Supplementary for Anatomy-VLM: A Fine-grained Vision-Language Model for Medical Interpretation

1. Zero-shot Disease Classification Table Expanded Version

Table 2 presents the complete zero-shot performance results for chest image disease classification across all 20 individual disease categories, expanding upon the summarized results shown in Table 1 of the main text. This comprehensive evaluation was conducted on a validation set of 6,223 chest images, testing multiple state-of-the-art vision-language models. The anatomical grouping strategy maps individual diseases to seven higher-level categories based on their primary anatomical involvement and pathophysiological mechanisms. Table 1 shows the mapping used to group the attributes and the metric calculation.

Class	Assigned attributes
Lung parenchyma & air–space disease	Consolidation, Lung Opacity, Pulmonary Edema/Hazy Opacity, Lung Lesion, Mass/Nodule (NOS)
Atelectasis & collapse	Atelectasis, Linear/Patchy Atelectasis, Lobar/Segmental Collapse
Pleural space & pleura	Pleural Effusion, Costophrenic Angle Blunting, Pleural/Parenchymal Scarring
Inflation & airway mechanics	Hyperaeration
Diaphragm & sub-diaphragmatic	Elevated Hemidiaphragm
Cardiomediastinal & hilar structures	Enlarged Cardiac Silhouette, Enlarged Hilum, Vascular Congestion, Tortuous Aorta, Vascular Calcification
Musculoskeletal & thoracic cage	Scoliosis, Spinal Degenerative Changes

Table 1. Chest ImaGenome attributes grouped into higher-level anatomical classes.

2. Anatomy to Disease Class Distribution

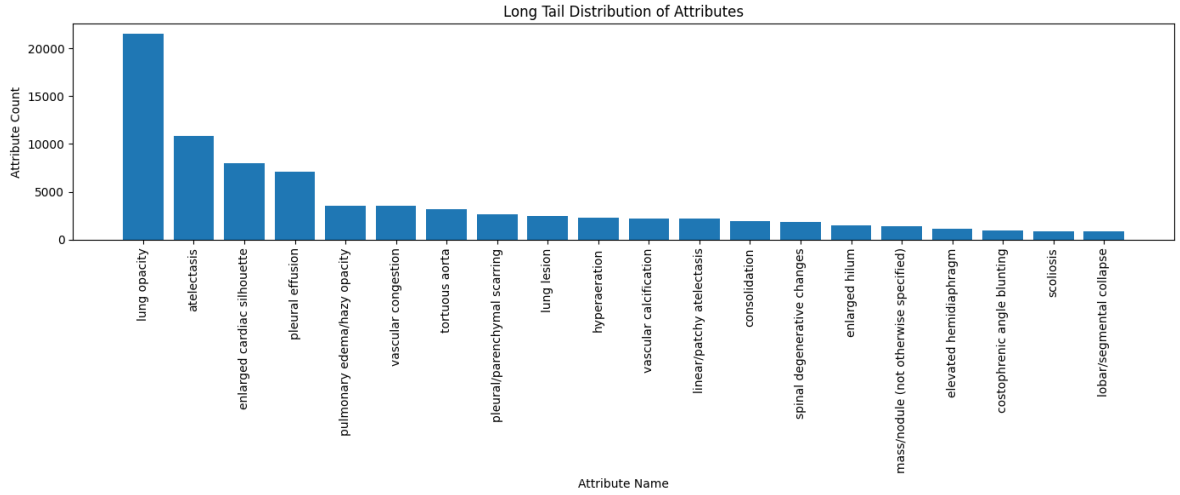
We analyzed the disease class frequency distribution in the Chest ImaGenome dataset, with results presented in Figure 1. The disease class labels were counted from the training set, revealing a long-tail distribution that reflects significant class imbalance across different conditions. Additionally, we examined the label distribution for each anatomical region provided by the Chest ImaGenome dataset, as shown in Figure 2. Note that Cavoatrial Junction, Right Atrium, Carina, and Abdomen are excluded from this analysis as they contain no matching disease findings, similar to the SVC region. The anatomical region analysis confirms that the long-tail distribution persists across different body regions, while also revealing complex disease patterns that provide meaningful training dynamics for Anatomy-VLM development.

Algorithm 1: Contrastive Text–Label Construction

Require: Findings $\mathcal{F} = \{f_i\}_{i=1}^{n}$ for n bounding boxes
Ensure: Sentences $T = \{t_i\}_{i=1}^{n}$, label matrix $L \in \{0,1\}^{n \times n}$

$L \leftarrow \mathbf{0}$; $T \leftarrow \emptyset$
Step 1: *Select one sub-sentence per box*
for $i = 1$ **to** n **do**
  if f_i has sub-sentences **then**
    $t_i \leftarrow$ random sub-sentence of f_i; $L[i,i] \leftarrow 1$
  else
    $t_i \leftarrow \epsilon$
  end if
end for
Step 2: *Perturb a random 20% of current sample*
for each i with $L[i,i] = 1$ **do**
  if rand() < 0.2 **then**
    optionally negate or rephrase t_i and flip $L[i,i]$ if negated
  end if
end for
Step 3: *Fill empty slots with attribute negatives*
$\mathcal{A} \leftarrow$ attributes found in positive T
for each i with $t_i = \epsilon$ **and** $\mathcal{A} \neq \emptyset$ **do**
  $t_i \leftarrow$ positive or negated sentence using random $a \in \mathcal{A}$; $L[i,i] \leftarrow 0$
end for
Step 4: *Mark duplicates*
for $i < j$ **do**
  if $t_i = t_j$ **then**
    increment $L[i,j]$ and $L[j,i]$
  end if
end for

Figure 1. Top 20/45 disease classes ranked to show a long tail distribution for Chest ImaGenome dataset.



Attribute	Zero-shot vision–language models															Black-box baselines						Concept-aligned (Ours)		
	CLIP			BioMedCLIP			BioViL			MedKLIP			CARZero			ResNet			ViT			Ours		
	BMAC	AUC	F1	BMAC	AUC	F1	BMAC	AUC	F1	BMAC	AUC	F1	BMAC	AUC	F1	BMAC	AUC	F1	BMAC	AUC	F1	BMAC	AUC	F1
Atelectasis	53.0	53.0	27.2	48.1	48.1	9.0	67.3	67.3	49.8	54.1	57.0	53.7	59.6	86.4	33.1	60.1	69.9	68.1	73.1	79.7	56.6	73.4	80.8	57.4
Consolidation	51.5	51.5	6.8	53.7	53.7	11.9	77.8	77.8	21.6	58.5	65.7	42.4	65.2	87.3	32.6	63.0	68.0	45.9	63.8	76.8	25.8	73.4	82.5	26.3
Costophrenic Angle Blunting	50.2	50.2	3.2	49.9	49.9	0.0	64.0	64.0	6.0	52.4	66.3	8.8	66.5	83.4	15.5	59.6	63.3	12.7	63.0	74.3	10.5	59.3	78.4	13.8
Elevated Hemidiaphragm	49.9	49.9	0.0	52.7	52.7	7.5	63.2	63.2	7.1	58.7	65.6	4.6	77.1	88.3	40.6	55.7	63.5	8.0	72.5	85.8	35.6	66.1	84.8	32.9
Enlarged Cardiac Silhouette	50.0	50.0	0.4	52.9	52.9	14.0	75.6	75.6	52.9	51.3	66.9	4.6	65.4	91.8	46.3	54.6	63.9	9.1	76.2	87.3	59.4	77.9	89.3	61.7
Enlarged Hilum	48.1	48.1	5.6	50.1	50.1	1.0	69.2	69.2	10.7	50.0	59.8	29.7	61.6	82.9	23.2	71.1	79.8	49.3	60.8	70.2	17.4	62.7	75.2	15.5
Hyperaeration	50.0	50.0	10.0	50.1	50.1	0.6	58.4	58.4	12.0	50.0	43.4	5.9	78.1	90.6	45.4	56.8	56.7	8.2	71.0	87.4	39.7	74.3	89.8	46.3
Linear/Patchy Atelectasis	49.7	49.7	3.9	48.9	48.9	2.4	57.3	57.3	11.5	41.7	37.3	7.0	63.4	75.4	20.5	57.4	64.2	16.9	60.0	66.7	15.9	66.4	72.1	18.2
Lobar/Segmental Collapse	50.7	50.7	3.5	63.1	63.1	13.7	73.3	73.3	9.7	55.9	57.5	11.6	65.7	88.6	25.1	59.8	60.9	12.6	59.9	84.7	22.1	61.6	87.0	22.3
Lung Lesion	50.9	50.9	10.7	52.0	52.0	8.7	59.1	59.1	13.8	50.0	54.7	3.9	57.0	69.0	19.7	58.3	69.0	9.9	57.0	62.6	16.4	59.0	66.6	17.4
Lung Opacity	49.0	49.0	64.2	54.2	54.2	17.0	74.9	74.9	69.3	50.0	59.5	10.6	55.0	87.4	18.7	51.4	49.6	10.8	76.9	85.1	77.5	77.5	86.2	78.3
Mass/Nodule (Not Otherwise Specified)	50.4	50.4	5.9	50.8	50.8	5.5	60.1	60.1	7.7	50.0	50.5	5.8	64.2	78.9	27.3	52.8	50.8	6.2	55.7	63.7	12.0	59.8	70.6	12.6
Pleural Effusion	50.2	50.2	27.8	52.2	52.2	8.6	81.4	81.4	59.8	50.2	65.7	28.0	62.6	93.9	40.0	71.2	78.6	46.4	82.9	90.8	66.5	80.8	91.4	67.1
Pleural/Parenchymal Scarring	50.0	50.0	11.4	50.3	50.3	3.7	57.6	57.6	13.9	50.7	54.3	11.5	57.6	80.2	23.1	56.2	57.9	13.5	61.3	69.6	20.3	67.4	75.2	24.2
Pulmonary Edema/Hazy Opacity	58.2	58.2	18.1	66.6	66.6	36.4	78.2	78.2	41.7	56.0	59.6	17.5	61.9	91.3	36.4	63.2	69.7	25.0	72.7	85.7	46.1	77.5	88.9	48.4
Scoliosis	50.0	50.0	4.5	50.2	50.2	1.9	48.1	48.1	4.1	49.2	46.6	4.4	74.2	83.2	27.6	58.2	63.7	11.3	62.5	80.0	31.1	60.8	78.1	22.0
Spinal Degenerative Changes	50.0	50.0	0.0	50.0	50.0	0.0	56.1	56.1	8.7	50.0	41.3	7.5	64.7	82.8	21.7	62.4	67.5	14.1	60.6	71.9	14.1	67.0	75.5	16.4
Tortuous Aorta	50.0	50.0	14.4	49.0	49.0	4.5	63.8	63.8	20.2	50.2	60.7	14.4	77.7	89.7	44.9	62.6	71.9	26.3	70.6	84.3	38.4	68.2	83.0	36.7
Vascular Calcification	50.0	50.0	0.0	50.0	50.0	0.0	66.0	66.0	16.0	50.0	49.8	10.2	72.5	89.1	38.4	58.8	64.9	17.3	71.0	81.8	28.9	66.8	83.9	30.2
Vascular Congestion	49.1	49.1	3.1	52.9	52.9	12.7	70.9	70.9	28.8	50.0	51.0	13.9	65.6	86.7	36.2	61.4	66.8	21.7	68.6	82.0	33.6	73.6	85.4	36.1
Average	50.6	50.6	11.0	52.4	52.4	8.0	66.1	66.1	23.3	51.4	55.7	14.8	65.8	85.4	30.8	59.7	65.0	21.7	67.0	78.5	33.4	68.7	81.2	34.2

Table 2. Chest ImaGenome zero-shot global disease-classification (Full result).

3. Fine-grained Contrastive Learning Label Generation Algorithm

We provide exact algorithmic procedure in algorithm 1 to detail the systematic procedure for generating fine-grained contrastive learning labels from radiological findings and their corresponding bounding boxes. The algorithm transforms input findings for n bounding boxes into sentences with corresponding label matrix through a four-step process designed to address key challenges in medical image-text alignment. The procedure begins by selecting representative sentences from available findings for each bounding box, establishing positive correspondences in the label matrix. To create meaningful contrastive pairs, the algorithm then generates negative samples by randomly selecting a subset of positive assignments and applying semantic perturbations such as negation or rephrasing. For bounding boxes lacking specific findings, the algorithm fills empty slots using a predefined attribute set derived from positive training examples, ensuring comprehensive coverage of radiological vocabulary. Finally, the algorithm maintains consistency by identifying duplicate sentences and ensuring they receive identical labels across all positions. This systematic approach captures the nuanced relationship between radiological findings and their textual descriptions while maintaining clinical validity. The algorithm's design reflects the complexity of medical image interpretation, where precise terminology and semantic understanding are crucial for accurate diagnosis. By generating both positive and negative examples with appropriate label assignments, the contrastive learning framework can effectively learn to distinguish between different radiological conditions and their corresponding textual descriptions, ultimately improving the performance of vision-language models in medical imaging applications.

Table 3. Zero-shot region-wise validation (bounding boxes) on Chest ImaGenome dataset

Region	**BMAC**	**AUC**	**F1**
Right lung	98.1	98.1	66.2
Right upper lung zone	84.3	90.1	6.3
Right mid lung zone	95.4	96.8	13.8
Right lower lung zone	94.2	95.5	15.8
Right hilar structures	97.8	98.5	58.4
Right apical zone	97.5	97.7	24.4
Right costophrenic angle	96.6	96.7	47.2
Right hemidiaphragm	92.4	93.8	5.5
Left lung	98.0	98.2	65.7
Left upper lung zone	84.9	90.6	4.5
Left mid lung zone	95.3	96.6	16.0
Left lower lung zone	93.9	95.4	17.7
Left hilar structures	97.9	98.4	58.5
Left apical zone	97.4	98.0	24.8
Left costophrenic angle	96.7	96.8	47.7
Left hemidiaphragm	90.8	93.1	2.8
Trachea	94.7	95.9	5.2
Spine	99.1	99.6	96.4
Right clavicle	98.2	99.3	32.9
Left clavicle	98.1	99.3	33.2
Aortic arch	91.7	94.7	6.7
Mediastinum	98.2	98.5	64.8
Upper mediastinum	96.9	97.9	48.8
SVC	92.9	94.4	2.6
Cardiac silhouette	99.7	99.7	98.2
Cavoatrial junction	88.1	92.4	0.4
Right atrium	90.6	91.6	1.3
Carina	94.5	96.3	2.3
Abdomen	94.1	94.2	11.2
Average	**94.8**	**96.1**	**30.3**

Figure 2. 25 anatomies to disease class fine-grained learning distribution for Chest ImaGenome dataset.