataset	BERTScore	ROUGE	METEOR	RadGraph-F1	CheXbert-F1
MS-CXR	0.758	0.729	0.402	0.732	0.744
VinDr-Inst	0.606	0.567	0.424	0.543	0.502

Table S9. Per-anatomy results for central anatomical structures with characteristic side predominance. (a) Results for the anatomy grounding task without flipping. (b–c) Anatomy grounding performance comparison between AnatomiX and RadVLM on flipped images. (d) Similarity between retrieved sentences $\hat{S}_t$ and ground-truth sentences in APM.

	Heart	Aortic arch structure	Descending aorta	Superior vena cava
(a) AnatomiX: Anatomy Grounding in Normal Images (no flipping)				
IoU	0.60	0.73	0.76	0.75
mAP	0.71	0.63	0.64	0.65
(b) AnatomiX: Anatomy Grounding in Horizontally Flipped Images				
IoU	0.79	0.73	0.76	0.70
mAP	0.69	0.61	0.64	0.57
(c) RadVLM: Anatomy Grounding in Horizontally Flipped Images				
IoU	0.59	0.05	0.04	0.06
mAP	0.45	0.02	0.00	0.00
(d) APM Sentence Retrieval Results				
CheXbert-14 F1	0.92	0.41	1.00	1.00
RadGraph F1	0.63	0.57	1.00	1.00
METEOR	0.72	0.70	0.99	1.00
(e) APM-wo-$\mathcal{M}$ Sentence Retrieval Results				
CheXbert-14 F1	0.92	0.31	1.00	1.00
RadGraph F1	0.60	0.51	1.00	1.00
METEOR	0.71	0.66	0.99	1.00
(f) APM-Dino Sentence Retrieval Results				
CheXbert-14 F1	0.92	0.38	1.00	1.00
RadGraph F1	0.61	0.57	1.00	1.00
METEOR	0.70	0.68	0.99	1.00
(g) APM-CLIP Sentence Retrieval Results				
CheXbert-14 F1	0.92	0.25	1.00	1.00
RadGraph F1	0.57	0.46	1.00	1.00
METEOR	0.67	0.64	0.99	1.00

Table S10. Anatomy-wise results for left-sided anatomical structures. (a) Results for the anatomy grounding task without flipping. (b–c) Anatomy grounding performance comparison between AnatomiX and RadVLM on flipped images. (d) Similarity between retrieved sentences $\hat{S}_t$ and ground-truth sentences in APM. Note: Some anatomical objects achieve a perfect score of 1.0 due to the limited number of possible sentences.

	Left lung	Left upper lung zone	Left mid lung zone	Left lower lung zone	Apical zone of left lung	Hilar area of left lung	Left costodiaphragmatic recess	Left hemidiaphragm	Left cardiophrenic sulcus	Left clavicle	Left upper abdominal quadrant	Left margin of heart
(a) AnatomiX: Anatomy Grounding in Normal Images (no flipping)												
IoU	0.88	0.86	0.77	0.80	0.81	0.67	0.61	0.74	0.60	0.74	0.79	0.62
mAP	0.81	0.76	0.67	0.70	0.70	0.66	0.51	0.62	0.53	0.63	0.77	0.68
(b) AnatomiX: Anatomy Grounding in Horizontally Flipped Images												
IoU	0.85	0.84	0.73	0.79	0.81	0.78	0.60	0.72	0.60	0.54	0.84	0.80
mAP	0.76	0.73	0.63	0.67	0.70	0.67	0.50	0.61	0.50	0.44	0.74	0.69
(c) RadVLM: Anatomy Grounding in Horizontally Flipped Images												
IoU	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.04
mAP	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
(d) APM Sentence Retrieval Results												
CheXbert-14 F1	0.58	0.11	0.15	0.31	0.32	0.39	0.34	0.85	1.00	0.75	1.00	1.00
RadGraph F1	0.49	0.49	0.45	0.55	0.70	0.57	0.71	0.76	1.00	0.64	1.00	1.00
METEOR	0.54	0.51	0.54	0.59	0.72	0.74	0.72	0.66	1.00	0.52	1.00	1.00
(e) APM-wo-$\mathcal{M}$ Sentence Retrieval Results												
CheXbert-14 F1	0.57	0.06	0.09	0.29	0.59	0.37	0.45	0.71	1.00	0.61	1.00	1.00
RadGraph F1	0.48	0.46	0.42	0.54	0.84	0.58	0.77	0.70	1.00	0.55	1.00	1.00
METEOR	0.52	0.49	0.52	0.59	0.82	0.74	0.75	0.56	1.00	0.44	1.00	1.00
(f) APM-Dino Sentence Retrieval Results												
CheXbert-14 F1	0.55	0.16	0.16	0.28	0.16	0.35	0.34	0.82	1.00	0.59	1.00	1.00
RadGraph F1	0.46	0.45	0.42	0.50	0.74	0.52	0.72	0.69	1.00	0.60	1.00	1.00
METEOR	0.51	0.50	0.52	0.57	0.70	0.73	0.71	0.54	1.00	0.44	1.00	1.00
(g) APM-CLIP Sentence Retrieval Results												
CheXbert-14 F1	0.54	0.05	0.11	0.31	0.54	0.37	0.37	0.84	1.00	0.80	1.00	1.00
RadGraph F1	0.46	0.53	0.45	0.56	0.81	0.54	0.73	0.62	1.00	0.55	1.00	1.00
METEOR	0.50	0.50	0.51	0.58	0.80	0.73	0.72	0.46	1.00	0.46	1.00	1.00

Table S11. Anatomy-wise results for right-sided anatomical structures. (a) Results for the anatomy grounding task without flipping. (b–c) Anatomy grounding performance comparison between AnatomiX and RadVLM on flipped images. (d) Similarity between retrieved sentences $\hat{S}_t$ and ground-truth sentences in APM. Note: Some anatomical objects achieve a perfect score of 1.0 due to the limited number of possible sentences.

	Right lung	Right upper lung zone	Right mid lung zone	Right lower lung zone	Apical zone of right lung	Hilar area of right lung	Right costodiaphragmatic recess	Right hemidiaphragm	Right cardiophrenic sulcus	Right clavicle	Right upper abdominal quadrant	Right atrial structure	Right heart border
(a) AnatomiX: Anatomy Grounding in Normal Images (no flipping)													
IoU	0.89	0.85	0.79	0.79	0.79	0.8	0.67	0.73	0.51	0.75	0.87	0.65	0.59
mAP	0.82	0.76	0.68	0.7	0.68	0.71	0.57	0.63	0.42	0.63	0.77	0.57	0.64
(b) AnatomiX: Anatomy Grounding in Horizontally Flipped Images													
IoU	0.89	0.78	0.74	0.81	0.77	0.79	0.65	0.73	0.47	0.38	0.85	0.61	0.75
mAP	0.81	0.68	0.62	0.7	0.67	0.68	0.55	0.61	0.35	0.27	0.74	0.51	0.63
(c) RadVLM: Anatomy Grounding in Horizontally Flipped Images													
IoU	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01	0.0	0.0	0.01	0.0	0.0
mAP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
(d) APM Sentence Retrieval Results													
CheXbert-14 F1	0.6	0.05	0.1	0.25	0.5	0.4	0.4	0.8	1.0	0.73	1.0	1.0	1.0
RadGraph F1	0.5	0.44	0.43	0.49	0.76	0.54	0.76	0.73	1.0	0.62	1.0	1.0	1.0
METEOR	0.55	0.5	0.53	0.57	0.75	0.74	0.74	0.61	1.0	0.49	1.0	1.0	1.0
(e) APM-wo-$\mathcal{M}$ Sentence Retrieval Results													
CheXbert-14 F1	0.57	0.03	0.09	0.28	0.48	0.39	0.49	0.6	1.0	0.7	1.0	1.0	1.0
RadGraph F1	0.48	0.41	0.39	0.52	0.64	0.55	0.78	0.62	1.0	0.59	1.0	1.0	1.0
METEOR	0.53	0.48	0.52	0.58	0.74	0.74	0.76	0.47	1.0	0.43	1.0	1.0	1.0
(f) APM-Dino Sentence Retrieval Results													
CheXbert-14 F1	0.57	0.05	0.1	0.24	0.35	0.38	0.41	0.63	1.0	0.78	1.0	1.0	1.0
RadGraph F1	0.47	0.47	0.37	0.47	0.74	0.51	0.73	0.61	1.0	0.61	1.0	1.0	1.0
METEOR	0.53	0.5	0.51	0.56	0.72	0.72	0.72	0.43	1.0	0.5	1.0	1.0	1.0
(g) APM-CLIP Sentence Retrieval Results													
CheXbert-14 F1	0.56	0.05	0.09	0.24	0.7	0.4	0.53	0.85	1.0	0.91	1.0	1.0	1.0
RadGraph F1	0.47	0.48	0.4	0.51	0.81	0.53	0.76	0.69	1.0	0.56	1.0	1.0	1.0
METEOR	0.53	0.5	0.51	0.57	0.79	0.73	0.75	0.57	1.0	0.45	1.0	1.0	1.0

Table S12. Anatomy-wise results for midline (central) anatomical structures. (a) Results for the anatomy grounding task without flipping. (b–c) Anatomy grounding performance comparison between AnatomiX and RadVLM on flipped images. (d) Similarity between retrieved sentences $\hat{S}_t$ and ground-truth sentences in APM.

	Trachea & main bronchus	Carina	Mediastinum	Superior mediastinum	Vertebral column	Cavoatrial	Abdominal cavity
(a) AnatomiX: Anatomy Grounding in Normal Images (no flipping)							
IoU	0.74	0.48	0.58	0.77	0.85	0.66	0.86
mAP	0.62	0.40	0.73	0.67	0.74	0.57	0.81
(b) AnatomiX: Anatomy Grounding in Horizontally Flipped Images							
IoU	0.75	0.47	0.70	0.77	0.76	0.38	0.84
mAP	0.64	0.35	0.59	0.66	0.65	0.29	0.77
(c) RadVLM: Anatomy Grounding in Horizontally Flipped Images							
IoU	0.43	0.05	0.60	0.68	0.74	0.00	0.81
mAP	0.31	0.01	0.47	0.56	0.62	0.00	0.67
(d) APM Sentence Retrieval Results							
CheXbert-14 F1	1.00	0.83	0.44	0.06	0.65	0.99	0.89
RadGraph F1	1.00	0.80	0.47	0.29	0.68	0.64	0.81
METEOR	0.99	0.86	0.45	0.44	0.62	0.57	0.82
(e) APM-wo-$\mathcal{M}$ Sentence Retrieval Results							
CheXbert-14 F1	1.00	0.91	0.43	0.09	0.69	0.99	0.88
RadGraph F1	1.00	0.87	0.43	0.18	0.60	0.47	0.80
METEOR	0.99	0.91	0.43	0.45	0.61	0.32	0.83
(f) APM-Dino Sentence Retrieval Results							
CheXbert-14 F1	1.00	0.89	0.40	0.05	0.57	0.99	0.88
RadGraph F1	1.00	0.88	0.44	0.21	0.55	0.50	0.82
METEOR	0.99	0.92	0.42	0.39	0.59	0.45	0.82
(g) APM-CLIP Sentence Retrieval Results							
CheXbert-14 F1	1.00	0.79	0.44	0.02	0.59	0.99	0.78
RadGraph F1	1.00	0.75	0.37	0.22	0.64	0.44	0.58
METEOR	0.99	0.82	0.37	0.43	0.55	0.22	0.59

S7. Attention Visualization

In this section, we visualize the cross-attention weights learned within the feature extraction module $\mathcal{M}$ using the anatomical object tokens O_A as queries and image patch embeddings I_p as keys. Fig. S9 shows that the model focuses on the correct region in the image with high accuracy leading to rich anatomical object tokens $\hat{\mathrm{O}}_A$ and overall anatomical understanding in the downstream tasks.

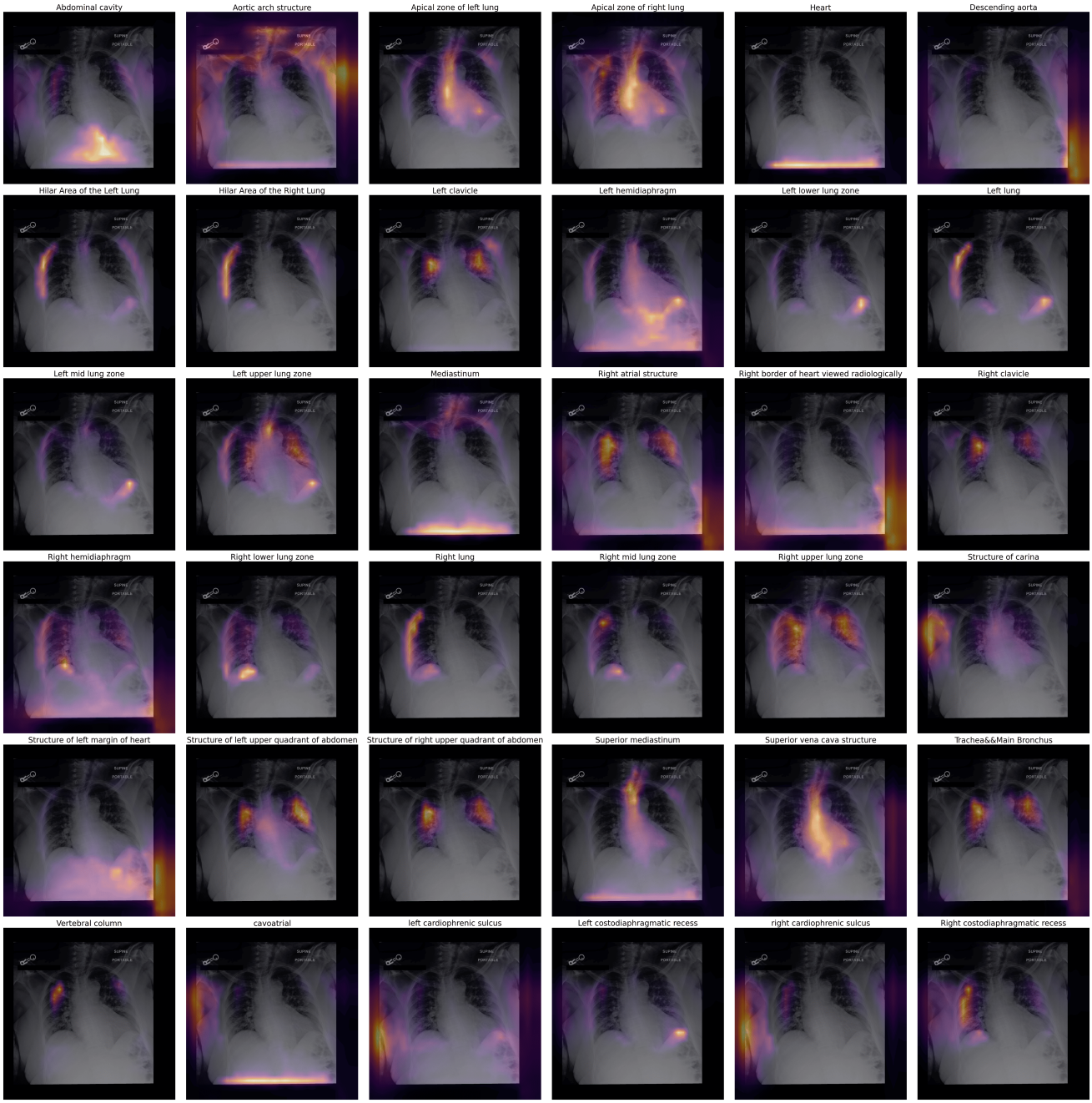

Figure S9. Cross-attention between anatomical object tokens and image embeddings in the feature extraction module $\mathcal{M}$. Subfigure titles indicate the corresponding anatomical object names. Best viewed when zoomed in.