

# VINDr-CXR-VQA: A VISUAL QUESTION ANSWERING DATASET FOR EXPLAINABLE CHEST X-RAY ANALYSIS WITH MULTI-TASK LEARNING

Dang H. Nguyen<sup>1,2,4</sup>

Hieu H. Pham<sup>2,4</sup>

Hao T. Nguyen<sup>3</sup>

Hieu H. Pham<sup>1,2,4\*</sup>

<sup>1</sup>College of Engineering and Computer Science, VinUniversity, 100000 Hanoi, Vietnam

<sup>2</sup>VinUni-Illinois Smart Health Center, VinUniversity, 100000 Hanoi, Vietnam

<sup>3</sup>Quan Su Radiology Department, Vietnam National Cancer Hospital, 100000 Hanoi, Vietnam

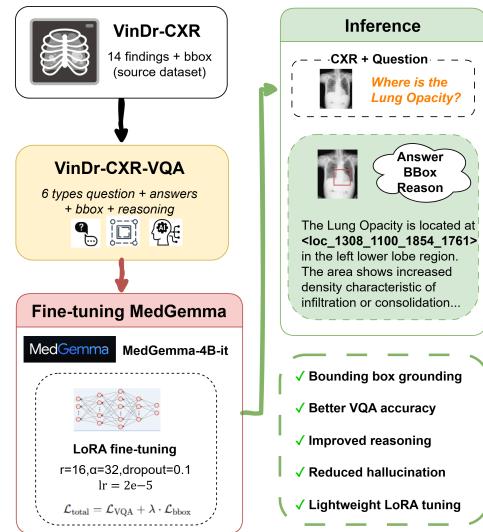
<sup>4</sup>The Computer Vision and Medical AI Lab, VinUniversity, 100000 Hanoi, Vietnam

## ABSTRACT

We present **VinDr-CXR-VQA**, a large-scale chest X-ray dataset for explainable Medical Visual Question Answering (Med-VQA) with spatial grounding. The dataset contains 17,597 question-answer pairs across 4,394 images, each annotated with radiologist-verified bounding boxes and clinical reasoning explanations. Our question taxonomy spans six diagnostic types: Where, What, Is there, How many, Which, and Yes/No, capturing diverse clinical intents. To improve reliability, we construct a balanced distribution of 41.7% positive and 58.3% negative samples, mitigating hallucinations in normal cases. Benchmarking with *MedGemma-4B-it*, a state-of-the-art medical VLM, demonstrates improved performance ( $F1 = 0.624$ , +11.8% over baseline) while enabling lesion localization. VinDr-CXR-VQA aims to advance reproducible and clinically grounded Med-VQA research. The dataset and evaluation tools are publicly available at [huggingface.co/datasets/Dangindev/VinDr-CXR-VQA](https://huggingface.co/datasets/Dangindev/VinDr-CXR-VQA).

## 1. INTRODUCTION AND RELATED WORK

Medical Visual Question Answering (Med-VQA) aims to answer clinical questions from medical images by combining visual recognition with natural language reasoning [5]. For clinical applicability, it is crucial that models not only generate diagnostic answers but also provide spatial justifications [6, 7]. Radiologists must be able to assess both *what* the model detects and *where* it localizes findings in order to validate AI outputs and ensure patient safety. Recent medical Vision-Language Models (VLMs), such as *MedGemma* [8] and *LLaVA-med* [9], show promise but still require explicit spatial grounding for clinical validation. Beyond these, recent efforts in medical Large Vision-Language Models (LVLMs) and benchmarks further expand capabilities and evaluation breadth, including *PaliGemma* [10], *RadFM* [11], and *XrayGPT* [12].



**Fig. 1: Overview of the VinDr-CXR-VQA pipeline.** The figure illustrates the multi-task learning approach, encompassing dataset creation (unifying Question & Answer, Bounding Box, and Reasoning) and the downstream fine-tuning of *MedGemma-4B-it*, enabling explainable VQA with accurate spatial grounding.

However, existing Med-VQA datasets fall short. Text-only benchmarks lack spatial annotations, while detection datasets (VinDr-CXR [13], ChestX-ray14 [14]) miss interactive Question Answering (QA) capabilities [13]. Even broader grounded efforts still lack precise VQA grounding for chest X-rays. To the best of our understanding, this critical gap necessitates a new dataset unifying clinical QA and spatial localization.

To this end, we present **VinDr-CXR-VQA**, a grounded Med-VQA dataset of 4,394 images and 17,597 QA pairs, verified by radiologists. The dataset spans six diagnostic types and includes bounding boxes and clinical reasoning. We ensured robust training by balancing positive (41.7%) and negative (58.3%) samples. VinDr-CXR-VQA is uniquely positioned to offer both spatial grounding and clinical explainability (Table 1 and Figure 1).

\*Correspondence: Huy-Hieu Pham (hieu.ph@vinuni.edu.vn)

Dataset	Num. of Images	Num. of QA Pairs	Modalities	Groundable	Explainable
VQA-RAD [1]	0.3K	3.5K	Diverse <sup>†</sup>	✗	✗
SLAKE [2]	0.6K	14K	Diverse <sup>†</sup>	✗	✗
PathVQA [3]	149K	33K	Pathology	✗	✗
MIMIC-CXR-VQA [4]	143K	377K	Chest X-ray	✗	✗
<b>Ours (VinDr-CXR-VQA)</b>	<b>4.4K</b>	<b>17.6K</b>	<b>Chest X-ray</b>	<b>✓</b>	<b>✓</b>

<sup>†</sup> Diverse: multiple body parts and imaging modalities.

<sup>‡</sup> Groundable: provides spatial annotations (e.g., bounding boxes or heatmaps) tied to the answer.

<sup>§</sup> Explainable: includes textual clinical reasoning or justification alongside the answer.

**Table 1:** Comparison of medical VQA datasets. VinDr-CXR-VQA is the only dataset that provides both spatial grounding (bounding boxes) and explainability (clinical reasoning).

### Our contributions are:

- We introduce **VinDr-CXR-VQA**, a chest X-ray dataset that combines spatial grounding (bounding boxes) with clinical question-answering and reasoning explanations.
- The dataset contains 17.6K QA pairs across 4.4K radiographs, covering six structured question types and a balanced distribution of normal and abnormal findings.
- All annotations, including bounding boxes and textual explanations, are curated and verified by board-certified radiologists.
- The dataset is publicly released to support the development and evaluation of explainable, clinically grounded Med-VQA models.

The remainder of this paper details the dataset construction, followed by our experimental setup and evaluation, and concludes with a discussion of future directions.

## 2. VINDR-CXR-VQA DATASET

### 2.1. Dataset Construction

VinDr-CXR [13] provides 18,000 posteroanterior chest X-ray images with expert-validated bounding box annotations across 22 thoracic diseases, establishing a gold standard for chest X-ray object detection. To reduce annotation noise and ensure sufficient training samples per class, we select 14 clinically relevant pathologies with higher prevalence, excluding rare or ambiguous findings. However, VinDr-CXR lacks the natural language supervision required for vision-language model training. To address this, we build VinDr-CXR-VQA by adding question-answer pairs and clinical reasoning, while preserving all expert annotations.

We selected 4,394 images from VinDr-CXR containing at least one expert-annotated pathology finding. The authors manually designed structured prompt templates for six question categories: *Where* (spatial localization), *What* (pathology identification), *Is\_there* (existence verification), *How\_many* (counting), *Which* (anatomical classification), and *Yes\_No* (binary confirmation). Each template directly references verified VinDr-CXR pathology labels and bounding boxes, grounding generated questions in validated clinical findings.

We utilize Google’s Gemini 2.5 Pro vision-language API [15] to systematically generate the required natural language content. For each image-template combination, the API receives the chest X-ray, along with its full VinDr-CXR annotation (which includes the pathology label and bounding box), as input. The API then generates three distinct components: (1) a coherent natural language question, (2) a precise answer incorporating spatial reference in `<loc_xmin_ymin_xmax_ymax>` format, and (3) a detailed clinical reasoning paragraph (150-250 words). This reasoning, which constitutes the core medical knowledge component and offers crucial explainability, is entirely generated by Gemini 2.5 Pro based on its extensive medical literature training data. The API articulates diagnostic significance, differential diagnoses, and clinical implications using appropriate medical terminology. Critically, the API only generates textual descriptions – all pathology labels and bounding box coordinates are directly copied from VinDr-CXR without modification, ensuring evaluation metrics assess model performance against verified clinical expertise.

Field	Source	Description
question	API	Natural language query
answer	API	Response with <code>&lt;loc&gt;</code> reference
reason	API	Clinical reasoning (150-250 words)
type	API	Question category (6 types)
difficulty	API	Easy/Medium
gt_finding	VinDr-CXR	Expert pathology label (unchanged)
gt_location	VinDr-CXR	Expert bounding box (unchanged)

**Table 2:** JSON Schema with Source Attribution

Table 2 presents the dataset’s JSON schema. Each sample contains seven fields: five generated by the API (question, answer, reason, type, difficulty) and two directly inherited from VinDr-CXR (gt\_finding, gt\_location). This clear separation between generated content and preserved ground truth is fundamental to the dataset’s reliability.

## 2.2. Quality Control and Clinical Validation

To ensure dataset reliability, we performed rigorous quality control via automated verification and clinical expert review.

**Automated Verification.** Automated Verification involves programmatic validation, ensuring 100% preservation of expert annotations. Scripts meticulously verified structural integrity across all seven JSON fields. Systematic cross-validation of all 4,394 images further confirmed zero mismatches against VinDr-CXR source files, ensuring 100% preservation of expert annotations.

**Clinical Expert Review.** Clinical Expert Review involved independent evaluation of 100 question-answer pairs (0.57% sample) by two board-certified radiologists (8 years experience). Reviewers assessed three dimensions (accuracy of reasoning, answer appropriateness, and spatial reference correctness). Inter-rater agreement was near-perfect ( $\kappa = 0.89$  [95% CI: 0.84, 0.93]). Initial disagreements (7 cases) were resolved via consensus, resulting in all 100 samples receiving "acceptable" ratings across all dimensions.

This multi-tier process provides strong evidence of trustworthiness, ensuring perfect ground truth preservation and medical accuracy via high inter-rater reliability.

## 2.3. Dataset Statistics and Composition

Attribute	Value
Source images (VinDr-CXR)	4,394
Generated Q&A pairs	17,597
Average Q&A per image	4.0
Pathology classes	14
Question types	6
Training split	3,735 images (85%)
Validation split	659 images (15%)
Test split	300 images
Difficulty distribution	Easy: 49.1%, Medium: 50.9%
Question type distribution	Balanced (16.4–17.0% each)
Multi-lesion complexity (validation)	54.8% with $\geq 2$ pathologies
Average bboxes per validation image	8.3

**Table 3:** VinDr-CXR-VQA Dataset Statistics

Table 3 summarizes the comprehensive dataset composition. From VinDr-CXR’s 18,000 images, we generated 17,597 question-answer pairs from 4,394 images (avg 4.0 pairs/image), featuring balanced distribution across six question types (16.4–17.0% each). The dataset spans 14 pathology classes with natural imbalance ranging from Cardiomegaly (28.0%) to Atelectasis (0.1%), accurately reflecting real-world clinical prevalence. This structure robustly enables assessment of model performance across both common and rare pathologies.

Data splits follow VinDr-CXR’s official partitioning with stratified sampling to maintain class proportions. Validation and test sets preserve multi-lesion complexity: 54.8% of validation images contain  $\geq 2$  pathologies with an average of 8.3

bounding boxes per image, reflecting realistic clinical scenarios where simultaneous multi-pathology detection and reasoning are required.

## 3. EXPERIMENTS

### 3.1. Implementation Details

We fine-tune MedGemma-4B-it [8], a 4B vision-language model pretrained on medical imaging, using Low-Rank Adaptation (LoRA) [16] for parameter-efficient training. LoRA is applied with rank  $r = 16$ , scaling factor  $\alpha = 32$ , and dropout 0.1, adding only 0.21% trainable parameters (8.9M parameters) while keeping the base model frozen. Optimization uses AdamW with learning rate  $2 \times 10^{-5}$ , weight decay 0.01, and batch size 4 with gradient accumulation steps of 4 (effective batch size 16). Training is conducted for 3 epochs over the 20,880-sample training set, requiring approximately 36 hours on a single NVIDIA A6000 GPU (48GB VRAM). Mixed precision training (fp16) is employed to reduce memory consumption. We save checkpoints at the end of each epoch and select the best model based on validation F1 score.

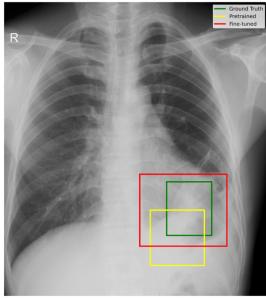
Our multi-task objective combines VQA and bounding box prediction:  $\mathcal{L}_{total} = \mathcal{L}_{VQA} + \lambda \cdot \mathcal{L}_{bbox}$ , with  $\lambda = 0.5$ . The VQA loss uses cross-entropy over answer tokens, while the bbox loss employs IoU-based regression. Bounding box coordinates are predicted via special tokens ( $<\text{loc}_x\text{x1}_y\text{y1}_x\text{x2}_y\text{y2}>$ ) embedded in the generated answer, enabling unified end-to-end training without architectural modifications.

### 3.2. Evaluation Metrics

For **VQA performance**, we evaluate text-based lesion classification using Accuracy, Precision, Recall, and F1 score on the validation set (659 images, 5,471 ground-truth bounding boxes), comparing the pretrained baseline (zero-shot, no finetuning on our dataset) against our fine-tuned model.

For **bounding box localization**, we **specifically adopt standard** detection metrics with IoU threshold 0.3: a predicted box is a true positive (TP) if it overlaps any ground-truth box with  $\text{IoU} \geq 0.3$ , otherwise a false positive (FP). Missed ground-truth boxes are false negatives (FN). We report Precision, Recall, F1, and mean IoU (mIoU) computed over TPs only. The IoU threshold of 0.3 represents a clinically meaningful minimum overlap.

To assess **clinical localization quality** on a per-image basis, we measure the proportion of validation images achieving at least one prediction with  $\text{IoU} \geq 0.5$  ("good" localization) and  $\text{IoU} \geq 0.3$  ("acceptable" localization), plus mean IoU over all predictions.

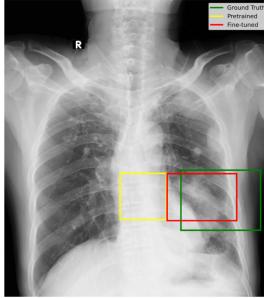


**Q: Where is the Lung Opacity?**

**Ground Truth:** The Lung Opacity is located at <loc\_1369\_1304\_1772\_1646>. Reasoning: This finding requires clinical evaluation.

**MedGemma Pretrained:** Based on the chest X-ray provided, the lung opacity is located at <loc\_1580\_1180\_2000\_1590> in the left lower lobe. It is difficult to definitively determine the cause without additional information. The location suggests a possible consolidation, pneumonia, or other infiltrative process. This is a preliminary interpretation based on a single image.

**MedGemma Fine-tuned:** The Lung Opacity is located at <loc\_1308\_1100\_1854\_1761> in the left lower lobe region. The area shows increased density characteristic of infiltration or consolidation. This finding requires clinical evaluation and correlation with patient symptoms.



**Q: Where is the Consolidation?**

**Ground Truth:** The Consolidation is located at <loc\_1472\_1526\_1992\_2215>. Reasoning: This finding requires clinical evaluation.

**MedGemma Pretrained:** Based on the chest X-ray provided, the consolidation is located at <loc\_1500\_1000\_1900\_1400> in the left lower lobe. The area shows increased density compared to surrounding lung tissue, suggesting air spaces filled with fluid or material, characteristic of consolidation. This is a preliminary interpretation requiring clinical correlation.

**MedGemma Fine-tuned:** The Consolidation is located at <loc\_1503\_1406\_1909\_2009> in the left lower lobe. The increased opacity indicates fluid or inflammatory material filling the alveolar spaces, consistent with consolidation. This finding requires clinical evaluation and further diagnostic workup.

**Fig. 2:** Model comparison on VinDr-CXR validation images. Bounding boxes: ground truth (green), Predictions from MedGemma Pretrained (yellow), and Predictions from MedGemma Fine-tuned (red). (A) Left Image: Infiltration with high IoU (0.715). The Fine-tuned model achieves excellent localization, significantly outperforming the pretrained baseline. (B) Right Image: Consolidation with moderate IoU (0.482). The Fine-tuned model demonstrates improved accuracy over the pretrained baseline. Both cases show the fine-tuned model’s superior spatial localization while maintaining clinical reasoning capability.

### 3.3. Quantitative Results

**VQA Performance.** Table 4 shows that fine-tuning on VinDr-CXR-VQA improves F1 from 0.558 to 0.624 (+11.8%), with consistent gains across all metrics.

Model	Acc.	Prec.	Recall	F1
Pretrained	0.558	0.421	0.827	0.558
Fine-tuned	<b>0.624</b>	<b>0.471</b>	<b>0.926</b>	<b>0.624</b>

**Table 4:** VQA performance on 659 validation images.

**Bounding Box Detection.** Table 5 reports bbox detection performance. While recall remains low (0.090), the model achieves a high mIoU of 0.615 on true positives, confirming its ability to produce accurate spatial grounding when predictions are correct.

Model	Prec.	Recall	F1	mIoU (TP)
Pretrained	0.096	0.068	0.070	0.036
Fine-tuned	<b>0.120</b>	<b>0.090</b>	<b>0.103</b>	<b>0.615</b>

**Table 5:** Bounding box detection performance ( $\text{IoU} \geq 0.3$ ).

**Localization Quality.** Table 6 summarizes spatial localization quality on a per-image basis. The fine-tuned model achieves "good" localization ( $\text{IoU} \geq 0.5$ ) in 22.8% of cases, and "acceptable" localization ( $\text{IoU} \geq 0.3$ ) in 48.6% of cases — substantially outperforming the pretrained baseline.

Metric	Pretrained	Fine-tuned
$\text{IoU} \geq 0.5$ (%)	1.1	<b>22.8</b>
$\text{IoU} \geq 0.3$ (%)	8.5	<b>48.6</b>
Mean IoU (all preds)	0.036	<b>0.133</b>

**Table 6:** Spatial localization quality (per image).

**Multi-task Effect.** Fine-tuning improves the primary VQA task while simultaneously enabling spatial grounding, demonstrating the effectiveness of VinDr-CXR-VQA as a multi-modal, clinically grounded supervision source.

### 3.4. Qualitative Results

Figure 2 illustrates examples of improved localization after fine-tuning. Ground truth boxes are shown in green, predictions from the pretrained model in yellow, and predictions from the fine-tuned model in red. The examples highlight the model’s ability to accurately localize abnormalities (e.g., infiltration, consolidation) while preserving clinical reasoning in its textual responses.

## 4. DISCUSSION AND CONCLUSION

VinDr-CXR-VQA presents a large-scale, publicly available dataset designed to support spatially-grounded medical VQA, consisting of 17,597 question-answer pairs over 4,394 chest X-rays. Each sample is annotated with expert-validated bounding boxes and clinical reasoning, enabling both answer generation and spatial localization. Fine-tuning MedGemma-4B-it on this dataset leads to a substantial improvement in VQA F1 (from 0.558 to 0.624, +11.8%) while introducing spatial grounding capability (mean IoU = 0.615), without compromising text-based performance.

Importantly, 22.8% of validation images achieve “good” localization ( $\text{IoU} \geq 0.5$ ), allowing visual verification of AI outputs—which is a critical step toward interpretable and clinically actionable VQA. Nonetheless, the model’s low recall (9.0%) reflects the lack of multi-lesion training examples, whereas validation images average 8.3 lesions each. This highlights a key direction for improvement: incorporating lesion-dense training data to effectively bridge the distribution gap and enhance generalization.

Despite current limitations, the dataset’s balanced structure (41.7% positive, 58.3% negative) and clinically relevant question taxonomy provide a strong foundation for safe and explainable medical AI. Future work should enrich lesion diversity, explore structured supervision for multi-instance detection, and extend clinical validation across broader populations and institutions.

## 5. COMPLIANCE WITH ETHICAL STANDARDS

This retrospective study utilized human subject data made publicly available through open access by the VinDr-CXR dataset [13], provided by the Vingroup Big Data Institute. According to the license provided with the dataset, ethical approval was not required. The dataset is released under the CC BY 4.0 license.

## 6. ACKNOWLEDGMENTS

This work was supported by the VinUni-Illinois Smart Health Center (VISHC). We thank the creators of the VinDr-CXR dataset for providing this valuable resource.

## 7. REFERENCES

- [1] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman, “A dataset of clinically generated visual questions and answers about radiology images,” *Scientific Data*, vol. 5, no. 1, pp. 1–10, 2018.
- [2] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu, “SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering,” in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2021, pp. 1650–1654.
- [3] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie, “PathVQA: 30000+ Questions for Medical Visual Question Answering,” 2020.
- [4] Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, et al., “EHRXQA: A multi-modal question answering dataset for electronic health records with chest X-ray images,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [5] Zhihong Lin, Donghao Zhang, Qingyi Tao, Danli Shi, Gholamreza Haffari, Qi Wu, Mingguang He, and Zongyuan Ge, “Medical visual question answering: A survey,” *Artificial Intelligence in Medicine*, vol. 143, pp. 102611, 2023.
- [6] Joy T. Wu, Nkechinyere N. Agu, Ismini Lourentzou, Arjun Sharma, Joseph A. Paguio, et al., “Chest imangenome dataset for clinical reasoning,” *Advances in Neural Information Processing Systems*, 2021.
- [7] Bo Liu, Ke Zou, Li-Ming Zhan, Zexin Lu, Xiaoyu Dong, et al., “Gemex: A large-scale, groundable, and explainable medical vqa benchmark for chest x-ray diagnosis,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2025, pp. 21310–21320.
- [8] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, et al., “MedGemma Technical Report,” 2025.
- [9] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao, “LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day,” in *Advances in Neural Information Processing Systems*, 2024, vol. 36.
- [10] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, et al., “Paligemma: A versatile 3b vlm for transfer,” *arXiv preprint*, vol. arXiv:2407.07726, 2024, Version v2, submitted 10 Jul 2024, revised 10 Oct 2024.
- [11] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie, “Towards generalist foundation model for radiology by leveraging web-scale 2D&3D medical data,” *Nature Communications*, vol. 16, no. 1, pp. 7866, 2025.
- [12] Omkar Chakradhar Thawakar, Abdelrahman M. Shaker, Sahal Shaji Mullappilly, et al., “XrayGPT: Chest radiographs summarization using large medical vision-language models,” in *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*. Aug. 2024, pp. 440–448, Association for Computational Linguistics.
- [13] Ha Q. Nguyen, Khanh Lam, Linh T. Le, Hieu H. Pham, Dat Q. Tran, Dung B. Nguyen, Dung D. Le, Chi M. Pham, Hang T. T. Tong, Diep H. Dinh, Cuong D. Do, Luu T. Doan, et al., “VinDr-CXR: An open dataset of chest X-rays with radiologist’s annotations,” *Scientific Data*, vol. 9, no. 1, pp. 429, 2022.
- [14] 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 (CVPR)*, 2017, pp. 2097–2106.
- [15] Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, et al., “Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities,” 2025.
- [16] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen, “LoRA: Low-rank adaptation of large language models,” in *International Conference on Learning Representations (ICLR)*, 2022.