Knowledge to Sight: Reasoning over Visual Attributes through Knowledge Decomposition for Abnormality Grounding

Supplementary Material

Jun Li[1,2] Che Liu[3] Wenjia Bai[3] Mingxuan Liu[4]
Rossella Arcucci[3] Cosmin I. Bercea[1,2*] Julia A. Schnabel[1,2,5,6*]

[1] Technical University of Munich [2] Munich Center for Machine Learning [3] Imperial College London
[4] University of Trento [5] Helmholtz Munich [6] King's College London

Contents

A. Dataset Details

In this section, we provide detailed statistics and data characteristics of the two datasets used in our experiments: VinDr-CXRR [5] and PadChest-GR [2]. This supplement presents additional information relevant to our evaluation.

A.1. Details of VinDr-CXR

We chose VinDr-CXR as the training source because it is annotated by experienced radiologists. We use all available image–abnormality pairs, resulting in a total of 18,195 samples covering 22 distinct thoracic findings. Each image may contain one or more annotated abnormalities. To address annotation inconsistencies across multiple radiologists, we adopt the weighted box fusion strategy [4], following prior work [1, 3], to merge overlapping bounding boxes into a unified set per abnormality per image. We follow the official split provided by the dataset, resulting in a training set of 16,087 samples and a test set of 2,108 samples. Figure 1 shows the distribution of the dataset in our experiments.

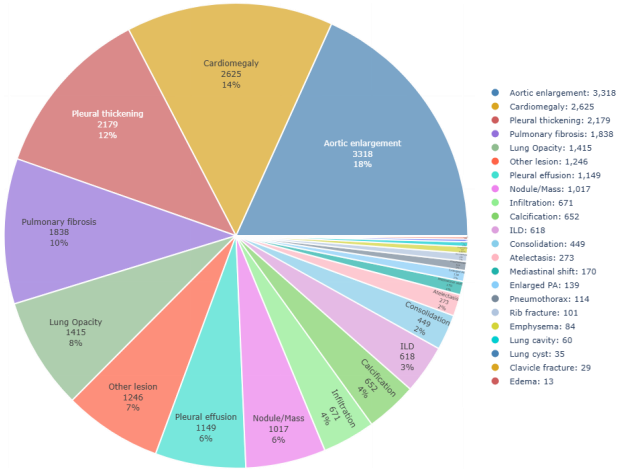

Figure 1. Distribution of the 22 abnormality types in the VinDr-CXR dataset. The dataset exhibits a long-tailed distribution. The dataset is split into a training set of 16,087 samples and a held-out test set of 2,108 samples.

A.2. Details of PadChest-GR

PadChest-GR [2] is used exclusively for evaluation. We construct two evaluation subsets from its official test split (covering 24 abnormality types): one for standard zero-shot generalization and another for out-of-distribution (OOD) evaluation. The zero-shot subset contains 641 image–abnormality pairs spanning 6 types, labeled with **known** categories that are also present in VinDr-CXR. The OOD subset includes 644 image–abnormality pairs across

*Shared senior authors.

18 types labeled with **unknown** categories, which are not seen during training. This split enables controlled evaluation of both generalization within known classes and robustness to novel, unseen concepts. Based on semantic and clinical correspondence with VinDr-CXR, we categorize 6 PadChest-GR labels as known classes (e.g., *cardiomegaly*, *atelectasis*, *nodule*), while the remaining 18 (e.g., *scoliosis*, *aortic atheromatosis*, *electrical device*) are treated as unknown. Figure 2 illustrates the distribution of the PadChest-GR categories, highlighting the long-tailed nature of both known and unknown subsets.

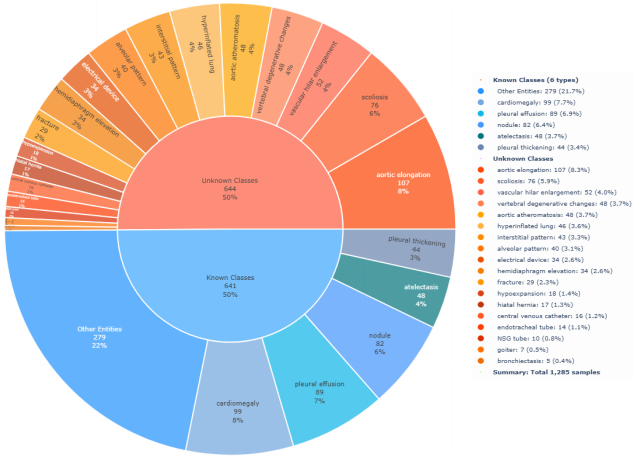

Figure 2. Distribution of the 24 abnormality types in the PadChest-GR dataset. Known classes are mapped to VinDr-CXR labels, while the rest are treated as unknown for OOD evaluation.

A.3. Annotation Preprocess for Florence-2

We follow the official formatting protocol of Florence-2 for converting bounding box annotations into language-based location prompts and answers.

Each bounding box is represented in the standard format as pixel coordinates (x_1, y_1, x_2, y_2) on an image of width W and height H. These coordinates are first normalized to the range $[0, 1000]$ using the following transformation:

$$\texttt{loc\_x} = \left\lfloor \frac{x}{W} \times 1000 \right\rfloor, \quad \texttt{loc\_y} = \left\lfloor \frac{y}{H} \times 1000 \right\rfloor, \quad (1)$$

where $\lfloor \cdot \rfloor$ denotes the floor operation to obtain discrete integer values. The normalized bounding box is then represented as a sequence of location tokens:

$$\texttt{Label}\ \langle \texttt{loc\_x1} \rangle \langle \texttt{loc\_y1} \rangle \langle \texttt{loc\_x2} \rangle \langle \texttt{loc\_y2} \rangle \quad (2)$$

During training and evaluation, each image–annotation pair is formatted into a prompt–answer style. The prompt queries the presence and location of abnormalities, and the answer consists of one or more spatial tokens corresponding to the bounding boxes.

Florence-2 (base) pair

Prompt: Locate disease {Disease}.
Answer:
Disease ⟨loc_145⟩ ⟨loc_300⟩ ⟨loc_812⟩ ⟨loc_940⟩
Disease ⟨loc_201⟩ ⟨loc_322⟩ ⟨loc_715⟩ ⟨loc_850⟩

In our K2Sight framework, we further decompose the underlying knowledge associated with each disease entity and extract visual-oriented descriptions. These include concise phrases describing shape, texture, and anatomical location for each abnormality. The final prompts are augmented with these definitions to enhance model during training. The training pairs follow the format below:

K2Sight-Lite Pair

Prompt: Locate disease {Disease}, which means {attribute-based description}.
Answer:
Disease ⟨loc_145⟩ ⟨loc_300⟩ ⟨loc_812⟩ ⟨loc_940⟩
Disease ⟨loc_201⟩ ⟨loc_322⟩ ⟨loc_715⟩ ⟨loc_850⟩

A.4. Annotation Preprocess for Qwen2-VL-Instruct

We follow the official guideline of Qwen2-VL-2B-Instruct for converting bounding boxes into text pairs. Similar to Florence-2, each bounding box is expressed as (x_1, y_1, x_2, y_2) and normalized by dividing each coordinate by the image width or height and multiplying by 1000:

$$\texttt{new\_x} = \left\lfloor \frac{x}{W} \times 1000 \right\rfloor, \quad \texttt{new\_y} = \left\lfloor \frac{y}{H} \times 1000 \right\rfloor. \quad (3)$$

Unlike Florence-2, Qwen2-VL-Instruct directly utilizes 2D bounding box coordinates in the output and does not require special format. The prompt structure is derived from the official Qwen2-VL-Instruct repository, where bounding box annotations are requested in JSON format.

Qwen2-VL (base) Pair

Prompt:
Return bounding boxes of 'Disease' areas as JSON format:

```
[{"bbox_2d": [x1, y1, x2, y2], "label": "
    label"}, ...]
```

Answer:

```
[{"bbox_2d": [276, 141, 484, 218], "label
    ": "Disease"},
{"bbox_2d": [552, 127, 767, 230], "label"
    : "Disease"}]
```

In our K2Sight, we enrich the original prompt–answer structure by appending the attribute-based description at the end. The format of the training pairs is shown below:

K2Sight-Base Pair

Prompt:
Return bounding boxes of 'Disease' areas as JSON format:

```
[{"bbox_2d": [x1, y1, x2, y2], "label": "
    label"}, ...]
```

Note: {attribute-based description}

Answer:

```
[{"bbox_2d": [276, 141, 484, 218], "label
    ": "Disease"},
{"bbox_2d": [552, 127, 767, 230], "label"
    : "Disease"}]
```

B. Further Implementation Details of K2Sight

B.1. K2Sight Framework Pseudocode

We further outline the complete pseudocode of the K2Sight framework. It consists of two core stages. In the first stage, the Knowledge Decomposition Constructor extracts and distills clinical definitions into prompts that are visually informative. The second stage, Semantic-Guided Fine-Tuning, fine-tunes vision-language models using these prompts to improve localization accuracy.

Algorithm 1: K2Sight Framework

Input: Abnormality definitions $\{d(a)\}_{a\in A}$; training dataset $D_{\text{train}} = \{(I_i, B_i)\}$
Output: Our K2Sight model M

Stage 1: Knowledge Decomposition Constructor
foreach $a \in A$ **do**
  Retrieve textual definition $d(a)$ from medical sources to get the medical definition;
  Generate structured prompt $\pi(a, d(a))$ incorporating visual attributes (shape, intensity, density, location);
  Sample $N = 5$ candidate visual descriptions $\{\tilde{k}_i(a)\}_{i=1}^{N}$ using GPT-4o with controlled decoding;
  Select best candidate $k(a) \leftarrow \text{HumanSelect}(\{\tilde{k}_i(a)\})$;
end
Construct prompt dictionary $K = \{a \mapsto k(a)\}$;
Stage 2: Semantic-Guided Fine-Tuning
foreach $(I_i, B_i) \in D_{train}$ **do**
  foreach *abnormality a in* B_i **do**
    Format instruction using $k(a)$ and convert bounding boxes to token sequence Y;
    Train M to predict Y given image I_i and prompt $k(a)$ via cross-entropy loss;
  end
end
return M

B.2. Clinical Definition Collection

We collect textual definitions for each abnormality class from authoritative radiology resources. Specifically, we extract formal descriptions from the official documentation of VinDr-CXR [5] and publicly available entries on Radiopaedia [6], a widely used collaborative radiology reference. All definitions are used under the Creative Commons BY-NC-SA 3.0 license for research purposes. Each definition is manually verified to ensure medical accuracy, visual interpretability, and alignment with radiographic appearance.

B.3. Visual Attribute Extraction Parameters

We use GPT-4o for visual attribute extraction in a zero-shot setting. For each abnormality definition, we generate $N = 5$ candidate descriptions per class using the following decoding configuration: temperature = 0.7, top-p = 0.7, repetition penalty = 1.1, and a maximum output length of 1024 tokens. Each generation is conditioned on a structured prompt template (see Sec. 3.1 in the main paper), which encourages the model to focus on four core visual aspects: shape, intensity, appearance pattern, and anatomical location. After generation, minimal text cleaning is applied to remove formatting artifacts or incomplete fragments. Each candidate set is then reviewed by a domain-aware annotator, who selects the most visually faithful, semantically aligned, and concise description. Candidates that are speculative, overly verbose, or dependent on latent clinical information are discarded. The final attribute-grounded prompts are consistently used across training and inference for all models in the K2Sight framework. Full lists of the selected descriptions for VinDr-CXR and PadChest-GR abnormality classes are provided in Table 3 and Table 4, respectively.

B.4. Training Details

Florence-2 Base and K2Sight Lite. Both models are trained on the VinDr-CXR training set, which contains 16,087 annotated image–abnormality pairs. To ensure a controlled comparison and to isolate the effect of knowledge-enhanced instructions, we adopt identical training configurations for both models. The fine-tuning process spans 20 epochs, using the AdamW optimizer with an initial learning rate of 3×10^{-6} and a weight decay coefficient of 0.01 to prevent overfitting. Training is executed on two A100 GPUs with a batch size of 16 per device (resulting in an effective batch size of 32), and distributed data parallelism is enabled via PyTorch Lightning. We employ mixed-precision (FP16) training for compute efficiency.

Qwen2-VL 3B and K2Sight Base. The Qwen2-VL (base) and our K2Sight Base model are both fully fine-tuned, meaning that the vision encoder, multimodal projector, and language decoder are jointly updated during training. Both variants use the same training configuration to ensure direct comparability; the only difference lies in the input format, in which K2Sight uses structured prompts incorporating attribute-based descriptions, while the baseline version uses only raw class names. Fine-tuning is conducted using DeepSpeed ZeRO-3 optimization on 4×A100 GPUs to enable efficient large-scale training with a reduced memory footprint. Each model is trained for 20 epochs, with a per-device batch size of 8 and a gradient accumulation step size of 8, yielding a total effective batch size of 256. The learning rate is initialized at 3×10^{-5} and follows a cosine decay schedule with a warmup ratio of 0.1. All training runs use bfloat16 precision to accelerate computation.

B.5. Inference Settings

To ensure a standardized and unbiased evaluation, we adopt a uniform decoding configuration across all model variants. In particular, we use greedy decoding with the temperature fixed at 0.0, disabling all stochastic sampling strategies such as top-k sampling. This deterministic decoding approach ensures consistency in output across runs and models. The maximum output length is set to 1024 tokens for both Florence-2 and Qwen2-VL variants, which is sufficient to accommodate all predicted bounding-box sequences. All inference experiments are conducted on a single A100 GPU per model to ensure resource parity. During evaluation, we use a batch size of 32, which provides a good trade-off between speed and memory usage while maintaining consistent throughput across all models.

C. Experiments

C.1. Comparison Model Checkpoints

All comparision models are using their publicly available checkpoints. These include general-purpose models such as Qwen2-VL [7] and InternVL3 [8], as well as domain-specific baselines like RadVLM [3] and MAIRA-2 [1]. All evaluations are conducted in a zero-shot or fine-tuned setting as described in the main experiments.

- **Qwen2-VL-2B-Instruct**: https://huggingface.co/Qwen/Qwen2-VL-2B-Instruct
- **Qwen2-VL-7B-Instruct**: https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct
- **InternVL3-2B**: https://huggingface.co/OpenGVLab/InternVL3-2B
- **RadVLM (7B)**: https://huggingface.co/KrauthammerLab/RadVLM

Table 1. Further evaluation for generalist models with our visual-oriented enhanced prompt. All settings are the same as in the main paper Sec. 4.1.

Model	VinDr-CXR			PadChest-GR		
	mAP_{30}	mAP_{50}	mAP_{75}	mAP_{30}	mAP_{50}	mAP_{75}
Qwen2-VL-7B	**2.27**	**1.07**	**0.02**	**1.91**	**0.26**	**0.06**
Qwen2-VL-2B	0.19	0.04	0.01	0.25	0.08	0.00
InternVL3-8B	0.18	0.03	0.00	0.64	0.24	0.00
InternVL3-2B	0.15	0.02	0.00	0.47	0.12	0.00

- **MAIRA-2 (13B)**: https://huggingface.co/microsoft/maira-2

C.2. Attribute-Conditioned Ablation Study

For the ablation study described in Section 4.2 of the main paper, we perform attribute-conditioned prompt masking to investigate the individual contribution of four canonical visual attributes: shape, density, intensity, and location. Specifically, we construct curated lexicons for each attribute category and remove all matched terms from the structured prompts at inference time. This simulates the absence of a specific attribute while keeping the rest of the prompt intact. For detailed lists of attribute-specific terms used in the masking procedure, please refer to Table C.4.

C.3. Further Results on Generalist VLMs with Our Enhanced Prompts

In this section, we further evaluate the performance of generalist VLMs using our enhanced visual-oriented prompts, distilled from medical knowledge via our K2Sight framework. Table 1 presents the evaluation results. Compared to directly using zero-shot prompts with only disease entity names, the performance improves slightly in some cases. For example, Qwen2-VL-7B improves from 1.52 to 2.27 in mAP_{30} and from 0.48 to 1.07 in mAP_{50} on VinDR-CXR dataset. However, some models show a performance drop. For instance, InternVL3-8B decreases from 0.24 to 0.18 in mAP_{30}, and from 0.04 to 0.03 in mAP_{50}.

Overall, relying solely on prompt engineering without further incorporating such descriptive prompts into semantically guided training is still insufficient to elevate generalist models to the performance level of medical specialists.

C.4. More Visualization Results

We provide additional qualitative results comparing K2Sight variants with MAIRA-2 and RadVLM. As shown in Figure 3 and Figure 4, K2Sight consistently produces more accurate and compact localizations, with higher alignment to radiologist-annotated regions. These examples further illustrate the benefit of attribute-guided prompts in improving grounding fidelity, especially for subtle or structurally complex abnormalities.

Table 2. Term sets used to mask individual visual attributes during inference-time ablation.

Attribute	Terms Used for Masking
Shape	round, oval, circular, spherical, elliptical, triangular, rectangular, linear, curved, straight, irregular, lobulated, spiculated, nodular, stellate, mass-like, lump-like, reticular, honeycomb, septal, branching, wedge-shaped, crescentic, patchy, diffuse, borders, contour, outline, edge, pattern, irregularity
Density	dense, solid, soft-tissue, fluid, liquid, gas, air-filled, air-containing, fat-density, calcified, calcific, ossified, consolidated, radiopaque, radiolucent, sclerotic, fibrotic, thick, thin, firm, density
Intensity	bright, white, hyperdense, hyperintense, high-signal, dark, black, hypodense, hypointense, low-signal, gray, greyish, hazy, faint, subtle, opaque, lucent, transparent, prominent, clear, ground-glass, increased, decreased, reduced, diminished
Location	within the lung, pleural cavity, pleural space, pulmonary artery, lung tissue, lung fields, in the lung, supradiaphragmatic, intrathoracic, extrathoracic, paramediastinal, paravertebral, mediastinum, mediastinal, costophrenic, retrocardiac, peripheral, perihilar, subpleural, unilateral, bilateral, central, aorta, heart, basal, posterior, anterior, ventral, dorsal, apical, middle, lower, upper, medial, lateral, right, left

References

[1] Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, et al. Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*, 2024.

[2] Daniel C 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. *arXiv preprint arXiv:2411.05085*, 2024.

[3] Nicolas Deperrois, Hidetoshi Matsuo, Samuel Ruipérez-Campillo, Moritz Vandenhirtz, Sonia Laguna, Alain Ryser, Koji Fujimoto, Mizuho Nishio, Thomas M Sutter, Julia E Vogt, et al. RadVLM: A multitask conversational vision-language model for radiology. *arXiv preprint arXiv:2502.03333*, 2025.

[4] Philip Müller, Georgios Kaissis, and Daniel Rueckert. Chex: Interactive localization and region description in chest X-ray. In *European Conference on Computer Vision*, pages 92–111. Springer, 2024.

[5] 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.

[6] Radiopaedia.org. Radiomics. https://radiopaedia.org/, 2023.

[7] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[8] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Table 3. Final attribute-based visual descriptions for each abnormality in VinDr-CXR.

Abnormality	**Attribute-based Visual Description**
Aortic Enlargement	Widening of the aorta visible as an enlarged artery on imaging.
Atelectasis	Collapsed lung tissue causing darkened or shrunken areas in the lung.
Cardiomegaly	Enlargement of the heart seen when the heart appears larger than normal.
Calcification	Calcium deposits in lung tissue visible as bright white spots.
Clavicle Fracture	A break in the collarbone seen as a gap or irregularity in the bone.
Consolidation	Lung tissue filled with fluid or cells causing dense solid areas on imaging.
Edema	Fluid accumulation in the lungs creating a hazy or clouded area.
Emphysema	Enlarged air spaces in the lungs appearing over-expanded or damaged.
Enlarged Pulmonary Artery	Widening of the pulmonary artery seen as an enlarged artery in the chest.
Interstitial Lung Disease (ILD)	Scarring or inflammation of the lung's interstitial tissue creating a reticular or nodular pattern.
Infiltration	Accumulation of substances or cells in the lung tissue visible as increased density or nodules.
Lung Cavity	Air-filled spaces within the lung often surrounded by dense tissue.
Lung Cyst	Fluid-filled spaces in the lung often round with thin walls.
Lung Opacity	An area of increased density in the lung fields typically appearing as a white or grayish patch.
Mediastinal Shift	Displacement of central chest structures like the heart to one side.
Nodule / Mass	A growth or lump in the lung which may appear as a well-defined or irregular shape.
Pulmonary Fibrosis	Scarring of the lung tissue creating a dense fibrous appearance.
Pneumothorax	Air trapped in the pleural space creating a gap or absence of lung tissue.
Pleural Thickening	Increased thickness of the pleura seen as a dense layer around the lung.
Pleural Effusion	Excess fluid in the pleural space appearing as a shadow around the lungs.
Rib Fracture	A break in one or more ribs appearing as a visible crack or displacement.
Other Lesion	An unusual mass or area in the lung with irregular borders or density.

Table 4. Final attribute-based visual descriptions for each abnormality in PadChest-GR.

Abnormality	**Attribute-based Visual Description**
Pleural Thickening	Increased thickness of the pleura seen as a dense layer around the lung.
Atelectasis	Collapsed lung tissue causing darkened or shrunken areas in the lung.
Pleural Effusion	Excess fluid in the pleural space appearing as a shadow around the lungs.
Cardiomegaly	Enlargement of the heart seen when the heart appears larger than normal.
Aortic Elongation	Lengthened and tortuous aorta, visible as an elongated curving structure.
Vertebral Degenerative Changes	Irregular vertebral margins with bony sclerosis and osteophytes.
Aortic Atheromatosis	Calcified deposits in the aortic wall appearing as bright, irregular opacities.
Nodule	A growth or lump in the lung which may appear as a well-defined or irregular shape.
Alveolar Pattern	Cloud-like, patchy opacities representing fluid or cellular accumulation in alveoli.
Hiatal Hernia	A soft-tissue mass or air-fluid level above the diaphragm, near the midline.
Scoliosis	Sideways curvature of the spine causing misalignment of vertebral bodies.
Hemidiaphragm Elevation	One side of the diaphragm appearing higher than the other, with convex shape.
Hyperinflated Lung	Abnormally increased lung volume with expanded air spaces.
Interstitial Pattern	Fine reticular or nodular opacities spread across the lung, indicating interstitial involvement.
Fracture	A break in the bone appearing as a radiolucent line or displacement.
Vascular Hilar Enlargement	Increased prominence of the pulmonary vessels near the lung hila.
NSG Tube	A thin radiopaque tube extending from the nasal cavity into the stomach.
Endotracheal Tube	A thin or opaque line in the middle of the trachea.
Hypoexpansion	Reduced lung inflation with increased density and narrow intercostal spaces.
Central Venous Catheter	A visible line inside large vein.
Electrical Device	A dense, well-defined metallic opacity, typically a pacemaker or defibrillator.
Bronchiectasis	Dilated bronchi with thick walls, appearing as tubular or cystic opacities.
Goiter	A soft tissue mass in the anterior neck, sometimes displacing the trachea.
Other lesions	An unusual mass or area in the lung with irregular borders or density.

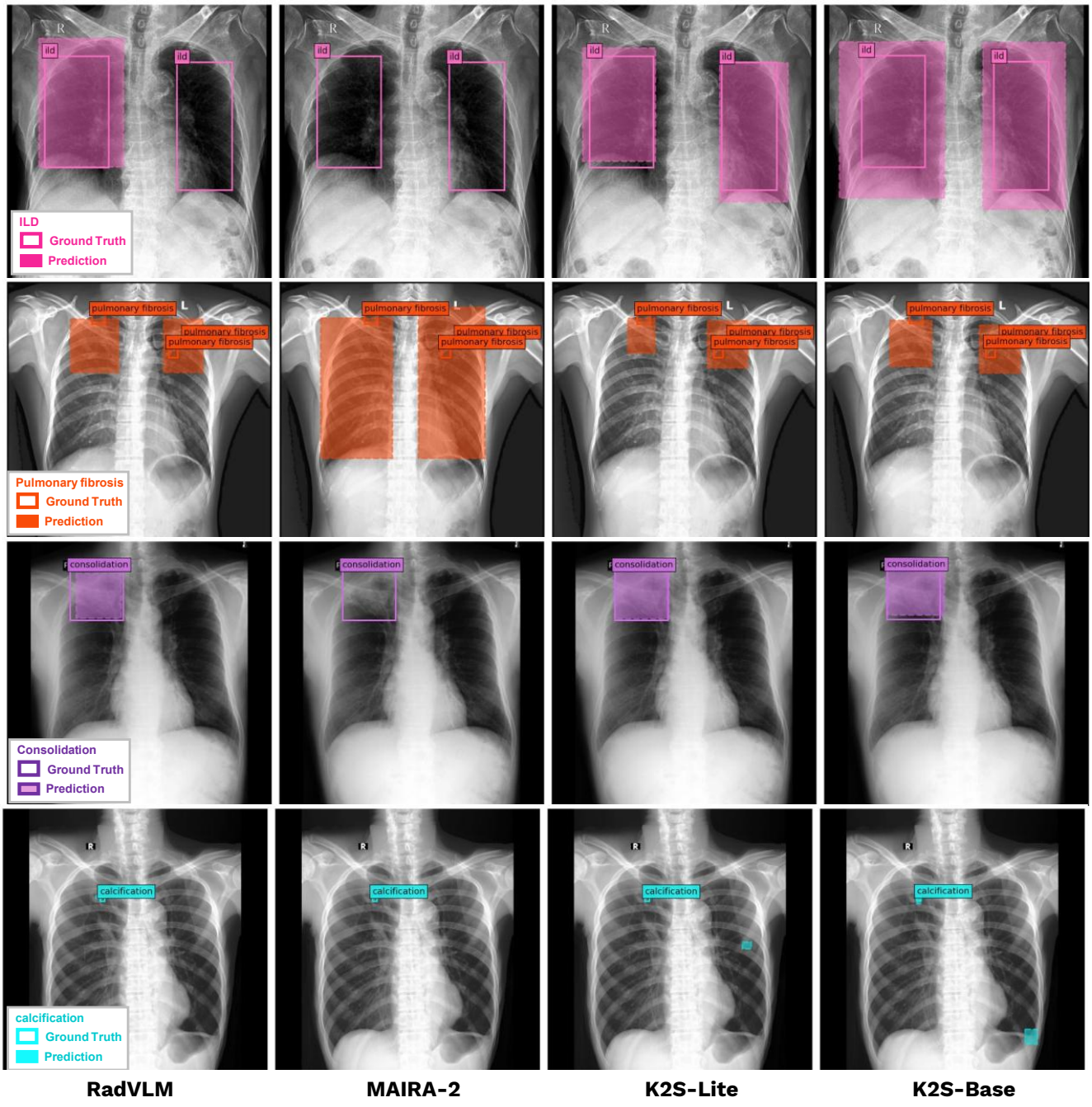

Figure 3. Qualitative examples of abnormality grounding across models.

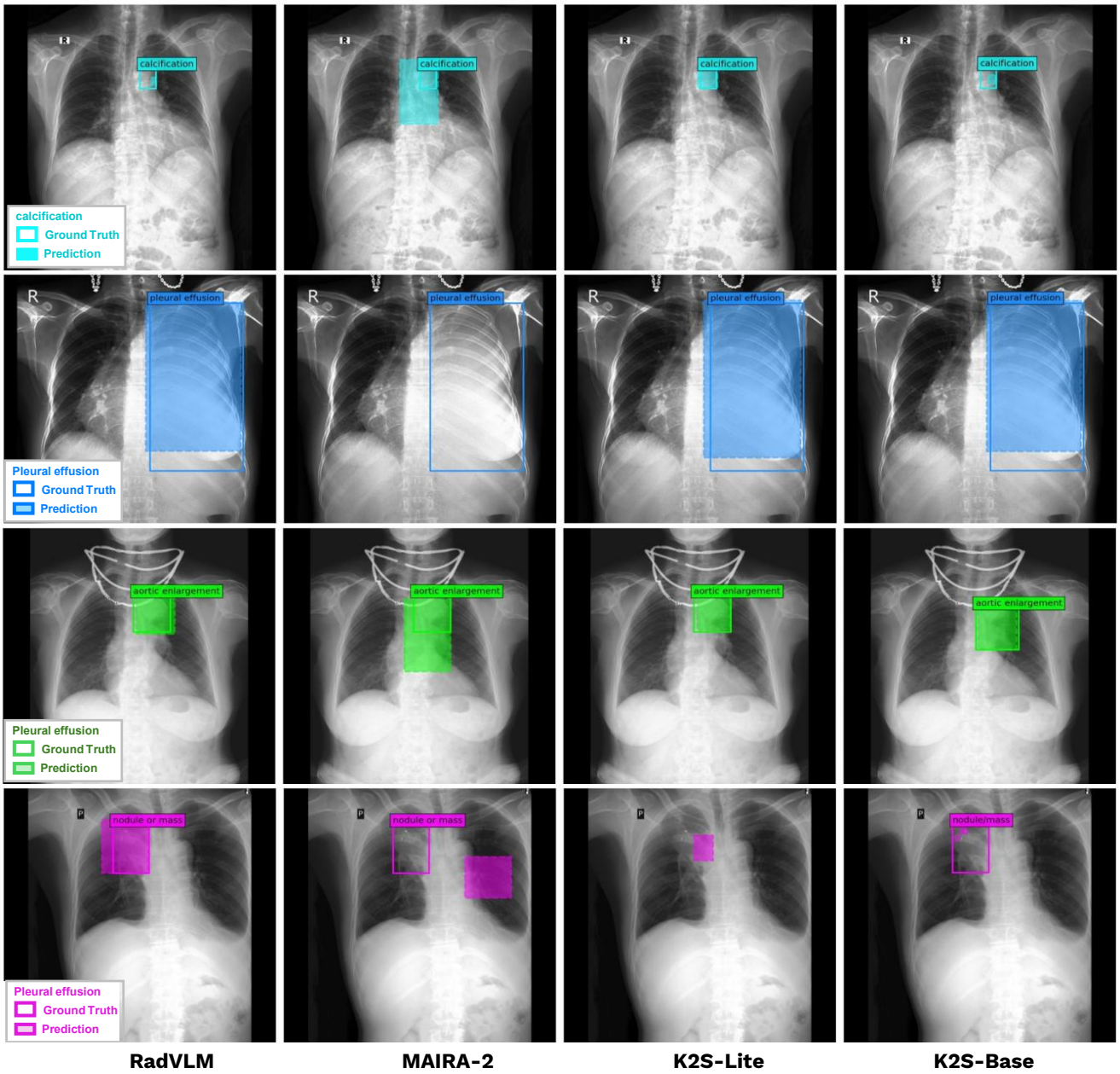

Figure 4. Qualitative examples of abnormality grounding across models.