RADSIGHT: TOWARDS PERCEPTUALLY RELIABLE MULTIMODAL RADIOLOGY IMAGE UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical multimodal large language models (MLLMs) are increasingly expected to perform complex image understanding tasks, yet their reliability is often compromised by frequent errors in visual interpretation. To systematically trace these failures, we traverse the hierarchy from high-level clinical tasks down to fundamental visual perception. We therefore introduce Perception-Bench, a large-scale benchmark that assesses medical MLLMs across six dimensions: attribute judgment, spatial grounding, spatial understanding, disease prediction, anomaly detection, and report generation, spanning both 2D and 3D radiology images. Our analysis on Perception-Bench reveals that existing MLLMs lack the ability to capture even the most basic lesion attributes, such as location, size, and density. This failure to ground clinical predictions in visual evidence reveals a critical yet overlooked source of diagnostic unreliability: deficient low-level visual perception. Motivated by this, we propose RadSight, a perception-driven MLLM with dedicated 2D and 3D visual encoders that preserve native imaging spatial structures. RadSight is trained on an 8.37-million-sample perception-oriented corpus through a four-stage curriculum: visual-language alignment, fine-grained visual perception, clinical diagnosis, and diagnostic interpretation. This curriculum explicitly develops perceptual capabilities before progressing to diagnosis and report generation. RadSight outperforms existing MLLMs across all six dimensions of Perception-Bench, with particularly strong gains in spatial grounding and clinical diagnosis. It also achieves consistent improvements on public 2D and 3D medical benchmarks, further demonstrating that robust low-level visual perception is a critical foundation for reliable clinical understanding.

1 INTRODUCTION

Recent advances in large language models and vision-language modeling have accelerated the development of medical multimodal large language models (MLLMs) (Moor et al., 2023a). Representative systems, including Lingshu (Xu et al., 2025), Hulu-Med (Jiang et al., 2025), and LLaVA-Med (Li et al., 2023), extend general-purpose vision-language models to biomedical and radiological settings. These systems are commonly developed through large-scale medical visual instruction tuning (Li et al., 2023; Zhang et al., 2023), visual-language alignment (Boecking et al., 2022; Li et al., 2023), and supervision over diverse clinical tasks (Tu et al., 2024). The resulting models support medical visual question answering (Liu et al., 2021; Zhang et al., 2023), abnormality recognition and diagnostic assistance (Xu et al., 2025; Jiang et al., 2025), and radiology report generation (Chen et al., 2020; Jin et al., 2024; Li et al., 2025). Collectively, these capabilities position MLLMs as flexible interfaces for interpreting medical images and producing clinically relevant responses.

Despite these advances, current medical MLLMs fall short when held to the standard of clinical reliability. A common failure mode is that, when given a CT volume and asked to generate a radiology report, existing models often overlook critical abnormal findings or even produce overly normal reports (Liu et al., 2024), as shown in Figure 1 (a). However, report generation is a high-level and sparse task: an incorrect report may arise from multiple sources, including weak language generation, insufficient diagnostic reasoning, or poor visual perception (Ostmeier et al., 2024; Zhao et al., 2024; Mahdavi et al., 2026). To trace the cause of these failures, we study this problem by treating task simplification as a diagnostic probe. We first convert report generation into clinical diagnosis by directly asking whether a specific disease is present. Surprisingly, existing models still produce

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

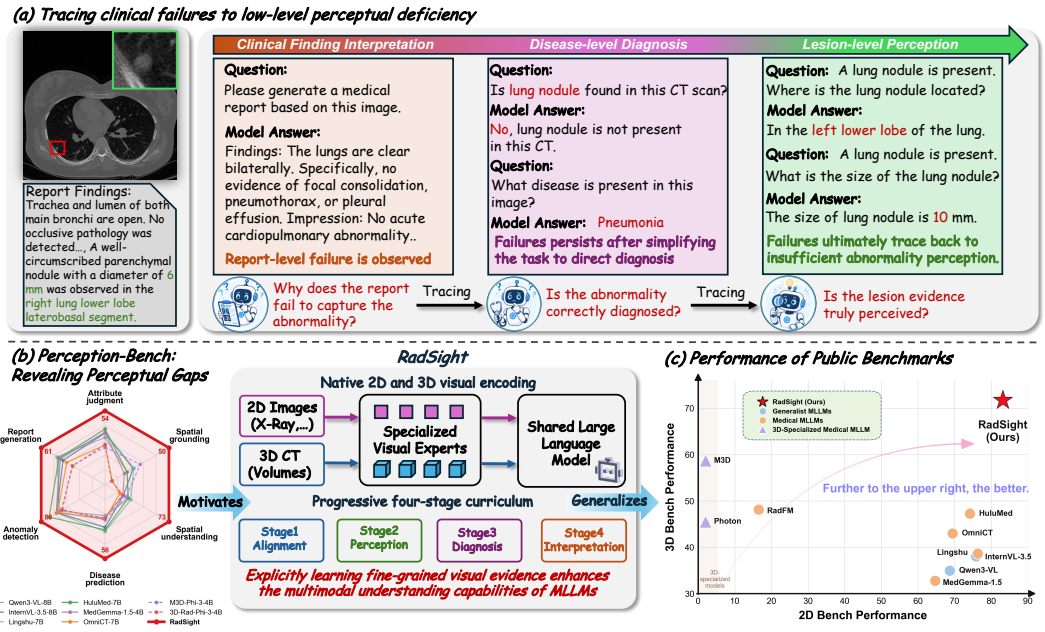


Figure 1: From perceptual failure analysis to improved medical multimodal understanding. (a) Tracing clinical failures to perceptual limitations. A representative medical MLLM, Lingshu (Xu et al., 2025), misses abnormal findings in report generation and continues to fail as the task is simplified to diagnosis and lesion-attribute QA, revealing limitations in fine-grained visual perception. (b) To address gaps in existing perceptual evaluation, we introduce *Perception-Bench*, which systematically exposes these deficiencies and motivates the development of RadSight. (c) RadSight outperforms the evaluated baselines on public medical benchmarks, indicating that stronger low-level visual perception supports better generalization.

incorrect answers. We then further reduce the task to lesion-level perception by asking about basic abnormality attributes, such as lesion location, size, and morphology. The models continue to fail on these simpler queries. This failure pattern reveals a critical bottleneck: *current medical MLLMs often struggle to perceive underlying visual abnormalities with sufficient clarity, thereby compromising their reliability across diagnosis, report generation, and more complex clinical tasks.*

To systematically evaluate existing medical MLLMs from high-level clinical tasks down to fundamental visual perception, we introduce **Perception-Bench**, a large-scale benchmark designed to assess fine-grained visual perception and its role in clinical understanding. Existing 2D (Liu et al., 2021; Hu et al., 2024; Zhang et al., 2023; Yamagishi et al., 2025) and 3D medical benchmarks (Gai et al., 2026; Chen et al., 2026; Monon et al., 2026; Yuan et al., 2026) remain fragmented, lacking a unified, systematic framework for evaluating low-level visual perception across both modalities. Perception-Bench addresses this gap through an evidence hierarchy comprising six dimensions. At the perception level, attribute judgment, spatial grounding, and spatial understanding examine whether clinically relevant visual evidence is correctly recognized and localized. At the diagnosis level, disease prediction and anomaly detection assess whether this evidence supports disease-level decisions. Finally, at the interpretation level, report generation assesses the integration of perceptual and diagnostic evidence into coherent clinical descriptions. As illustrated in Figure 1 (b), this comprehensive evaluation reveals that existing MLLMs suffer from substantial deficiencies in visual perception across multiple dimensions.

The deficiencies revealed by Perception-Bench motivate **RadSight**, a perception-driven medical MLLM designed around two principles. First, visual encoding should preserve the native spatial structures of 2D radiographs and 3D CT volumes. RadSight therefore employs dedicated 2D and 3D visual encoders to capture fine-grained in-plane features and volumetric anatomical representations, respectively, and projects their outputs into a shared language embedding space. Second, high-level clinical interpretation should be learned after the model has acquired explicit evidence-level capabilities. To this end, we construct an 8.37-million-sample perception-oriented corpus and adopt a four-stage curriculum: visual-language alignment, fine-grained visual perception, clinical diagno-

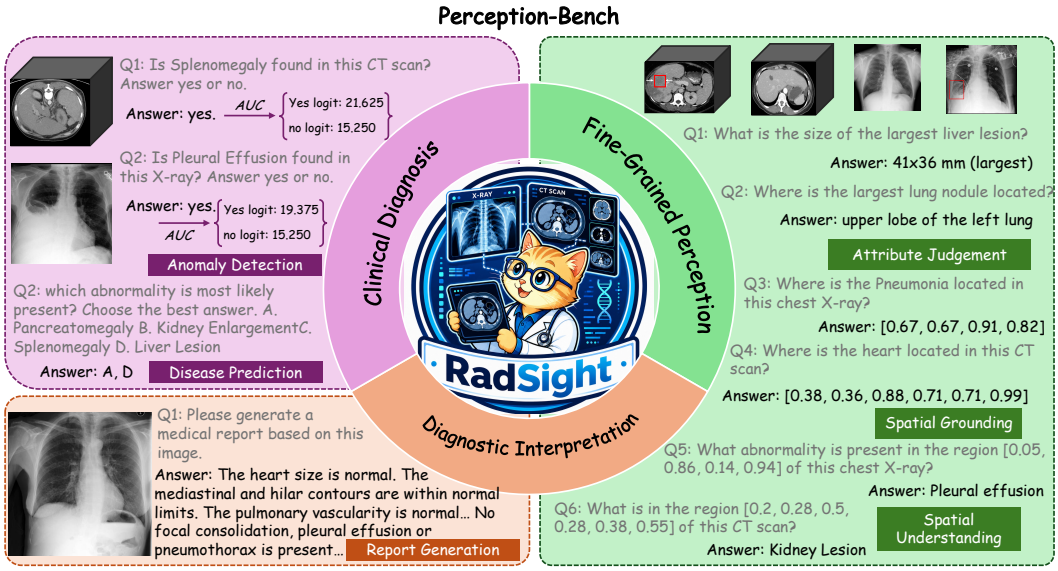


Figure 2: Task definitions of Perception-Bench.

sis, and diagnostic interpretation. This curriculum turns lesion attributes and spatial correspondence into explicit intermediate learning objectives instead of expecting them to emerge implicitly from diagnosis or report supervision.

Experiments on Perception-Bench demonstrate that RadSight improves performance across the full evaluation hierarchy relative to generalist, 2D medical, and 3D medical MLLMs. Further evaluations on public 2D and 3D benchmarks show its transfer and task adaptation capabilities. Together, these results support our central hypothesis: **explicitly learning fine-grained visual evidence strengths medical multimodal understanding.**

Our contributions are summarized as follows:

• **Perception failure tracing**: We identify and analyze a fundamental limitation of current medical MLLMs: their high-level clinical failures often originate from insufficient low-level visual perception, as revealed by progressively simplifying tasks from report generation to diagnosis and lesion-attribute perception.

• **Perception-Bench**: We introduce Perception-Bench, a 1.13-million-sample benchmark that systematically evaluates medical MLLMs across six dimensions, covering fine-grained perception, spatial grounding, diagnosis, anomaly detection, and report generation over both 2D medical images and 3D CT volumes.

• **RadSight**: We develop a medical MLLM with native visual encoding and a four-stage perception-driven curriculum trained on 8.37 million samples. Experiments show consistent improvements in perceptual, diagnostic, and report-generation performance on Perception-Bench and public medical benchmarks.

2 Perception-Oriented Dataset

In this section, we introduce the task definitions and construction process of Perception-Bench, as illustrated in Figure 2 and Figure 3. Our goal is to establish a comprehensive benchmark for evaluating low-level visual perception in medical MLLMs across 2D images and 3D volumes, spanning the hierarchy from fundamental visual perception to high-level clinical understanding. To this end, we define six representative VQA tasks: attribute judgment, spatial grounding, spatial understanding, anomaly detection, disease prediction, and report generation.

We construct a unified annotation pipeline that converts heterogeneous medical data into structured clinical concepts, lesion attributes, and spatial regions. As illustrated in Figure 3a, the pipeline integrates two complementary sources of supervision: report-derived semantic annotations and image-

derived spatial annotations. These unified annotations are subsequently used to construct Perception-Bench for evaluation and a separate perception-oriented corpus for training RadSight.

2.1 Task Definitions

Task 1: Attribute Judgment. Attribute judgment evaluates whether a model can recognize clinically meaningful visual properties of a specified abnormality. Given an attribute question about a lesion or imaging finding, the model produces a short free-text response describing its location, size, morphology, density, margin, count, or severity. For example, a question may ask for the size of a liver lesion or the lobe-level location of a lung nodule. Anatomical locations are expressed in clinical terms rather than image coordinates. **Task 2: Spatial Grounding.** Spatial grounding evaluates whether the model can anchor a named clinical concept to the correct image region. Given an anatomical structure, lesion, or imaging finding, it predicts one or more normalized bounding boxes in 2D radiographs or 3D CT volumes. The task accommodates multiple spatially separated instances of a concept and assesses explicit concept-to-region correspondence rather than image-level recognition alone. **Task 3: Spatial Understanding.** Spatial understanding evaluates the inverse association between visual regions and clinical semantics. Given a marked 2D or 3D region, the model identifies the corresponding anatomical structure, lesion, or imaging finding. Whereas spatial grounding asks where a named concept occurs, this task asks what the marked visual evidence represents, testing region-level recognition. **Task 4: Disease Prediction.** Disease prediction evaluates whether perceptual evidence can support a broad clinical recognition decision. Given an image or volume and a set of candidate diseases or imaging findings, the model selects every candidate supported by the visual input. Both single-answer and multiple-answer cases are included, requiring the model to discriminate among plausible alternatives rather than simply answer a binary query. **Task 5: Anomaly Detection.** Anomaly detection performs binary verification for a single queried abnormality or imaging finding. The model answers "Yes" or "No" according to whether the target is present in the image or volume, and the confidence assigned to "Yes" enables disease-wise AUC evaluation. In contrast to disease prediction, this formulation isolates the ability to detect one clinical condition at a time. **Task 6: Report Generation.** Report generation assesses high-level clinical interpretation from an input image or volume. The model generates a free-text radiology report that summarizes clinically salient findings and, when applicable, a diagnostic impression. By integrating visual attributes, spatial evidence, and disease-level judgments, this task measures whether the model can produce a coherent and clinically meaningful description.

2.2 Unified Annotation Construction

Data Source. Our perception-oriented corpus is constructed from 12 publicly available medical datasets spanning 2D chest X-rays and 3D CT volumes, resulting in 8.37 million training samples with complementary semantic and spatial supervision. The source datasets provide paired medical images and radiology reports, segmentation masks, and region annotations for constructing clinical concepts, lesion attributes, and spatial regions. More details are provided in Appendix B.1.

Report-derived Semantic Annotations. Free-text radiology reports contain clinically relevant findings but lack standardized representations of organs, diseases, and lesion attributes. We convert each report into structured annotations through a two-stage pipeline, with quality control at both the disease-label and QA-pair levels. First, Qwen3-Max (Bai et al., 2025) decomposes each report into organ-level findings and extracts diseases or imaging findings together with their supporting evidence and status (*present*, *absent*, *suspected*, or *historical*). Synonymous expressions are normalized into a common disease vocabulary, which is screened by six human reviewers and then rechecked by an LLM to remove ambiguous, redundant, or clinically inconsistent labels. Second, for each retained finding, Qwen3-Max extracts report-supported attributes, including location, size, shape, density, margin, count, and type, and converts each attribute–value pair into a question–answer instance. For quality assessment, 500 randomly sampled QA pairs are evaluated by DeepSeek-V4-Pro (Xu et al., 2026), Kimi-K3 (Team et al., 2026), GLM-5.2 (Zeng et al., 2026), and Qwen3-Max under a shared rubric covering factual accuracy (whether the answer is supported by the source report) and question–answer consistency (whether it addresses the queried attribute). We compare each grader with the other three; as shown in Figure 5, Qwen3-Max achieves the strongest overall cross-grader consistency. Detailed extraction prompts, disease-label screening procedures, validation criteria, and output constraints are provided in Appendix B.4.

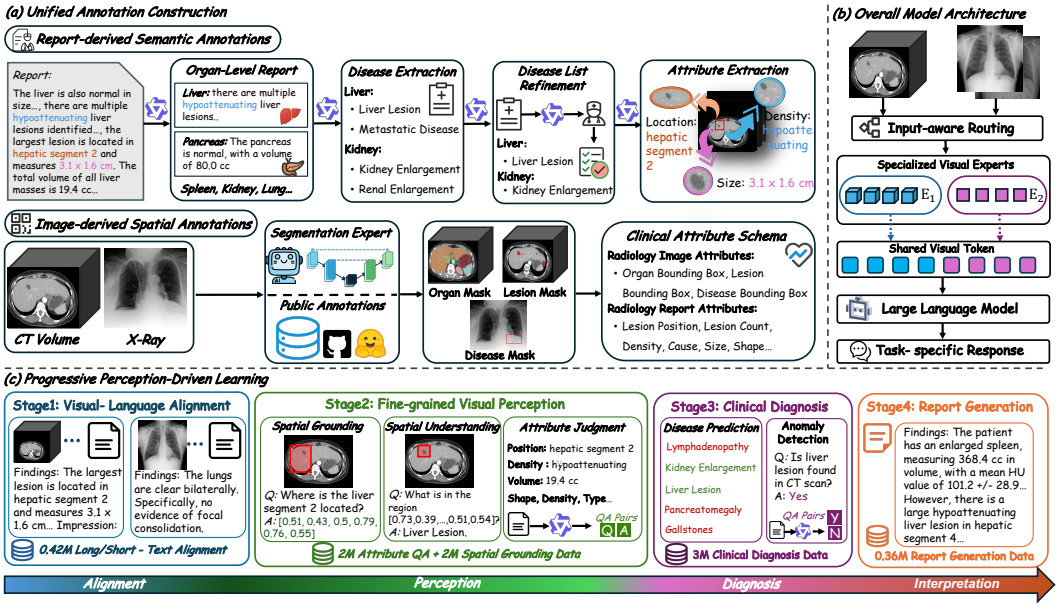


Figure 3: (a) Unified Annotation Construction. Report derived semantic annotations and image derived spatial annotations are converted into a unified clinical attribute schema covering lesion attributes and spatial regions. (b) Overall Model Architecture. RadSight uses modality specific 2D and 3D visual encoders with projectors that map visual features into a shared LLM space. (c) Progressive Perception-Driven Learning. RadSight follows a four stage curriculum from visual language alignment to perception, diagnosis, and clinical interpretation.

Image-derived Spatial Annotations. We construct spatial annotations from segmentation masks and native bounding boxes to support spatial grounding and spatial understanding (Tasks 2 and 3). All spatial regions are represented by normalized coordinates, using (x_1, y_1, x_2, y_2) for 2D images and $(x_1, y_1, z_1, x_2, y_2, z_2)$ for 3D volumes. For CT volumes, organ bounding boxes are derived from segmentation masks, while disconnected lesion components are processed separately to preserve instance-level annotations. These spatial annotations are paired with their corresponding anatomical or pathological concepts to construct bidirectional supervision, enabling models to localize specified regions and identify the concepts associated with given regions. Detailed preprocessing and annotation procedures are provided in Appendix C.1.

3 RadSight

RadSight builds on the central observation from Perception-Bench: reliable clinical interpretation should be grounded in explicit visual evidence. Accordingly, its architecture comprises two main components: a modality-native visual encoding module and a large language model (Figure 3b). Training follows a progressive perception-driven curriculum (Figure 3c). The visual encoding module preserves the distinct spatial structures of 2D radiographs and 3D CT volumes, while the curriculum develops fine-grained perceptual capabilities before progressing to disease-level reasoning and high-level report generation.

3.1 Modality Native Visual Encoding

Let $X^{\mathrm{2D}} \in \mathbb{R}^{H\times W\times C}$ denote a 2D medical image and $X^{\mathrm{3D}} \in \mathbb{R}^{D\times H\times W\times C}$ denote a CT volume with D slices. For modality $m \in \{\mathrm{2D}, \mathrm{3D}\}$, RadSight routes X^m through the corresponding visual encoder E^m, followed by a modality-specific projector P^m:

$$Z^m = P^m(E^m(X^m)) \in \mathbb{R}^{N_m \times d_{\mathrm{LLM}}}, \qquad m \in \{2D, 3D\}, \tag{1}$$

where Z^m denotes the projected visual token sequence for modality m, N_m is the number of visual tokens, and d_{LLM} is the embedding dimension of the shared language model.

The 2D pathway preserves fine-grained in-plane visual information, whereas the 3D pathway captures volumetric anatomical representations that integrate rich intra-slice features and inter-slice dependencies. The projected visual tokens are concatenated with the instruction embeddings and processed by a shared language model. This unified generative interface supports open-ended responses, categorical predictions, binary decisions, spatial coordinates, and radiology reports without introducing task-specific prediction heads.

3.2 Progressive Perception-Driven Learning

Using the unified annotations described in Section 2.2, we train RadSight with a four-stage curriculum that progresses from visual-language alignment to fine-grained perception, clinical diagnosis, and diagnostic interpretation. The tasks used in Stages 2–4 follow the definitions in Section 2.1. Each stage is initialized from the preceding checkpoint. The sample counts below refer to newly introduced stage-specific data and exclude replayed examples; detailed data distributions and training configurations are provided in Appendix C.4.

Stage 1: Visual-Language Alignment. We align the 2D and 3D visual encoders with the LLM using approximately 423K medical image–text pairs, including full radiology reports and concise descriptions of imaging findings. The LLM is frozen, while the visual encoders and their projectors are optimized to produce representations compatible with its embedding space.

Stage 2: Fine-Grained Visual Perception. We jointly fine-tune RadSight on 4.56M perception samples, comprising approximately 2.19M attribute judgment, 1.19M spatial grounding, and 1.18M spatial understanding instances. This stage introduces explicit supervision for fine-grained visual attributes and bidirectional correspondence between clinical concepts and image regions.

Stage 3: Clinical Diagnosis. We introduce 3.00M diagnosis samples, including 0.95M disease prediction and 2.05M anomaly detection instances, to connect the perceptual representations learned in Stage 2 with disease-level semantics. Positive labels are obtained from curated report-derived disease lists, while negative candidates are sampled from findings annotated as absent.

Stage 4: Diagnostic Interpretation. Finally, we train on approximately 363K image-to-report pairs spanning 2D radiographs and 3D CT volumes. This stage consolidates the perceptual and diagnostic capabilities acquired in the preceding stages into coherent, visually grounded radiology reports.

3.3 Training Objectives

All tasks in RadSight are formulated as instruction-following generation and optimized using a unified autoregressive language modeling objective. At stage s, $\mathcal{D}_s$ is the training set, $H_m = [Z^m; T]$ concatenates visual tokens Z^m with instruction embeddings T, and $\mathbf{y} = (y_1, \ldots, y_{N_y})$ is the target response. We minimize the negative log-likelihood of the response tokens only, conditioned on H_m:

$$\mathcal{L}_s = -\mathbb{E}_{(H_m,\mathbf{y})\sim\mathcal{D}_s}\left[\frac{1}{N_y}\sum_{t=1}^{N_y}\log p_{\Theta_s}\left(y_t \mid H_m, y_{<t}\right)\right], \qquad s \in \{1,2,3,4\}, \tag{2}$$

where Θ_s denotes the parameters at stage s. The same likelihood covers free text, categorical labels, binary decisions, and normalized box coordinates, so no task-specific head is used. Stages 2–4 are initialized from the preceding checkpoint and optimized over the trainable subset Θ_s^{train}:

$$\Theta_s^{(0)} \leftarrow \Theta_{s-1}^*, \qquad \Theta_s^* = \arg\min_{\Theta_s^{\text{train}}} \mathcal{L}_s, \qquad s \in \{2,3,4\}, \tag{3}$$

where Θ_s^* is the resulting checkpoint. Stage 1 updates only the visual encoders and projectors, with the LLM frozen; Stages 2–4 update all components.

4 Experiments

4.1 Experimental Settings

Implementation Details. We adopt the SigLIP vision encoder (Zhai et al., 2023) for 2D radiographs and the visual encoder from RADAR (Zhang et al., 2026) for volumetric CT scans. We use

Table 1: Performance on Perception-Bench. LLM, 2D-Pos., 3D-Pos., Acc., RaTE, and CT-F1: LLM score, 2D disease-region recognition, 3D organ/lesion recognition, accuracy, RaTEScore, and CT-Rate F1. **Bold** and underlined entries mark the best and second-best results.

	Task 1	Task 2		Task 3		Task 4	Task 5			Task 6	
Model	**LLM**	**2D-mIoU**	**3D-mIoU**	**2D-Pos.**	**3D-Pos.**	**Acc.**	**AUC**	**Sens.**	**Spec.**	**RaTE**	**CT-F1**
Generalist MLLMs											
Qwen3-VL-8B (Bai et al., 2025)	34.54	17.14	–	23.35	30.54	23.27	57.36	53.46	63.45	46.19	9.39
InternVL-3.5-8B (Wang et al., 2025)	37.77	21.86	–	10.60	35.81	32.07	61.10	64.81	55.14	40.36	3.96
2D Medical MLLMs											
Lingshu-7B (Xu et al., 2025)	38.26	22.36	–	22.07	37.24	35.31	61.40	66.23	54.80	44.04	7.66
HuluMed-7B (Jiang et al., 2025)	40.81	17.17	–	26.56	26.56	34.10	70.39	69.01	64.09	44.23	23.92
MedGemma-1.5-4B (Sellergren et al., 2026)	38.46	15.02	–	20.04	18.25	25.26	61.37	67.18	55.35	42.88	11.62
OmniCT-7B (Lin et al., 2026)	29.15	5.54	–	13.70	20.14	26.21	59.72	63.20	54.28	36.82	28.90
3D Medical MLLMs											
M3D-Phi-3-4B (Bai et al., 2024)	27.47	–	27.36	–	24.29	26.64	53.63	69.10	41.50	31.65	3.59
3D-Rad-Phi-3-4B (Gai et al., 2026)	29.54	–	5.15	–	24.01	25.12	55.09	62.67	50.30	30.98	6.11
RadSight-4B	52.45	34.19	**64.68**	48.25	96.63	**56.16**	**81.10**	**78.02**	**72.35**	60.73	**36.45**
RadSight-8B	**54.15**	**35.20**	64.31	**50.11**	**96.64**	55.81	80.47	77.19	72.10	**61.24**	36.00

Table 2: Performance comparison with existing MLLMs on public 2D medical benchmarks.

Model	**SLAKE**		**PMC-VQA**	**OmniMedVQA**	**RadFig-VQA**			**Avg**
	Open	**Closed**	**Choice**	**Choice**	**Easy**	**Medium**	**Hard**	
Generalist MLLMs								
Qwen3-VL-8B (Bai et al., 2025)	65.50	74.76	46.05	75.96	74.48	72.92	71.26	68.70
InternVL-3.5-8B (Wang et al., 2025)	71.38	78.59	54.72	89.21	78.56	79.09	79.27	75.83
Medical MLLMs								
RadFM-14B (Wu et al., 2025)	31.73	18.58	14.46	16.28	16.55	8.67	9.29	16.51
Lingshu-7B (Xu et al., 2025)	78.22	81.94	55.15	82.12	80.56	77.66	78.82	76.35
HuluMed-7B (Jiang et al., 2025)	66.45	76.08	60.95	81.13	77.67	78.02	78.99	74.18
MedGemma-1.5-4B (Sellergren et al., 2026)	75.84	68.28	47.89	69.81	64.51	62.88	63.71	64.70
OmniCT-7B (Lin et al., 2026)	59.70	69.62	49.36	79.25	75.27	76.31	76.57	69.44
RadSight-4B	80.21	85.05	60.81	99.39	82.05	**81.79**	83.23	81.79
RadSight-8B	**82.27**	**88.16**	**62.30**	**99.51**	**84.05**	81.75	**84.24**	**83.18**

the language components of Qwen3-VL-4B and Qwen3-VL-8B as the LLM backbones, yielding RadSight-4B and RadSight-8B, respectively. Further details on data processing, model architecture, and training configurations are provided in Appendix C.

Evaluation Protocols. We evaluate RadSight on Perception-Bench and public medical benchmarks. On Perception-Bench, we directly evaluate baseline MLLMs without additional fine-tuning, as most have been trained on overlapping source datasets and support diverse medical tasks. This overlap is particularly common for 2D benchmarks, most of whose underlying datasets have been used in the pretraining or instruction tuning of existing medical MLLMs. To assess transfer and task adaptation, we evaluate RadSight on public benchmarks both without task-specific fine-tuning and after fine-tuning on the corresponding training splits, as specified for each experiment. These benchmarks include SLAKE (Liu et al., 2021), PMC-VQA (Zhang et al., 2023), OmniMedVQA (Hu et al., 2024), and RadFig-VQA (Yamagishi et al., 2025) for 2D images, and 3D-Rad (Gai et al., 2026), DeepTumorVQA (Chen et al., 2026), and CT-Spatial-VQA (Monon et al., 2026) for 3D CT volumes.

Evaluation Metrics. We evaluate attribute judgment using mean Qwen3-Max scores (0, 0.5, and 1 for incorrect, partially correct, and correct responses), spatial grounding using mIoU, and spatial understanding and disease prediction using accuracy. For anomaly detection, we report macro-averaged AUC as the primary metric, alongside sensitivity and specificity. Report generation uses RaTEScore (Zhao et al., 2024) and Clinical Efficacy (CE) F1 (Hamamci et al., 2026b). See Appendix Sections D.2, D.4, D.5, and D.6 for details.

4.2 Main Results

Performance on Perception-Bench. Table 1 presents results on Perception-Bench, which evaluates medical MLLMs from low-level perception to high-level clinical interpretation. Three main observations emerge: **(i) Existing medical MLLMs exhibit clear deficiencies in low-level perception.** Despite their medical image understanding capabilities, both generalist and medical MLLMs

Table 3: Benchmark-adapted results on 3D-Rad and DeepTumorVQA. Exist., S.Temp., and L.Temp. denote existence detection, static temporal diagnosis, and longitudinal temporal diagnosis. † denotes use of the training split; unmarked baseline entries use no benchmark-specific supervision.

Model	3D-Rad			DeepTumorVQA				
	Exist.	S.Temp.	L.Temp.	Recog.	Meas.	Vis.R.	Med.R.	Overall
Generalist MLLMs								
Qwen3-VL-8B (Bai et al., 2025)	54.14	14.20	37.70	50.70	27.70	35.60	21.80	34.40
InternVL-3.5-8B (Wang et al., 2025)	65.09	15.49	43.06	54.52	22.99	37.02	30.13	36.41
2D Medical MLLMs								
Lingshu-7B (Xu et al., 2025)	78.70	14.20	14.60	58.48	29.46	37.17	35.43	39.59
HuluMed-7B (Jiang et al., 2025)	81.60†	58.00†	79.40†	60.06	25.21	32.21	32.28	36.44
MedGemma-1.5-4B (Sellergren et al., 2026)	66.68	17.47	19.32	36.78	26.05	28.85	31.13	30.32
OmniCT-7B (Lin et al., 2026)	67.92†	**64.74**†	80.40†	50.62	24.72	30.60	27.35	32.84
3D Medical MLLMs								
RadFM-14B (Wu et al., 2025)	29.20	44.11	42.99	62.30†	70.10†	66.30†	27.70†	58.90†
M3D-Phi-3-4B (Bai et al., 2024)	82.43†	49.30†	74.77†	66.20†	**70.30**†	65.80†	29.10†	59.80†
Photon-3B (Fang et al., 2026)	83.07†	52.86†	77.01†	51.21	19.62	32.67	27.35	32.77
RadSight-4B	**85.14**†	63.80†	**81.69**†	**85.62**†	69.90†	**68.98**†	73.54†	**73.40**†
RadSight-8B	**85.14**†	63.94†	80.89†	83.74†	68.76†	68.56†	**74.28**†	72.76†

remain limited in attribute judgment, spatial grounding, and spatial understanding. For example, Lingshu and HuluMed score only 38.26 and 40.81 on attribute judgment, while M3D achieves only 27.36 on spatial grounding, suggesting that visual perception remains a bottleneck for reliable clinical understanding. **(ii) Existing 2D and 3D medical models face modality-specific limitations.** 2D medical MLLMs lack native volumetric grounding, whereas dedicated 3D models offer limited support for several 2D perception tasks. RadSight supports both 2D images and 3D CT volumes through modality-specific visual encoding, achieving strong performance across both modalities. **(iii) RadSight achieves consistent gains from perception to clinical interpretation.** RadSight-4B achieves 56.16 accuracy on disease prediction and 81.10 AUC on anomaly detection across 82 diseases, with the latter exceeding the strongest baseline by 10.71 points. It also achieves the highest report generation F1 of 36.45. Overall, Perception-Bench exposes perceptual limitations in existing medical MLLMs, which RadSight substantially mitigates through perception-driven learning.

Performance on Public 2D and 3D Benchmarks. Tables 2 and 3 report results on public 2D and 3D medical benchmarks. On 2D benchmarks, RadSight-8B achieves the best average score of 83.18, outperforming Lingshu-7B by 6.83 points, while RadSight-4B also reaches a strong average of 81.79. On 3D CT benchmarks, RadSight consistently outperforms both dedicated 3D models and slice based medical MLLMs across most tasks. It achieves the best performance on two of the three 3D-Rad subtasks, while RadSight-4B obtains the highest DeepTumorVQA overall score of 73.40, exceeding the strongest baseline by 13.60 points. These results demonstrate strong performance across both 2D and 3D medical understanding tasks.

Table 4: Zero-shot results on public 3D benchmarks.

Model	DeepTumorVQA					CT-Spatial-VQA
	Recog.	Meas.	Vis.R.	Med.R.	Overall	Acc.
Qwen3-VL-8B	50.70	27.70	35.60	21.80	34.40	38.19
InternVL-3.5-8B	54.52	22.99	37.02	30.13	36.41	37.43
Lingshu-7B	58.48	**29.46**	37.17	35.43	39.59	40.16
HuluMed-7B	**60.06**	25.21	32.21	32.28	36.44	42.96
MedGemma-1.5-4B	36.78	26.05	28.85	31.13	30.32	38.24
OmniCT-7B	50.62	24.72	30.60	27.35	32.84	36.39
RadFM-14B	39.10	21.70	25.40	26.30	27.60	31.49
M3D-Phi-3-4B	4.60	25.00	25.20	14.20	18.90	30.57
Photon-3B	51.21	19.62	32.67	27.35	32.77	33.25
RadSight-4B	59.32	27.83	**39.86**	**58.01**	**44.82**	**44.33**

Zero-Shot Transfer to DeepTumorVQA and CT-Spatial-VQA. Although these benchmarks share source datasets with our training corpus, their evaluation patients and QA pairs are excluded from training. Without benchmark-specific adaptation, RadSight-4B achieves an overall score of 44.82 on DeepTumorVQA and an accuracy of 44.33 on CT-Spatial-VQA, outperforming the strongest evaluated baselines by 5.23 and 1.37 points, as shown in Table 4. These results support the transfer of learned representations to new QA pairs on held-out patients.

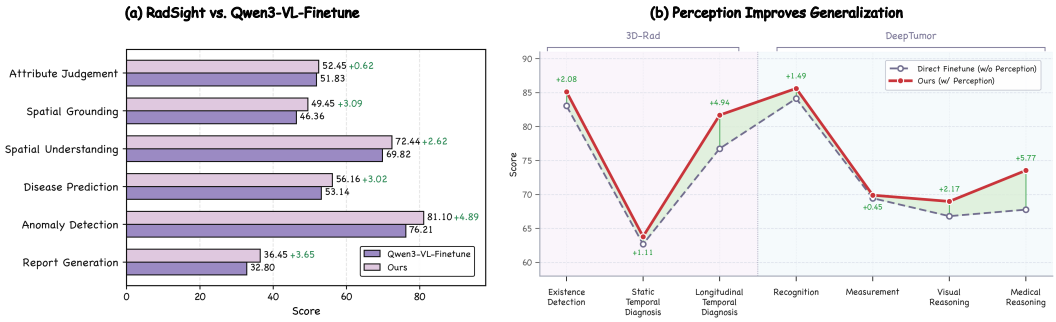


Figure 4: Analysis of RadSight Design Choices. (a) Effect of modality specific visual encoding. (b) Effect of perception driven learning on generalization.

4.3 Ablation Study

Effect of Perceptual Supervision on Diagnosis and Interpretation. Table 5 evaluates the contribution of perceptual supervision to diagnosis and report generation while keeping visual–language alignment fixed. Perception alone raises report-generation CE F1 from 15.90 to 26.21. Although neither perception nor diagnosis alone improves zero-shot CT-Spatial-VQA over alignment alone (41.17 and 37.10 vs. 41.66), combining them reaches 44.33. Adding perception to diagnosis also improves disease-prediction accuracy (54.04 to 56.16), anomaly-detection AUC (77.36 to 81.10), and report-generation CE F1 (32.38 to 36.45). These results demonstrate their complementarity and, together with the curriculum-versus-joint comparison in Appendix E.1, support the effectiveness of our progressive perception-driven curriculum.

Table 5: Effect of perceptual and diagnostic supervision. D.P: disease prediction; A.D: anomaly detection.

Training Supervision			Diagnosis		Interpretation		Public Bench
Align	Perception	Diagnosis	D.P	A.D	RaTE	CT-Rate	CT-Spatial-VQA
✓	-	-	–	–	58.45	15.90	41.66
✓	✓	-	–	–	59.43	26.21	41.17
✓	-	✓	54.04	77.36	60.50	32.38	37.10
✓	✓	✓	**56.16**	**81.10**	**60.73**	**36.45**	**44.33**

Effect of Modality Specific Visual Encoding. To assess the effectiveness of RadSight's modality specific visual encoding, we compare it with Qwen3-VL-Finetune, which is directly fine tuned on the same perception oriented training corpus and processes CT volumes as sequences of 2D slices. As shown in Figure 4(a), RadSight consistently outperforms Qwen3-VL-Finetune across all evaluated dimensions, with spatial grounding mIoU increasing from 46.35 to 49.45. These results highlight the benefits of directly encoding volumetric CT scans with a dedicated 3D vision encoder, which captures volumetric anatomical structures and spatial relationships across slices. The observed improvements support our design choice of employing modality specific visual encoders to preserve the distinct spatial characteristics of 2D medical images and 3D CT volumes.

Generalization of Learned Perceptual Representations. To assess the transferability of perception driven learning, we further evaluate RadSight on public benchmarks with task formulations distinct from our training corpus. As shown in Figure 4 (b), RadSight consistently outperforms models without perception oriented training across diverse 3D CT tasks, with notable gains in visual and medical reasoning. In particular, it improves medical reasoning on DeepTumorVQA by 5.77 points. These results show that perception driven learning yields transferable visual representations that benefit higher-level clinical CT understanding beyond the original training tasks.

5 Discussion and Conclusion

We identify fine-grained visual perception as a foundation for reliable medical multimodal understanding. Perception-Bench reveals perceptual limitations in existing MLLMs, motivating RadSight's native 2D/3D encoding and progressive perception-driven learning. RadSight improves perception, diagnosis, and clinical interpretation, with ablations confirming complementary benefits of perceptual and diagnostic supervision. Remaining limitations include imprecise lesion localization and reliance on retrospective public data and automated evaluation. Future work will improve localization, extend modality coverage, and validate on external clinical data with radiologist assessment.

AI USE STATEMENT

Generative AI tools were used in the data-construction and evaluation procedures described in this paper and to assist with manuscript preparation. Specifically, Qwen3-Max was used to extract structured findings and attribute-level question–answer pairs from radiology reports and to evaluate attribute-judgment responses. An LLM performed a secondary review of the manually curated disease vocabulary, while DeepSeek-V4-Pro, Kimi-K3, GLM-5.2, and Qwen3-Max were used for cross-grader validation of 500 sampled question–answer pairs. Generative AI also assisted with language editing and LaTeX drafting. Six human reviewers screened the disease vocabulary, and the authors reviewed the study design, experimental procedures, results, and final manuscript. The authors take responsibility for the paper and its accompanying materials.

ETHICS STATEMENT

This study uses retrospective, publicly available radiology datasets and is intended for research evaluation rather than clinical deployment or autonomous medical decision-making. Its findings should be interpreted in light of the populations, annotations, and reporting practices represented in the source datasets, as well as the limitations of automated annotation and evaluation.

REPRODUCIBILITY STATEMENT

Appendix B.1 documents the data sources, and Appendix B.2 describes patient-level splits and leakage prevention. Appendix C provides data-processing, architecture, benchmark, and training details, including stage-specific hyperparameters, while Appendix D specifies the evaluation protocols. These details, together with the reported annotation prompts and validation procedures, are provided to support reproducibility.

REFERENCES

Fan Bai, Yuxin Du, Tiejun Huang, Max Q-H Meng, and Bo Zhao. M3d: Advancing 3d medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*, 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Pedro RAS Bassi, Mehmet Can Yavuz, Ibrahim Ethem Hamamci, Sezgin Er, Xiaoxi Chen, Wenxuan Li, Bjoern Menze, Sergio Decherchi, Andrea Cavalli, Kang Wang, et al. Radgpt: Constructing 3d image-text tumor datasets. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 23720–23730, 2025.

Ricardo Bigolin Lanfredi, Mingyuan Zhang, William F Auffermann, Jessica Chan, Phuong-Anh T Duong, Vivek Srikumar, Trafton Drew, Joyce D Schroeder, and Tolga Tasdizen. Reflacx, a dataset of reports and eye-tracking data for localization of abnormalities in chest x-rays. *Scientific data*, 9(1):350, 2022.

Louis Blankemeier, Joseph Paul Cohen, Ashwin Kumar, Dave Van Veen, Syed Jamal Safdar Gardezi, Magdalini Paschali, Zhihong Chen, Jean-Benoit Delbrouck, Eduardo Reis, Cesar Truyts, et al. Merlin: A vision language foundation model for 3d computed tomography. *Research Square*, pp. rs–3, 2024a.

Louis Blankemeier, Joseph Paul Cohen, Ashwin Kumar, Dave Van Veen, Syed Jamal Safdar Gardezi, Magdalini Paschali, Zhihong Chen, Jean-Benoit Delbrouck, Eduardo Reis, Cesar Truyts, et al. Merlin: A vision language foundation model for 3d computed tomography. *Research Square*, pp. rs–3, 2024b.

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*, pp. 1–21. Springer, 2022.

Yixiong Chen, Wenjie Xiao, Pedro RAS Bassi, Boyan Wang, Liang He, Xinze Zhou, Sezgin Er, Ibrahim Ethem Hamamci, Zongwei Zhou, and Alan Yuille. Deeptumorvqa: A hierarchical 3d ct benchmark for stage-wise evaluation of medical vlms and tool-augmented agents. *arXiv preprint arXiv:2605.09679*, 2026.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 1439–1449, 2020.

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.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.

Chengyu Fang, Heng Guo, Zheng Jiang, Chunming He, Xiu Li, and Minfeng Xu. Photon: Speedup volume understanding with efficient multimodal large language models. *arXiv preprint arXiv:2603.25155*, 2026.

Xiaotang Gai, Jiaxiang Liu, Yichen Li, Zijie Meng, Jian Wu, and Zuozhu Liu. 3d-rad: A comprehensive 3d radiology med-vqa dataset with multi-temporal analysis and diverse diagnostic tasks. *Advances in Neural Information Processing Systems*, 38, 2026.

Ibrahim Ethem Hamamci, Sezgin Er, and Bjoern Menze. Ct2rep: Automated radiology report generation for 3d medical imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 476–486. Springer, 2024a.

Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, et al. Developing generalist foundation models from a multimodal dataset for 3d computed tomography. *arXiv preprint arXiv:2403.17834*, 2024b.

Ibrahim Ethem Hamamci, Sezgin Er, Suprosanna Shit, Hadrien Reynaud, Dong Yang, Pengfei Guo, Marc Edgar, Daguang Xu, Bernhard Kainz, and Bjoern Menze. Better tokens for better 3d: Advancing vision-language modeling in 3d medical imaging. *Advances in Neural Information Processing Systems*, 38:135074–135102, 2026a.

Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, et al. Generalist foundation models from a multimodal dataset for 3d computed tomography. *Nature Biomedical Engineering*, pp. 1–19, 2026b.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22170–22183, 2024.

Shih-Cheng Huang, Zepeng Huo, Ethan Steinberg, Chia-Chun Chiang, Matthew P Lungren, Curtis P Langlotz, Serena Yeung, Nigam H Shah, and Jason A Fries. Inspect: a multimodal dataset for pulmonary embolism diagnosis and prognosis. *arXiv preprint arXiv:2311.10798*, 2023.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in neural information processing systems*, 35:36722–36732, 2022a.

Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in neural information processing systems*, 35:36722–36732, 2022b.

Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, et al. Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*, 2025.

Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. Promptmrg: Diagnosis-driven prompts for medical report generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 2607–2615, 2024.

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.

Bennett Landman, Zhoubing Xu, Juan Igelsias, Martin Styner, Thomas Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI multi-atlas labeling beyond cranial vault—workshop challenge*, volume 5, pp. 12. Munich, Germany, 2015.

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*, 5(1):180251, 2018.

Cheng-Yi Li, Kao-Jung Chang, Cheng-Fu Yang, Hsin-Yu Wu, Wenting Chen, Hritik Bansal, Ling Chen, Yi-Ping Yang, Yu-Chun Chen, Shih-Pin Chen, et al. Towards a holistic framework for multimodal llm in 3d brain ct radiology report generation. *Nature Communications*, 16(1):2258, 2025.

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. *Advances in Neural Information Processing Systems*, 36: 28541–28564, 2023.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Tianwei Lin, Wenqiao Zhang, Sijing Li, Yuqian Yuan, Binhe Yu, Haoyuan Li, Wanggui He, Hao Jiang, Mengze Li, Xiaohui Song, et al. Healthgpt: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. *arXiv preprint arXiv:2502.09838*, 2025.

Tianwei Lin, Zhongwei Qiu, Wenqiao Zhang, Jiang Liu, Yihan Xie, Mingjian Gao, Zhenxuan Fan, Zhaocheng Li, Sijing Li, Zhongle Xie, et al. Omnict: Towards a unified slice-volume lvlm for comprehensive ct analysis. *arXiv preprint arXiv:2602.16110*, 2026.

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)*, pp. 1650–1654. IEEE, 2021.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024.

Zahra Mahdavi, Zahra Khodakaramimaghsoud, Hooman Khaloo, Sina Bakhshandeh Taleshani, Erfan Hashemi, Javad Mirzapour Kaleybar, and Omid Nejati Manzari. Med-vcd: Mitigating hallucination for medical large vision language models through visual contrastive decoding. *Computers in Biology and Medicine*, 200:111347, 2026.

Mashrafi Monon, Umaima Rahman, Asif Hanif, Numan Saeed, and Mohammad Yaqub. Lost in volume: The ct-spatialvqa benchmark for evaluating semantic-spatial understanding of 3d medical vision-language models. *arXiv preprint arXiv:2605.08787*, 2026.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023a.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine learning for health (ML4H)*, pp. 353–367. PMLR, 2023b.

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.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson Md, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, et al. Green: Generative radiology report evaluation and error notation. In *Findings of the association for computational linguistics: EMNLP 2024*, pp. 374–390, 2024.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

François Remy, Kris Demuynck, and Thomas Demeester. Biolord-2023: semantic textual representations fusing large language models and clinical knowledge graph insights. *Journal of the American Medical Informatics Association*, 31(9):1844–1855, 2024.

Andrew Sellergren, Chufan Gao, Fereshteh Mahvar, Timo Kohlberger, Fayaz Jamil, Madeleine Traverse, Alberto Tono, Bashir Sadjad, Lin Yang, Charles Lau, et al. Medgemma 1.5 technical report. *arXiv preprint arXiv:2604.05081*, 2026.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew Lungren. Combining automatic labelers and expert annotations for accurate radiology report labeling using bert. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 1500–1519, 2020.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y Charles, et al. Kimi k3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026.

National Lung Screening Trial Research Team. The national lung screening trial: overview and study design. *Radiology*, 258(1):243–253, 2011.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *Nejm Ai*, 1(3):AIoa2300138, 2024.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5):e230024, 2023.

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*, 16(1):7866, 2025.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*, 2025.

Yosuke Yamagishi, Shouhei Hanaoka, Yuta Nakamura, Tomohiro Kikuchi, Akinobu Shimizu, Takeharu Yoshikawa, and Osamu Abe. Radfig-vqa: A multi-imaging-modality radiology benchmark for evaluating vision-language models in clinical practice. In *International Workshop on Agentic AI for Medicine*, pp. 229–238. Springer, 2025.

An Yan, Julian McAuley, Xing Lu, Jiang Du, Eric Y Chang, Amilcare Gentili, and Chun-Nan Hsu. Radbert: adapting transformer-based language models to radiology. *Radiology: Artificial Intelligence*, 4(4):e210258, 2022.

William J Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950.

Ruifeng Yuan, Wanxing Chang, Weiwei Cao, Bowen Shi, Zhongyu Wei, Ling Zhang, and Jianpeng Zhang. Ct-finebench: A diagnostic fidelity benchmark for fine-grained evaluation of ct report generation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 44045–44058, 2026.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, et al. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

Jianpeng Zhang, Weiwei Cao, and Wanxing Chang. RADAR: An expert-level generalist ai for abdominal CT diagnosis, 2026. URL `https://doi.org/10.5281/zenodo.21271172`. Model checkpoint available at `https://huggingface.co/radar-generalist/RADAR`.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023.

Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Ratescore: A metric for radiology report generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15004–15019, 2024.

APPENDIX

A RELATED WORK

Medical Multimodal Large Language Models. Medical multimodal large language models (MLLMs) extend generative vision-language models to medical applications through domain-specific pretraining, vision-language alignment, and instruction tuning. Early systems such as

LLaVA-Med (Li et al., 2023) and Med-Flamingo (Moor et al., 2023b) primarily established medical visual question answering and few-shot multimodal assistance over 2D biomedical images. More recent generalist models, including HealthGPT (Lin et al., 2025), Hulu-Med (Jiang et al., 2025), MedGemma (Sellergren et al., 2026), and Lingshu (Xu et al., 2025), broaden the task scope to medical image interpretation, diagnosis, and report generation. Much of this progress, however, has been developed around 2D image interfaces or slice-based representations, which do not explicitly preserve volumetric context when a CT study is represented by a limited set of slices. Volumetric MLLMs such as M3D (Bai et al., 2024), CT-CHAT (Hamamci et al., 2026b), Merlin (Blankemeier et al., 2024b), and Photon (Fang et al., 2026) instead encode 3D scans and support tasks including volumetric question answering, diagnosis, localization, and report generation. RadFM (Wu et al., 2025) further provides a generalist interface for both 2D and 3D radiology data, while OmniCT explicitly studies joint slice-volume CT understanding. These studies substantially expand the input and task coverage of medical MLLMs, but unified modeling alone does not ensure that high-level predictions are grounded in fine-grained visual evidence. RadSight instead explicitly learns lesion-level attributes and spatial correspondence before disease-level reasoning and report generation, while preserving the native spatial structure of both 2D and 3D inputs.

Medical Vision-Language Benchmarks. Medical vision-language benchmarks have evolved from task-specific 2D question answering toward broader multimodal and volumetric evaluation. VQA-RAD (Lau et al., 2018), SLAKE (Liu et al., 2021), and PathVQA (He et al., 2020) evaluate question answering over radiology, semantically annotated medical images, and pathology images, respectively. PMC-VQA (Zhang et al., 2023), OmniMedVQA (Hu et al., 2024), and RadFig-VQA (Yamagishi et al., 2025) increase the scale or diversity of modalities and clinical questions. These benchmarks are valuable for measuring medical knowledge and image-level understanding, but they provide limited coverage of native volumetric perception and explicit bidirectional spatial correspondence. Recent 3D benchmarks address complementary aspects of this problem: M3D-Bench (Bai et al., 2024) evaluates multiple 3D medical tasks, 3D-RAD (Gai et al., 2026) and DeepTumorVQA (Chen et al., 2026) emphasize volumetric and tumor-related reasoning, CT-SpatialVQA (Monon et al., 2026) targets semantic-spatial understanding, and CT-FineBench (Yuan et al., 2026) evaluates fine-grained factual consistency in CT reports. Taken together, prior benchmarks provide strong coverage of individual capabilities, but the benchmarks considered here do not jointly trace the full evidence hierarchy from low-level attributes and spatial correspondence to diagnosis and report generation across both native 2D and 3D inputs. Perception-Bench complements these resources with a unified six-dimensional evaluation of attribute judgment, spatial grounding, spatial understanding, disease prediction, anomaly detection, and report generation, thereby exposing where perceptual errors emerge and how they propagate to higher-level clinical outputs.

B Data Construction Details

B.1 Data Sources

Table 6 summarizes the datasets used in this work, spanning both 2D chest X-ray and 3D CT modalities. We categorize each dataset by its annotation type: *Image-Report* denotes datasets with paired radiology reports, which we leverage for perception-oriented QA generation via our two-stage prompting pipeline, and *Image-Mask* denotes datasets with pixel/voxel-level segmentation annotations, which we use for grounding data construction.

For the 2D X-ray modality, MIMIC-CXR (Johnson et al., 2019) and IU-Xray (Demner-Fushman et al., 2016) serve as our primary report-based sources, covering a diverse range of thoracic pathologies across over 235,000 studies. PadChest-GR (de Castro et al., 2025), VinDr-CXR (Nguyen et al., 2022), MS-CXR (Boecking et al., 2022), and REFLACX (Bigolin Lanfredi et al., 2022) provide pixel-level annotations for anatomical structures and lesion regions, enabling spatial grounding supervision.

For the 3D CT modality, Merlin (Blankemeier et al., 2024a), CT-RATE (Hamamci et al., 2024b), Inspect (Huang et al., 2023), and AMOS-MM (Ji et al., 2022a) provide paired radiology reports covering chest and abdominal findings, from which we extract structured disease annotations and attribute-level QA pairs. AbdomenAtlas 3.0 (Bassi et al., 2025) contributes voxel-level segmentation

Table 6: Summary of datasets used in this work. "-" indicates that the corresponding statistic is not applicable or not reported by the original dataset.

Modality	Dataset	#Patients	#Studies	Type	URL
2D (X-ray)	MIMIC-CXR (Johnson et al., 2019)	65,379	227,835	Image-Report	https://physionet.org/content/mimic-cxr/
	IU-Xray (Demner-Fushman et al., 2016)	–	7,470	Image-Report	https://www.kaggle.com/datasets/raddar/chest-xrays-indiana-university
	PadChest-GR (de Castro et al., 2025)	–	4,555	Image-Mask	https://bimcv.cipf.es/bimcv-projects/padchest-gr/
	VinDr-CXR Nguyen et al. (2022)	–	15,000	Image-Mask	https://vindr.ai/datasets/cxr
	MS-CXR (Boecking et al., 2022)	–	1026	Image-Mask	https://physionet.org/content/ms-cxr/1.1.0/
	REFLACX (Bigolin Lanfredi et al., 2022)	–	3,056	Image-Mask	https://physionet.org/content/reflacx/
3D (CT)	Merlin (Blankemeier et al., 2024a)	18,317	25,494	Image-Report	https://stanfordaimi.azurewebsites.net/datasets/60b9c7ff-877b-48ce-96c3-0194c8205c40
	CT-RATE (Hamamci et al., 2024b)	21,304	50,188	Image-Report	https://huggingface.co/datasets/ibrahimhamamci/CT-RATE
	Inspect (Huang et al., 2023)	19,402	23,248	Image-Report	https://stanfordaimi.azurewebsites.net/datasets?term=INSPECT
	AbdomenAtlas 3.0 (Bassi et al., 2025)	–	9,262	Image-Report	https://huggingface.co/datasets/AbdomenAtlas/AbdomenAtlas3.0Mini
	AMOS-MM (Ji et al., 2022a)	–	1,690	Image-Report	https://era-ai-biomed.github.io/amos/dataset.html#download
	National Lung Screening Trial(NLST) (Team, 2011)	26,722	73,116	–	https://www.cancerimagingarchive.net/wp-content/uploads/manifest-NLST_allCT.tcia

masks for abdominal organs and lesions. The National Lung Screening Trial (NLST) (Team, 2011) is used for pretraining, providing a large-scale collection of over 73,000 low-dose chest CT studies.

In total, our data curation pipeline covers approximately 150,000 patients and over 440,000 studies, ensuring broad anatomical coverage and pathological diversity across both modalities.

B.2 Patient-Level Dataset Splits and Leakage Prevention

Except for NLST (Team, 2011), we followed the official splits provided by each source dataset. All datasets were partitioned at the patient level, ensuring that all studies from the same patient were assigned exclusively to one subset. For NLST, we created patient-level training and held-out subsets using a 9:1 ratio.

Notably, both our training corpus and DeepTumorVQA (Chen et al., 2026) include data from AbdomenAtlas 3.0 (Bassi et al., 2025). For this shared source, we adopted exactly the same patient-level partition as DeepTumorVQA. Consequently, no patient assigned to the DeepTumorVQA evaluation split was included in our training data, preventing patient-level data leakage between training and evaluation.

B.3 Report Parsing and Annotation Validation

We employ a two-stage prompting pipeline to convert free-text radiology reports into structured disease-level annotations with fine-grained attributes. In the first stage, the large language model (LLM) is instructed to perform organ-level decomposition followed by disease identification. Specifically, the model first segments the original report into organ-level sub-reports by mapping all descriptive sentences to standardized anatomical entities (e.g., "lung," "liver," "heart"), resolving anaphora and duplicating cross-organ sentences as needed to ensure each sub-report is self-contained. The model then identifies all diseases or clinical findings within each organ sub-report, normalizing synonymous or laterality-specific terms into standardized disease names (e.g., "left lower lobe opacity" and "bilateral opacities" are unified as "Lung Opacity"). Each identified finding is recorded with its associated organ, supporting evidence sentences, and a status label from present, absent, suspected, historical. In the second stage, conditioned on each extracted disease and a predefined set of candidate attributes, the model performs attribute extraction and question–answer (QA) pair generation. The model first determines which candidate attributes are explicitly described in the report for the target disease, then extracts the corresponding values, and finally constructs natural-language QA pairs grounded solely in the report text. Strict constraints are imposed throughout both stages to prevent hallucination: the model is required to use only information explicitly stated in the report, avoid merging or inferring attributes, and ensure each extracted value is directly supported by the source text. The diseases extracted from each dataset are summarized in Table 7. The complete prompt templates for both stages are provided in Figure 6 and Figure 7, respectively.

We apply quality control at both the disease-label and QA-pair levels. After synonym normalization, a team of six human reviewers manually screens the disease vocabulary to remove ambiguous, redundant, and clinically inconsistent labels. An LLM then performs a second-pass review of the manually curated vocabulary to identify residual errors, and only labels retained after both reviews are used in subsequent attribute extraction. For QA-level validation, we randomly sample 500 generated pairs and ask DeepSeek-V4-Pro, Kimi-K3, GLM-5.2, and Qwen3-Max to evaluate each pair using the same two criteria: factual accuracy, which measures whether the answer is explicitly supported by the source report, and question–answer consistency, which measures whether the answer

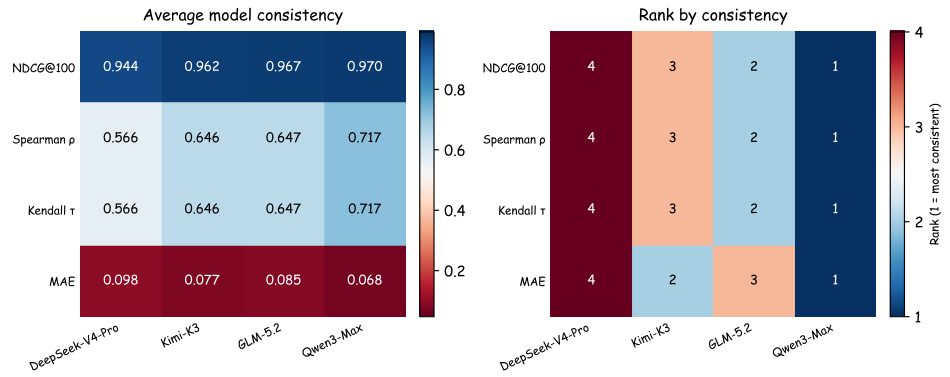


Figure 5: Cross-grader consistency on 500 randomly sampled report-derived QA pairs. For each grader, we report its average agreement with the other three graders using NDCG@100, Spearman correlation, Kendall correlation, and MAE; higher values indicate stronger agreement except for MAE. Qwen3-Max exhibits the strongest overall consistency.

addresses the requested attribute. We compare each grader with the other three using NDCG@100, Spearman correlation, Kendall correlation, and mean absolute error (MAE). Definitions of these agreement metrics are provided in Section D.1. As summarized in Figure 5, Qwen3-Max exhibits the strongest overall cross-grader consistency.

B.4 Prompt Design of Perception-oriented QA Generation

We employ a two-stage prompting pipeline to convert free-text radiology reports into structured disease-level annotations with fine-grained attributes. In the first stage, the large language model (LLM) is instructed to perform organ-level decomposition followed by disease identification. Specifically, the model first segments the original report into organ-level sub-reports by mapping all descriptive sentences to standardized anatomical entities (e.g., "lung," "liver," "heart"), resolving anaphora and duplicating cross-organ sentences as needed to ensure each sub-report is self-contained. The model then identifies all diseases or clinical findings within each organ sub-report, normalizing synonymous or laterality-specific terms into standardized disease names (e.g., "left lower lobe opacity" and "bilateral opacities" are unified as "Lung Opacity"). Each identified finding is recorded with its associated organ, supporting evidence sentences, and a status label from present, absent, suspected, historical.

In the second stage, conditioned on each extracted disease and a predefined set of candidate attributes, the model performs attribute extraction and question–answer (QA) pair generation. The model first determines which candidate attributes are explicitly described in the report for the target disease, then extracts the corresponding values, and finally constructs natural-language QA pairs grounded solely in the report text. Strict constraints are imposed throughout both stages to prevent hallucination: the model is required to use only information explicitly stated in the report, avoid merging or inferring attributes, and ensure each extracted value is directly supported by the source text. The diseases extracted from each dataset are summarized in Table 7. The complete prompt templates for both stages are provided in Figure 6 and Figure 7, respectively.

B.5 Training Data Distribution Details

We provide a detailed breakdown of the training data distribution across all four stages in Figures 8–11. As shown in Figure 8, our training pipeline follows a curriculum learning strategy with progressively specialized objectives. Stage 1 (423K samples) performs vision-language alignment on radiology image-report pairs, with 78.7% from 2D X-rays and 21.3% from 3D CT volumes. Stage 2 (4.6M samples) is the largest stage, focusing on fine-grained perception tasks including attribute-level VQA, spatial grounding, and visual understanding, with the modality ratio shifting toward 3D (70.9%). Stage 3 (3.0M samples) targets clinical diagnosis through anomaly detection and disease prediction, with a roughly balanced 2D/3D split (49.4% vs. 50.6%). Stage 4 (363K samples) trains

Table 7: Summary of diseases extracted from each dataset.

Modality	Dataset	Diseases	#
2D	MIMIC-CXR	Atelectasis, Atherosclerosis, Cardiomegaly, Consolidation, Edema, Emphysema, Enlarged Cardiomediastinum, Fracture, Hernia, Lung Lesion, Lung Opacity, Pleural Effusion, Pneumonia, Pneumothorax, Pulmonary Edema, Support Devices	16
	IU-Xray	Atelectasis, Atherosclerosis, Cardiomegaly, Degenerative Changes, Emphysema, Fracture, Lung Opacity, Pleural Effusion, Scoliosis, Support Devices	10
3D	CT-RATE	Aortic Aneurysm, Arterial Wall Calcification, Atelectasis, Bronchiectasis, Cardiomegaly, Consolidation, Coronary Artery Disease, Coronary Calcification, Emphysema, Fibrosis, Gallstones, Hepatic Steatosis, Hepatomegaly, Hernia, Interlobular Septal Thickening, Kidney Stone, Kyphosis, Liver Lesion, Lung Nodule, Lung Opacity, Lymphadenopathy, Medical Material, Mosaic Attenuation Pattern, Osteophyte, Peribronchial Thickening, Pericardial Effusion, Pleural Effusion, Pneumonia, Pulmonary Fibrotic Sequela, Renal Cyst, Scoliosis, Splenomegaly, Thyroid Nodule	33
	Merlin	Abscess, Anasarca, Aortic Valve Calcification, Appendicitis, Ascites, Atelectasis, Atherosclerosis, Biliary Ductal Dilation, Bladder Wall Thickening, Bowel Obstruction, Bronchiectasis, Cardiomegaly, Cirrhosis, Colitis, Coronary Calcification, Degenerative Disc Disease, Diverticulitis, Diverticulosis, Emphysema, Fracture, Free Air, Gallbladder Wall Thickening, Gallstones, Granuloma, Hemangioma, Hematoma, Hepatic Steatosis, Hepatomegaly, Hernia, Hydronephrosis, Lung Nodule, Lung Opacity, Lymphadenopathy, Osteopenia, Ovarian Cyst, Pancreatic Atrophy, Pancreatic Cyst, Pericardial Effusion, Pleural Effusion, Post-operative, Prostatomegaly, Renal Atrophy, Renal Calculus, Renal Cyst, Renal Hypodensities, Scoliosis, Splenomegaly, Submucosal Edema, Surgically Absent Gallbladder, Thrombosis, Uterine Fibroids, Varices	52
	Inspect	Ascites, Atelectasis, Bronchiectasis, Cardiomegaly, Emphysema, Fracture, Hepatic Steatosis, Hernia, Lung Nodule, Lung Opacity, Lymphadenopathy, Pericardial Effusion, Pleural Effusion, Pneumonia, Pneumothorax, Pulmonary Edema, Pulmonary Embolism, Pulmonary Hypertension, Splenomegaly, Thyroid Nodule	20
	AbdomenAtlas	Colon Lesion, Hepatomegaly, Kidney Enlargement, Kidney Lesion, Liver Lesion, Pancreas Lesion, Pancreatomegaly, Splenomegaly	8
	AMOS-MM	Ascites, Gallstones, Hepatic Steatosis, Hydronephrosis, Lung Opacity, Lymphadenopathy, Prostatomegaly, Renal Cyst, Renal Hypodensities	9

on medical report generation from both chest X-rays (75.1%) and CT volumes (24.9%). Figure 9 presents the distribution of Stage 2 attribute VQA data. Among the 12 question categories, Location (25.1%) is the most frequent, followed by Abnormality Type (12.3%), Size (10.6%), Texture (8.1%), and Severity (8.0%), reflecting the clinical priorities in radiology interpretation. Figure 10 illustrates the spatial grounding data, which spans 49 3D organ categories, 35 2D radiographic findings, and 3 lesion types (liver, kidney, and pancreatic lesions with 20K samples each). The organ categories cover major anatomical structures from the thorax to the pelvis, providing comprehensive spatial supervision for 3D localization. Figure 11 shows the disease distribution of Stage 3 anomaly detection data across 82 categories.

B.6 Perception-Bench Distribution

Perception-Bench contains 1,129,302 evaluation instances across six tasks. As shown in Figure 12(a), anomaly detection is the largest component, with 451,184 instances (40.0%). Spatial grounding and spatial understanding contain 196,484 (17.4%) and 196,437 (17.4%) instances, respectively, followed by attribute judgment with 181,744 (16.1%), disease prediction with 87,777 (7.8%), and report generation with 15,676 (1.4%) instances.

The remaining panels characterize the internal composition of the perception and report-generation tasks. Figure 12(b) reports the most frequent 3D anatomical structures and 2D radiographic findings used for spatial grounding. Figure 12(c) shows that the attribute-judgment subset spans 12 attribute types, with location (22.8%) and abnormality type (20.9%) being the most frequent. Figure 12(d) summarizes report lengths for 4,448 2D reports and 11,228 3D reports. The 3D reports are longer on average than the 2D reports (192 versus 57 words).

The anomaly-detection subset contains 451,184 evaluation instances covering 82 abnormality and imaging-finding categories. Figure 13 shows a long-tailed category distribution. The five most frequent categories contain 69,205 instances (15.3%), the body categories ranked 6–55 contain 325,799 instances (72.2%), and the tail categories ranked 56–82 contain 56,180 instances (12.5%). This

distribution documents the breadth and frequency variation of the clinical concepts evaluated by Perception-Bench.

C Experiments and Implementation Details

C.1 Data Processing Details

3D Volume Preprocessing. For 3D CT datasets that do not provide voxel-level annotations, we first run TotalSegmentator (Wasserthal et al., 2023) on each CT volume to generate dense anatomical segmentation masks. The masks and their corresponding CT volumes are resampled to an isotropic in-plane resolution of $1.0 \times 1.0\,\mathrm{mm}^2$ with a slice thickness of $5.0\,\mathrm{mm}$, using trilinear interpolation for images and nearest-neighbor interpolation for masks. CT intensities are then clipped to $[-1000, 1000]$ HU and linearly normalized to $[0, 1]$. We compute the tight bounding box of all non-zero mask voxels, extend it by 5 voxels along the depth axis and 20 voxels along the height and width axes to preserve surrounding anatomical context, and zero-pad the cropped volume to a fixed spatial size of $96 \times 256 \times 384$ (depth $\times$ height $\times$ width).

3D Bounding Box Extraction. Given the processed segmentation masks, we extract 3D bounding boxes using two strategies depending on the anatomical category. For normal organs, we compute the tight axis-aligned bounding box enclosing all voxels of each label. For lesion classes (*e.g.*, kidney lesion, liver lesion, and pancreatic lesion), where multiple spatially disjoint instances may coexist in a single volume, we perform 3D connected component analysis with 26-connectivity to identify individual lesion instances. Connected components with fewer than 10 voxels are discarded as noise, and only the three largest instances (by voxel count) are retained for each lesion class. A per-instance bounding box is then computed for each retained component. All bounding box coordinates are normalized to $[0, 1]$ relative to the volume dimensions before being used as grounding supervision.

2D Bounding Box Extraction. All 2D chest X-ray images are resized to 448×448 pixels. We extract disease-level bounding boxes from each dataset's native annotation format: for REFLACX, we parse the per-radiologist ellipse annotations and the chest region bounding box; for VinDr-CXR, we select annotations from the highest-priority radiologist with fallback to alternative annotators when the primary is absent, and use the consensus labels for the test set; for PadChest, we convert the normalized coordinates in the grounding annotations to pixel space. Across all datasets, bounding boxes belonging to the same disease category within a single image are aggregated into one sample, with duplicate boxes removed. All bounding box coordinates are then normalized to $[0, 1]$ relative to the resized image dimensions.

C.2 Model Architecture Details

Our model adopts a dual-encoder multimodal large language model architecture to jointly process 2D radiographic images and 3D CT volumes. For 2D inputs (*e.g.*, chest X-rays), we use the SigLIP-NaViT encoder from VideoLLaMA3 (Zhang et al., 2025), whose patch features are projected into the LLM embedding space via a two-layer MLP with GELU activation. For 3D CT volumes, we employ a convolutional encoder based on the PlainConvUNet architecture from nnU-Net (Isensee et al., 2021). Features from the deepest encoder stage are projected via a $1 \times 1 \times 1$ convolution, flattened into a 1D token sequence, and mapped into the LLM space through a two-layer MLP projector. The 3D encoder is initialized from a pretrained segmentation model to leverage learned volumetric representations. We use Qwen3-VL (Bai et al., 2025) as the language backbone, with 4B and 8B variants explored in our experiments. Visual tokens from either encoder are substituted into the corresponding placeholder positions in the input sequence, enabling the LLM to jointly attend to visual and textual information in a unified autoregressive framework.

C.3 Public Benchmarks Details

To verify the perceptual capabilities of our model, we jointly fine-tune the Stage 4 checkpoint on all public benchmarks and evaluate on their respective test sets. Table 8 summarizes the benchmarks used, covering both 2D and 3D medical VQA tasks with open-ended, closed-ended, and multi-choice answer formats.

Table 8: Statistics of downstream evaluation benchmarks.

Modality	Dataset	Train	Test	Answer Type
2D	SLAKE (Liu et al., 2021)	9,835	2,094	Open + Closed
	OmniMedVQA (Hu et al., 2024)	142,410	35,581	Multi-Choice
	PMC-VQA (Zhang et al., 2023)	152,603	33,430	Multi-Choice
	RadFig (Yamagishi et al., 2025)	228,155	10,139	Multi-Choice
3D	3D-Rad (Gai et al., 2026)	134,701	33,712	Open + Closed
	DeepTumor (Chen et al., 2026)	428,050	10,000	Multi-Choice
	CT-Spatial-VQA (Monon et al., 2026)	–	8,953	Open

Table 9: Training hyperparameters for each stage. "–" indicates the module is frozen. BS: batch size. LR: learning rate. Optim.: optimizer. Sched.: learning rate scheduler.

Model	Stage	Data Size	Epochs	Global BS	Per-device BS	LLM LR	Projector LR	Encoder LR	Optim. / Sched. / Warmup
4B	1	423K	2	1024	8	–	1e-3	1e-5	AdamW / Cosine / 0.03
	2	4.56M	1	1024	8	1e-5	1e-5	2e-6	AdamW / Cosine / 0.03
	3	3.93M	1	1024	8	1e-5	1e-5	2e-6	AdamW / Cosine / 0.03
	4	2.12M	2	1024	8	1e-5	1e-5	2e-6	AdamW / Cosine / 0.03
8B	1	423K	2	512	8	–	1e-3	1e-5	AdamW / Cosine / 0.03
	2	4.56M	1	512	4	1e-5	1e-5	2e-6	AdamW / Cosine / 0.03
	3	3.93M	1	512	4	1e-5	1e-5	2e-6	AdamW / Cosine / 0.03
	4	2.12M	2	512	4	1e-5	1e-5	2e-6	AdamW / Cosine / 0.03

C.4 Training Details

We adopt a four-stage curriculum training strategy, progressively introducing tasks of increasing complexity. Table 9 summarizes the training configuration of each stage for both the 4B and 8B model variants. All stages are trained on 32×H20 GPUs with DeepSpeed ZeRO-Stage 1, gradient checkpointing, and BF16 mixed precision. To mitigate catastrophic forgetting of previously acquired capabilities, we employ a data replay strategy that mixes a subset of earlier-stage data into subsequent training stages. In Stage 3, we replay 20% of Stage 2 data while training on the full anomaly detection dataset. In Stage 4, we further replay 10% of Stage 2 and 20% of Stage 3 data alongside the report generation data. For the anomaly detection subset in Stage 3 replay, we apply disease-level balanced sampling to maintain a 1:1 positive-to-negative ratio for each disease category. When the current-stage data is smaller than the replayed data, we upsample it via cyclic repetition to match the total volume.

D Evaluation Metrics

D.1 Cross-Judge Agreement Metrics

We use five complementary metrics to assess ranking and scoring agreement among LLM judges.

NDCG@100 (Normalized Discounted Cumulative Gain): Measures how well one judge's top-100 ranking aligns with another judge's scores, assigning greater weight to higher-ranked predictions.

Spearman's Rank Correlation (ρ): Measures agreement in the overall ordering of predictions through the correlation between the ranks assigned by two judges.

Kendall's Tau (τ): Measures whether two judges consistently order pairs of predictions in the same direction.

MAE (Mean Absolute Error): Measures the average absolute difference between two judges' scores, capturing disagreement on the original scoring scale.

Quadratic Weighted Kappa (κ_w): Measures agreement beyond chance while accounting for the ordinal scoring scale, penalizing larger score differences more heavily.

Higher values indicate stronger agreement for all metrics except MAE, for which lower values indicate closer agreement.

D.2 PROMPT-BASED LLM EVALUATION FOR ATTRIBUTE JUDGMENT

Prompt Template for Attribute Judgment LLM Evaluation

Please adhere strictly to the following rules:

Score 1 (Correct / Fully Acceptable):
The predict answer is fully correct and semantically equivalent to the reference answer, or provides a correct, more detailed version of it.

1. **Exact or Semantic Match:** The answer is identical or uses different wording (synonyms, rephrasing) to convey the exact same meaning.
   Example: Ref: not pathological, Pred: not pathologically enlarged.
2. **Correct but More Specific (Acceptable):** The predict answer is a correct, but more specific, instance of the reference answer.
   Example: Ref: left lower lobe, Pred: left lower lobe of the lung.
3. **Contains Correct Additional Detail (Acceptable):** The predict answer includes all key information from the reference and adds other correct, relevant details.
   Example: Ref: sliding type, Pred: sliding type, Type I.
4. **Numerical Answers:** When both answers are numbers, the predict answer matches exactly or is within a very close tolerance (e.g., $\pm 10\%$) of the reference answer.
   Example: Ref: 10 mm, Pred: 9.5 mm.
5. **Multi-dimensional Size Measurements:** When the reference answer contains a multi-dimensional measurement (e.g., "A $\times$ B cm") and the predict answer provides only a single dimension, use the largest dimension of the reference as the basis for comparison, then apply the standard numerical tolerance rules.
   Example: Ref: 1.9 $\times$ 0.5 cm, Pred: 1.8 cm $\rightarrow$ compare 1.8 cm against 1.9 cm (the largest dimension) $\rightarrow$ difference is $\sim$5%, within $\pm 10\%$ $\rightarrow$ Score 1.
   Example: Ref: 3.0 $\times$ 1.2 cm, Pred: 2.8 cm $\rightarrow$ compare 2.8 cm against 3.0 cm $\rightarrow$ difference is $\sim$7%, within $\pm 10\%$ $\rightarrow$ Score 1.
   Example: Ref: 3.0 $\times$ 1.2 cm, Pred: 1.1 cm $\rightarrow$ compare 1.1 cm against 3.0 cm (largest) and 1.2 cm (second) $\rightarrow$ 1.1 cm is close to 1.2 cm (second dimension), not the largest $\rightarrow$ Score 0.

Score 0.5 (Partially Correct):
The predict answer is on the right track but is flawed by being incomplete or too general.

1. **Correct but Incomplete:** The predict answer provides correct information but omits some key elements from the reference answer.
   Example: Ref: ground-glass, consolidation, Pred: ground-glass.
   Example: Ref: right lung middle lobe, left lingular segment, Pred: right lung middle lobe.
2. **Correct but Overly General:** The predict answer is a correct but less specific version of the reference answer. It captures the essence but loses important detail.
   Example 1 (Qualitative): Ref: 6 mm, Pred: millimetric.
   Example 2 (Location): Ref: left lung lower lobe superior segment, right lung lower lobe, Pred: both lungs, more prominently in the lower lobes.
3. **Numerical Answers:** The predict answer is numerically in the same ballpark (same order of magnitude) but outside the strict tolerance for a score of 1. It captures part of the quantitative information but is imprecise.
   Example: Ref: 8 mm, Pred: 10 mm. (Correct order of magnitude, 30% $>$ difference $>$ 10%).
4. **Multi-dimensional Size Measurements:** When the reference contains a multi-dimensional measurement and the predict answer provides only a single dimension that matches the largest dimension within the 0.5-score tolerance range (10%$\sim$30% difference), assign Score 0.5.
   Example: Ref: 2.0 $\times$ 0.8 cm, Pred: 1.6 cm $\rightarrow$ compare against 2.0 cm $\rightarrow$ difference is 20%, within 10%$\sim$30% range $\rightarrow$ Score 0.5.

5. **Correct Laterality but Wrong Lobe:** The predict answer correctly identifies the side (left/right) but specifies a wrong lobe or segment within that side. Since the laterality is correct but the intra-lung localization is wrong, this is partially correct.
   Example 1: Ref: right lung upper lobe, Pred: right lung lower lobe → correct side (right), wrong lobe → Score 0.5.
   Example 2: Ref: left lung lower lobe, Pred: left lung upper lobe → correct side (left), wrong lobe → Score 0.5.
   Example 3: Ref: right lung middle lobe, Pred: right lung lower lobe → correct side (right), wrong lobe → Score 0.5.
   Note: This rule only applies when the laterality is unambiguously correct and only the lobe/segment is wrong. If the side itself is wrong or expanded to bilateral, apply Score 0 rules instead.

Score 0 (Incorrect):
The predict answer is factually wrong, irrelevant, or fails to answer the question.

1. **Contradiction:** It directly contradicts the reference answer.
   Example: Ref: pathological, Pred: not pathologically enlarged.
2. **Wrong Information:** It provides a completely different piece of information.
   Example: Ref: atelectasis, Pred: infectious process.
3. **Hallucination / Irrelevant:** It provides information not supported by the context or states that an answer cannot be determined when the reference provides one.
   Example: Ref: focal, Pred: The report does not provide this information.
4. **Numerical Answers:** The predict answer is of a different order of magnitude or significantly incorrect (difference $> 30\%$).
   Example: Ref: 10 mm, Pred: 1 mm or 50 mm.
5. **Wrong Laterality or Wrong Side:** The reference specifies a particular side (e.g., left), but the predict answer states the opposite side (e.g., right) or expands it to bilateral/both sides. Changing or broadening the laterality is always incorrect and must be scored 0.
   Example 1: Ref: left lung, Pred: right lung. (opposite side → Score 0)
   Example 2: Ref: left lung, Pred: bilateral / both lungs. (unilateral expanded to bilateral → Score 0)
   Example 3: Ref: right kidney, Pred: both kidneys. (unilateral expanded to bilateral → Score 0)
6. **Multi-dimensional Size Measurements:** When the predict answer provides only a single dimension that does not match the largest dimension of the reference (difference $> 30\%$), assign Score 0.
   Example: Ref: 3.0×1.2 cm, Pred: 1.1 cm → does not match largest dimension (3.0 cm), difference $> 30\%$ → Score 0.

Instruction
You are an expert evaluator for medical question answering systems. Your task is to score a predict answer against a reference answer for a given question. You must evaluate the semantic accuracy and completeness of the predict answer with the above scoring criteria. Your final output must consist of a score from the set $\{0, 0.5, 1\}$. You need to think first, and then give the final score. The final score should be enclosed in `<score>` and `</score>`.
Example: your thinking process. `<score>` 1 or 0.5 or 0 `</score>`

Input
Question: `<question>`
Reference Answer: `<ref>`
Predict Answer: `<prd>`

D.3 Cross-Judge Agreement for Attribute Judgment

To assess the reliability of Qwen3-Max as the primary judge for attribute judgment in Table 1, we compare its scores with those of three additional evaluators: DeepSeek-V4-Pro, Kimi-K3, and GLM-5.2. All four judges independently evaluate the same predictions using the rubric in Section D.2, assigning scores of 0, 0.5, or 1. Each judge receives only the question, reference answer, and prediction; neither the image nor the identity of the evaluated model is provided.

The audit comprises 200 RadSight predictions sampled from the attribute-judgment split of Perception-Bench, stratified across size (32), location (32), count (24), density (24), morphology (20), margin (16), severity (16), distribution (16), and type (20). Within each category, the sample includes near-matches, partial overlaps, and clear mismatches, ensuring that the analysis covers a range of answer quality rather than only unambiguous correct responses. We assess cross-judge agreement using the five metrics defined in Section D.1.

Figure 14 shows strong agreement across evaluators. The left panel reports each judge's average pairwise NDCG@100, Spearman ρ, Kendall τ, and MAE against the other three judges. Qwen3-Max achieves 0.970, 0.916, 0.900, and 0.048, respectively, closely matching the agreement levels of the alternative judges. Across all four evaluators, Spearman correlations range from 0.901 to 0.919 and MAEs from 0.046 to 0.052. Qwen3-Max therefore exhibits consistency comparable to the rest of the evaluator panel.

The right panel further confirms this agreement using pairwise quadratic weighted κ, which ranges from 0.892 to 0.937. For Qwen3-Max, agreement with DeepSeek-V4-Pro, Kimi-K3, and GLM-5.2 reaches κ values of 0.893, 0.908, and 0.937, with exact agreement rates of 91.5%, 92.0%, and 93.5%, respectively. All four judges assign identical scores to 170 of the 200 predictions (85.0%). Among the remaining 30 items, the maximum difference between judges is 0.5 for 22 items and 1 for eight items. Qwen3-Max differs from the majority of the other three judges on nine items (4.5%).

Together, these results support Qwen3-Max as a reasonable primary evaluator under the adopted rubric: its scores show strong agreement with alternative judges, and its disagreement levels are comparable to those observed among the other evaluators.

D.4 Grounding Metrics

We evaluate the spatial grounding capability of our model using both 2D and 3D bounding box predictions, depending on the imaging modality. For 2D modalities (*e.g.*, X-ray), each bounding box is represented as $\mathbf{b} = (x_1, y_1, x_2, y_2)$, where (x_1, y_1) and (x_2, y_2) denote the top-left and bottom-right corners, respectively. For 3D volumetric modalities (*e.g.*, CT), bounding boxes are extended to $\mathbf{b} = (x_1, y_1, z_1, x_2, y_2, z_2)$, where the additional coordinates specify the extent along the axial (depth) dimension.

Intersection over Union (IoU). Given a predicted bounding box $\mathbf{b}^p$ and a ground-truth bounding box $\mathbf{b}^g$, IoU is defined as

$$\mathrm{IoU}(\mathbf{b}^p, \mathbf{b}^g) = \frac{|\mathbf{b}^p \cap \mathbf{b}^g|}{|\mathbf{b}^p \cup \mathbf{b}^g|}, \tag{4}$$

where $|\cdot|$ denotes area for 2D boxes and volume for 3D boxes.

Detection Metrics. For spatial detection, we use standard one-to-one box matching. A predicted box is counted as a true positive if its IoU with the assigned ground-truth box is at least $\tau = 0.3$. Unmatched predictions and ground-truth boxes are counted as false positives and false negatives, respectively. Precision, Recall, and F1-score are computed as

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \quad \mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \quad \mathrm{F1} = \frac{2 \cdot \mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}. \tag{5}$$

For localization quality, we report mIoU as the mean IoU over all evaluated localization instances. The IoU threshold τ is used only for detection metrics and does not filter the continuous IoU values used to compute mIoU.

D.5 Clinical Evaluation Metrics

Why is AUC used as the evaluation metric for anomaly detection? Many previous anomaly-detection studies primarily report accuracy (ACC) as the evaluation metric (Gai et al., 2026). However, ACC depends strongly on both the class distribution and the decision threshold, and may therefore overestimate diagnostic performance under severe class imbalance.

Taking the CT-RATE dataset as an example, we formulate 18 disease-specific questions for each patient. Constructing the training set according to the natural label distribution results in a positive-to-negative ratio of approximately $1{:}4.2$. Under this distribution, a model can achieve high ACC by favoring the majority negative class, even when it rarely identifies positive cases. As shown in Figure 15 a, the model trained with the original imbalanced distribution achieves an overall ACC of 83.75%. However, its positive accuracy is only 30.38%, compared with a negative accuracy of 96.45%. For four diseases, the model produces no positive predictions, resulting in an F1 score of zero. The apparently high ACC therefore primarily reflects the dominance of negative samples rather than reliable diagnostic ability.

We further reconstruct the training set using a balanced $1{:}1$ positive-to-negative ratio. After balancing, positive accuracy increases from 30.38% to 71.92%, and the disease-level F1 scores improve, including recovery from zero for the four previously undetected diseases. Nevertheless, overall ACC decreases from 83.75% to 71.65%. This counter-intuitive result shows that ACC may penalize a model for becoming more sensitive to clinically relevant positive cases. By contrast, the macro-average area under the receiver operating characteristic curve (AUC) remains nearly unchanged: 79.69% under imbalanced training and 79.04% after balancing, corresponding to a difference of only 0.65 pp. The disease-level comparison further shows that the median absolute change in positive accuracy is 51.2 pp (Figure 15 b), whereas the corresponding change in AUC is only 0.5 pp (Figure 15 c).

Unlike ACC, which evaluates binary predictions at a single operating threshold, AUC measures how consistently positive cases are ranked above negative cases across possible thresholds. It is therefore less sensitive to a particular decision threshold and to the class prevalence of the evaluation set. These results do not imply that AUC is mathematically invariant to the training distribution; rebalancing may still alter the learned ranking function. Rather, they demonstrate that, in our experiments, the underlying discriminative ability remains stable while the threshold-dependent binary decisions change substantially. We use AUC as the primary metric for anomaly detection.

Area Under the ROC Curve (AUC). For binary clinical diagnosis tasks (*e.g.*, determining whether a specific disease is present), we frame the evaluation as a binary classification problem. The model produces a confidence score $p \in [0, 1]$ representing the estimated probability that the target pathology is present ("yes"). Given a decision threshold t, samples with $p \geq t$ are classified as positive. By varying t across all possible values, we obtain the Receiver Operating Characteristic (ROC) curve, which plots the True Positive Rate (TPR, *i.e.*, Sensitivity) against the False Positive Rate (FPR, *i.e.*, $1 -$ Specificity):

$$\mathrm{TPR}(t) = \frac{\mathrm{TP}(t)}{\mathrm{TP}(t) + \mathrm{FN}(t)}, \quad \mathrm{FPR}(t) = \frac{\mathrm{FP}(t)}{\mathrm{FP}(t) + \mathrm{TN}(t)}, \tag{6}$$

where TP, FP, TN, and FN denote the number of true positives, false positives, true negatives, and false negatives at threshold t, respectively. The AUC is then defined as the area under this curve:

$$\mathrm{AUC} = \int_0^1 \mathrm{TPR}\big(\mathrm{FPR}^{-1}(x)\big)\, dx. \tag{7}$$

An AUC of 1.0 indicates perfect discrimination, while 0.5 corresponds to random chance. Importantly, AUC is a threshold-independent metric, making it robust to class imbalance — a common characteristic of clinical datasets where positive (diseased) cases are often significantly outnumbered by negative (healthy) ones.

To derive clinically interpretable operating points, we additionally adopt the Youden's index (Youden, 1950) to select the optimal binarization threshold:

$$t^* = \arg\max_t \Big\{ \mathrm{TPR}(t) - \mathrm{FPR}(t) \Big\} = \arg\max_t \Big\{ \mathrm{Sensitivity}(t) + \mathrm{Specificity}(t) - 1 \Big\}. \tag{8}$$

At the optimal threshold t^*, we report Sensitivity and Specificity as secondary metrics:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}. \tag{9}$$

All metrics are computed per disease category within each imaging modality, and we report the macro-averaged AUC across diseases as the primary aggregate metric.

Multiple-Choice Accuracy (ACC). For clinical knowledge questions presented in multiple-choice format, we evaluate accuracy by comparing the model's selected option against the ground-truth answer. For single-answer questions, a prediction is scored 1 if it matches the correct option and 0 otherwise. For multi-answer questions, we adopt a partial-credit scoring scheme: if the predicted option set $\mathcal{P}$ contains any incorrect option (*i.e.*, $\mathcal{P} \not\subseteq \mathcal{G}$, where $\mathcal{G}$ is the ground-truth set), the score is 0; otherwise, the score is proportional to the number of correctly selected options:

$$\text{Score} = \begin{cases} 0, & \text{if } \mathcal{P} \not\subseteq \mathcal{G}, \\ \dfrac{|\mathcal{P} \cap \mathcal{G}|}{|\mathcal{G}|}, & \text{otherwise.} \end{cases} \tag{10}$$

The overall accuracy is the mean score across all samples.

D.6 Report Generation Evaluation Metrics.

Unlike most prior works that adopt NLG metrics (*e.g.*, BLEU (Papineni et al., 2002), ROUGE (Lin, 2004)) for report evaluation, we note that such n-gram overlap metrics are ill-suited for clinical settings: they treat all tokens equally, thus failing to capture the impact of negation words (*e.g.*, "no evidence of") that reverse diagnostic conclusions, and penalize valid medical synonyms (*e.g.*, "heart enlargement" *vs.* "cardiomegaly") despite identical clinical semantics. Instead, we adopt **Clinical Efficacy (CE) F1** and **RaTEScore** (Zhao et al., 2024), which evaluate factual correctness at the pathology-label and entity levels, respectively.

Clinical Efficacy F1. We adopt Clinical Efficacy (CE) metrics to assess whether the generated report conveys the same clinical findings as the reference, by comparing pathology label vectors rather than surface text. The ground-truth labels are obtained directly from dataset annotations. For CT-RATE (Hamamci et al., 2024b), 18 thoracic pathology categories are provided; for MIMIC-CXR (Johnson et al., 2019), 14 pathology labels are extracted by CheXbert (Smit et al., 2020). Predicted labels are extracted from the generated reports using a RadBERT-based classifier (Yan et al., 2022) (for CT-RATE) or CheXbert (for MIMIC-CXR). We report Macro F1, Micro F1, and sample-wise F1 to measure the agreement between ground-truth and predicted label vectors.

RaTEScore. RaTEScore (Zhao et al., 2024) is an entity-aware metric that extracts medical entities from both the reference and candidate reports via a fine-tuned NER model, encodes them with BioLORD (Remy et al., 2024), and computes entity-level precision and recall through type-weighted cosine similarity matching with negation penalties. The final score is the harmonic mean of precision and recall.

E Additional Experimental Results

E.1 Ablation Studies

Impact of Perceptual Curriculum Learning on 2D X-ray Report Generation. Since publicly available 2D medical benchmarks typically aggregate multiple imaging modalities (e.g., X-ray, MRI, CT), isolating the contribution of each training stage to a specific modality is challenging. Given that our 2D training data is predominantly composed of chest X-rays, we conduct the ablation study on MIMIC-CXR to ensure a consistent and controlled evaluation, as shown in Figure 16. We selectively remove Stage 2 and Stage 3 from the training pipeline. Removing both stages yields the lowest performance (Sample F1: 27.60), confirming that directly fine-tuning from alignment is insufficient. Adding Stage 2 alone improves Sample F1 to 29.30 (+1.70), while adding Stage 3 alone yields a larger gain (34.55, +6.95), indicating that diagnostic reasoning contributes more than perceptual grounding. The full model achieves the best result (Sample F1: 35.95, Macro F1: 28.81),

validating that progressively building from perception to diagnosis enables increasingly effective report generation, and confirming that perceptual curriculum learning is also effective for 2D X-ray.

Curriculum Learning vs. Joint Mixed Training. To validate the necessity of our progressive training strategy, we compare the four-stage curriculum learning pipeline against a joint mixed training baseline that trains on all tasks simultaneously without stage separation. As shown in Figure 17, curriculum learning consistently outperforms joint mixed training across all ten evaluation metrics spanning perception, diagnosis, and interpretation. The most pronounced gains appear in CT-Rate F1 (+20.8) and AUC (+11.3), indicating that staged training is especially beneficial for tasks requiring complex clinical reasoning. Even for perception-level metrics such as 3D mIoU (+5.3) and 2D mIoU (+4.9), the progressive strategy yields notable improvements, suggesting that learning lower-level visual grounding before higher-level reasoning helps the model develop more robust spatial representations. These results confirm that simply mixing all training data cannot substitute for a carefully designed curriculum, and that progressively building from alignment through perception and diagnosis to interpretation is essential for achieving strong performance across the full spectrum of medical vision-language tasks.

E.2 Full Experimental Results

E.2.1 Spatial Grounding

Lesion-specific 3D spatial grounding. Attribute judgment and spatial grounding assess complementary but distinct capabilities. Attribute judgment asks the model to describe lesion properties such as anatomical location, size, and morphology in clinical language, whereas spatial grounding requires the prediction of one or more precise 3D bounding boxes. The latter is substantially more challenging for small, irregular, or multifocal lesions distributed throughout a large CT volume. Table 10 reports lesion-specific grounding results for RadSight, with mIoU ranging from 2.60 to 8.30 across liver, kidney, and pancreatic lesions. These results indicate that coordinate-level lesion grounding remains difficult under limited lesion-specific supervision and should be interpreted separately from both attribute-level QA and the aggregate 3D grounding score, which primarily reflects anatomical-structure localization. Figure 21 provides representative qualitative examples but does not imply consistently accurate localization of small lesions.

Table 10: Spatial grounding performance on lesion detection.

Disease	mIoU
Pancreatic lesion	8.30
Kidney lesion	7.60
Liver lesion	2.60

Zero-Shot Spatial Grounding on Additional Benchmarks. Table 12 reports the zero-shot spatial grounding performance on BTCV (Landman et al., 2015) and AMOS-22 (Ji et al., 2022b), two widely-used abdominal organ segmentation benchmarks. Without any task-specific fine-tuning, our RadSight significantly outperforms M3D on both datasets, achieving 47.54 mIoU and 80.67 F1 on BTCV, and 44.57 mIoU and 75.62 F1 on AMOS-22, improving F1 by 51.47 points on BTCV and 25.89 points on AMOS-22. These results demonstrate the strong generalization ability of our model for 3D anatomical localization in unseen datasets.

E.2.2 Report Generation

Report Generation on CT-Rate. Table 13 presents the report generation performance on CT-RATE, evaluated by Clinical Efficacy (CE) metrics using a RadBERT-based classifier. Our method achieves the highest F1 score of 36.5, outperforming the best baseline BTB3D by a large margin of 10.7. Notably, while prior methods such as CT2Rep and CT-CHAT attain competitive Precision, their Recall remains below 16, indicating a tendency to generate conservative reports that miss many positive findings. In contrast, our method achieves a substantially higher Recall of 32.0 while maintaining the best Precision of 45.3, demonstrating a better balance between diagnostic coverage and correctness. We provide the per-disease breakdown of our method in Table 14.

Table 11: Per-disease performance of abnormality detection on the combined X-ray and CT test set. We report the area under the ROC curve (AUC), sensitivity (Sen), and specificity (Spe) for each of the 82 evaluated diseases. The global macro-averaged AUC is 81.10%.

Disease	AUC	Sen	Spe	Disease	AUC	Sen	Spe
Abscess	86.44	83.33	76.36	Kidney Lesion	92.09	85.71	82.65
Anasarca	92.99	89.84	82.63	Kidney Stone	71.39	52.00	90.64
Aortic Aneurysm	82.93	100.0	60.75	Kyphosis	76.57	85.71	56.71
Aortic Valve Calcification	90.01	87.63	79.11	Liver Lesion	85.23	72.15	84.91
Appendicitis	89.70	83.94	83.14	Lung Lesion	68.23	85.48	42.51
Arterial Wall Calcification	94.39	91.76	85.65	Lung Nodule	69.56	77.48	51.13
Ascites	84.15	82.09	70.63	Lung Opacity	72.85	66.45	66.44
Atelectasis	76.61	74.64	64.42	Lymphadenopathy	74.35	70.02	66.78
Atherosclerosis	87.09	81.06	78.12	Medical Material	90.99	80.38	85.70
Biliary Ductal Dilation	92.81	80.08	90.72	Mosaic Attenuation Pattern	84.32	76.98	77.89
Bladder Wall Thickening	73.19	67.20	69.49	Osteopenia	70.40	67.83	61.65
Bowel Obstruction	93.87	86.26	90.19	Osteophyte	73.57	78.95	64.02
Bronchiectasis	73.31	77.46	57.01	Ovarian Cyst	82.83	82.54	66.91
Cardiomegaly	88.19	88.64	73.89	Pancreas Lesion	82.53	82.73	68.38
Cirrhosis	97.34	95.88	92.58	Pancreatic Atrophy	80.90	77.54	71.40
Colitis	80.53	85.25	67.24	Pancreatic Cyst	71.87	52.05	80.19
Colon Lesion	98.85	95.00	97.68	Pancreatomegaly	79.91	69.05	74.51
Consolidation	75.81	70.83	67.30	Peribronchial Thickening	79.02	84.00	60.76
Coronary Artery Disease	83.56	91.49	72.51	Pericardial Effusion	83.63	66.42	85.89
Coronary Calcification	89.89	83.73	80.57	Pleural Effusion	94.14	90.10	84.26
Degenerative Changes	65.41	62.50	60.89	Pneumonia	74.53	83.30	54.37
Degenerative Disc Disease	67.88	95.65	30.72	Pneumothorax	81.25	83.54	65.03
Diverticulitis	88.59	82.17	78.56	Post-operative	89.18	78.58	85.39
Diverticulosis	78.69	75.31	68.68	Prostatomegaly	89.96	89.53	77.88
Edema	82.62	83.89	65.89	Pulmonary Edema	87.06	83.13	75.13
Emphysema	81.46	86.91	62.73	Pulmonary Embolism	64.95	60.47	59.61
Enlarged Cardiomediastinum	67.92	79.47	49.29	Pulmonary Fibrotic Sequela	68.67	68.62	60.69
Fibrosis	65.17	44.44	80.09	Pulmonary Hypertension	80.87	69.23	77.65
Fracture	68.23	51.75	75.26	Renal Atrophy	88.77	85.71	78.15
Free Air	72.56	58.84	74.46	Renal Calculus	75.56	55.96	84.49
Gallbladder Wall Thickening	89.23	95.24	71.02	Renal Cyst	82.81	73.93	75.27
Gallstones	90.00	74.44	90.85	Renal Hypodensities	64.84	69.36	52.11
Granuloma	65.32	54.39	68.69	Scoliosis	68.93	87.38	40.69
Hemangioma	73.46	61.31	72.51	Splenomegaly	94.88	90.87	86.09
Hematoma	69.67	92.00	42.11	Submucosal Edema	74.24	69.84	69.21
Hepatic Steatosis	84.59	70.16	80.62	Support Devices	88.66	78.47	85.65
Hepatomegaly	87.42	74.12	84.13	Surgically Absent Gallbladder	97.15	97.06	93.06
Hernia	78.47	76.01	67.29	Thrombosis	77.18	68.00	71.92
Hydronephrosis	88.09	77.72	85.07	Thyroid Nodule	63.80	58.65	63.09
Interlobular Septal Thickening	85.73	81.45	75.56	Uterine Fibroids	84.72	85.94	70.03
Kidney Enlargement	90.81	87.81	76.73	Varices	94.97	94.87	82.90
Global Mean:	**AUC = 81.10**	**Sen = 78.02**	**Spe = 72.35**				

Table 12: Zero-shot spatial grounding performance on BTCV and AMOS-22. F1 is computed at IoU threshold $\tau = 0.3$.

Model	BTCV-mIoU	BTCV-F1	AMOS-22-mIoU	AMOS-22-F1
M3D-Phi-3-4B	19.78	29.20	29.29	49.73
RadSight-4B	**47.54**	**80.67**	**44.57**	**75.62**

Report Generation on MIMIC-CXR. Table 15 compares the report generation performance of various models on the MIMIC-CXR dataset, evaluated by sample-wise CheXbert F1, RaTEScore, and macro-averaged CheXbert F1 across 14 pathology categories. Among all compared methods, RadSight-8B achieves the best Sample F1 of 36.32 and the highest Macro F1 of 29.87, while MedGemma-1.5-4B obtains the best RaTEScore of 56.56. General-purpose vision-language models (*e.g.*, Qwen3-VL-8B, InternVL-3.5-8B) lag behind domain-specific medical models, highlighting the importance of medical knowledge for clinical report generation. Notably, all methods exhibit

Table 13: Report generation performance on CT-RATE. Results of compared methods are from (Hamamci et al., 2026a).

Method	Prec	Recall	F1
CT2Rep (Hamamci et al., 2024a)	43.5	12.8	16.0
CT-CHAT (Hamamci et al., 2026b)	45.0	15.8	18.4
Merlin (Blankemeier et al., 2024b)	29.5	11.2	16.0
BTB3D (Hamamci et al., 2026a)	26.0	26.0	25.8
Ours	**45.3**	**32.0**	**36.5**

Table 14: Per-disease Clinical Efficacy results on CT-RATE.

Disease	**F1**	**Precision**	**Recall**
Medical material	29.41	58.82	19.61
Arterial wall calcification	70.00	76.69	64.39
Cardiomegaly	53.33	64.86	45.28
Pericardial effusion	16.67	26.53	12.15
Coronary artery wall calc.	65.65	69.30	62.37
Hiatal hernia	17.72	27.72	13.02
Lymphadenopathy	23.16	42.00	15.99
Emphysema	40.18	36.15	45.21
Atelectasis	40.21	38.76	41.78
Lung nodule	46.57	50.96	42.88
Lung opacity	56.05	70.92	46.33
Pulmonary fibrotic sequela	21.44	45.31	14.04
Pleural effusion	74.85	81.70	69.06
Mosaic attenuation pattern	16.76	28.30	11.90
Peribronchial thickening	9.73	18.33	6.63
Consolidation	38.94	53.33	30.66
Bronchiectasis	17.42	22.55	14.20
Interlobular septal thickening	18.00	23.08	14.75
Macro Average	**36.45**	**46.41**	**31.68**

relatively low Macro F1 scores, reflecting the inherent difficulty of covering diverse and long-tailed pathology categories in chest X-ray report generation.

Table 15: Report generation performance on MIMIC-CXR.

Method	Sample F1	RaTEScore	Macro F1
Qwen3-VL-8B	26.89	48.97	22.00
InternVL-3.5-8B	31.05	51.14	26.81
Lingshu-7B	29.01	51.55	22.83
HuluMed-7B	34.63	50.11	<u>28.83</u>
MedGemma-1.5-4B	<u>36.19</u>	**56.56**	28.66
OmniCT-7B	22.59	47.47	19.08
RadSight-4B	35.95	52.13	28.81
RadSight-8B	**36.32**	<u>52.16</u>	**29.87**

F CASE STUDY

We present qualitative examples across medical image understanding tasks and examine how perceptual learning supports high-level clinical interpretation.

F.1 Qualitative Results Across Tasks

Figure 18 presents fine-grained attribute judgment examples, including lesion location and size estimation. In these cases, RadSight provides descriptions that more closely match the reference, while baseline predictions include inaccurate measurements or incomplete anatomical descriptions. Figure 19 illustrates 2D spatial grounding for pneumothorax and atelectasis in chest X-rays, with higher overlap between RadSight's predictions and the reference boxes in the displayed cases. Figures 20 and 21 extend these examples to organ and lesion localization in CT volumes. The lesion examples also reveal localization errors, consistent with the remaining challenges in precise lesion grounding (Table 10). Figure 22 illustrates spatial understanding, where RadSight identifies the finding, lesion, or organ associated with a specified bounding box. Figure 23 presents report generation examples for chest X-rays and CT scans, illustrating how visual findings are expressed in clinical language.

F.2 From Visual Perception to Clinical Interpretation

Figure 24 illustrates how perceptual learning supports high-level medical understanding through report generation. Without perceptual supervision, the displayed report focuses on normal structures and negative findings while omitting several clinically relevant abnormalities. With perceptual supervision, RadSight reports emphysematous changes, atelectasis with specific anatomical locations, millimetric pulmonary nodules, atheromatous plaques, and a sliding hiatal hernia, yielding closer agreement with the reference. This comparison suggests that learning to recognize fine-grained visual abnormalities provides a stronger basis for clinical interpretation. Together with the quantitative ablations, it highlights the importance of perceptual learning for translating visual evidence into more complete high-level clinical outputs.

Prompt Design for Organ-Level Disease Extraction

You are a medical report analysis assistant. I will provide you with a CT/X-Ray imaging report, and you are to process it according to the following steps. Do not provide any summaries or explanations.

Step 1: Decompose the report into organ-level sub-reports according to the following requirements:

1. Carefully read the entire CT report and identify all organs or anatomical structures mentioned.
2. Use only high-level, standardized organ names (e.g., "lung", "liver", "kidney", "heart", "spine"). Do NOT use laterality-specific or region-specific terms as separate organs. For example, "left lower lobe", "right upper lobe" should all be grouped under "lung"; "right arm", "left leg" should be grouped under their parent structure (e.g., "bone" or "soft tissue").
3. For each identified organ, extract all sentences or clauses that describe it, forming a dedicated sub-report for that organ.
4. Ensure that each organ-level sub-report is self-contained and complete, including all relevant descriptions such as location, size, shape, density, and any abnormalities mentioned for that organ.
5. If a sentence involves multiple organs, duplicate it into each relevant organ's sub-report.
6. Do not use pronouns; perform anaphora resolution to ensure each sentence is explicit and clear.
7. Output the result as a JSON object where each key is the standardized organ name and the value is a list of sentences describing that organ. Use `<step1>` and `</step1>` as tags.

Step 2: Based on the organ-level sub-reports from Step 1, identify all diseases. For each organ, adhering to the following requirements:

1. Use only high-level, standardized disease names (e.g., Pleural Effusion, Lung Opacity, Cardiomegaly, Pneumonia, Lung nodule). Merge synonymous or overly specific findings into a single standard term. For example: "left lower lobe opacity", "right lung opacity", "bilateral opacities" $\rightarrow$ "Lung Opacity". "small pleural effusion", "bilateral pleural effusion" $\rightarrow$ "Pleural Effusion". "PICC line", "A port catheter extending to the right atrium", "central venous catheter" $\rightarrow$ "Medical material". Do NOT create separate disease entries for left/right side or mild/severe variations.
2. For each identified disease or finding, record the following fields:

`Field`	**Description**
`organ`	The standardized organ name (from Step 1)
`disease`	The standardized disease or finding name
`evidence`	A list of sentences from Step 1 that support this diagnosis
`status`	One of: `present`, `absent`, `suspected`, or `historical`

3. If an organ shows no abnormalities, record it with an empty disease list.
4. Output as a JSON list of organ-disease objects. Use `<step2>` and `</step2>` as tags.

Figure 6: Prompt Design for Organ-Level Disease Extraction.

Prompt Template for Disease Attribute Extraction and QA Generation

CT Report: `<report>`
Disease/Finding: `<disease>`
Candidate Attributes: `<candidate_features>`
The CT report is already known to mention the disease/finding `<disease>`. Your task is to identify which attributes of this disease/finding are explicitly described in the report, extract their values, and then construct question-answer pairs based on the extracted information.

Please follow the 3 steps below and output each step on a new line only. Do not output any explanation, summary, or extra text.

General requirements:

1. Only use information explicitly stated in the CT report.
2. Do not infer, assume, or hallucinate any attribute or value.
3. Only select attributes from the provided Candidate Attributes.
4. If multiple findings are described, only extract attributes that correspond to the target disease/finding `<disease>`.
5. Keep attribute values concise and accurate.
6. Do not merge multiple attributes into one.
7. Each extracted attribute should correspond to one best value only.

Step 1: Identify which candidate attributes of `<disease>` are explicitly described in the report.
If none are described, output exactly: `<step1>[]</step1>`
Otherwise, output exactly a JSON list of attribute names, for example:
`<step1>["Location", "Severity", "Morphology"]</step1>`

Step 2: Extract the corresponding value for each attribute identified in Step 1.
If Step 1 is empty, output exactly: `<step2>[]</step2>`
Otherwise, output exactly a JSON list of objects in this format:
`<step2>[{"feature": "Location", "value": "left lower lobe"}, {"feature": "Severity", "value": "mild"}]</step2>`
Each feature should appear only once.
The value must be directly supported by the report.

Step 3: For each feature-value pair in Step 2, construct one question-answer pair.
If Step 2 is empty, output exactly: `<step3>[]</step3>`
Otherwise, output exactly a JSON list of objects in this format:
`<step3>[{"feature": "Location", "question": "Where is the atelectasis located?", "answer": "left lower lobe"}]</step3>`

Requirements for Step 3:

1. Each question must ask about one specific attribute of the target disease/finding.
2. The question must not contain the answer or its paraphrase.
3. The answer must exactly match the value extracted in Step 2.
4. Questions should be natural and medically meaningful.

Figure 7: Prompt template for disease attribute extraction and QA generation.

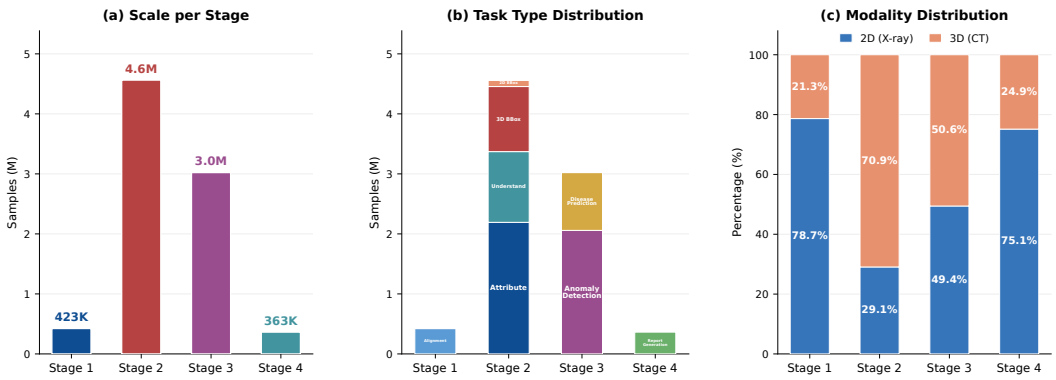


Figure 8: Overview of training data across four stages. **(a)** Total sample count per stage. **(b)** Task type distribution: Stage 1 focuses on vision-language alignment, Stage 2 on perception tasks (attribute QA, understanding, and 2D/3D bounding box localization), Stage 3 on clinical diagnosis (anomaly detection and disease prediction), and Stage 4 on medical report generation. **(c)** Modality distribution (2D X-ray vs. 3D CT) at each stage.

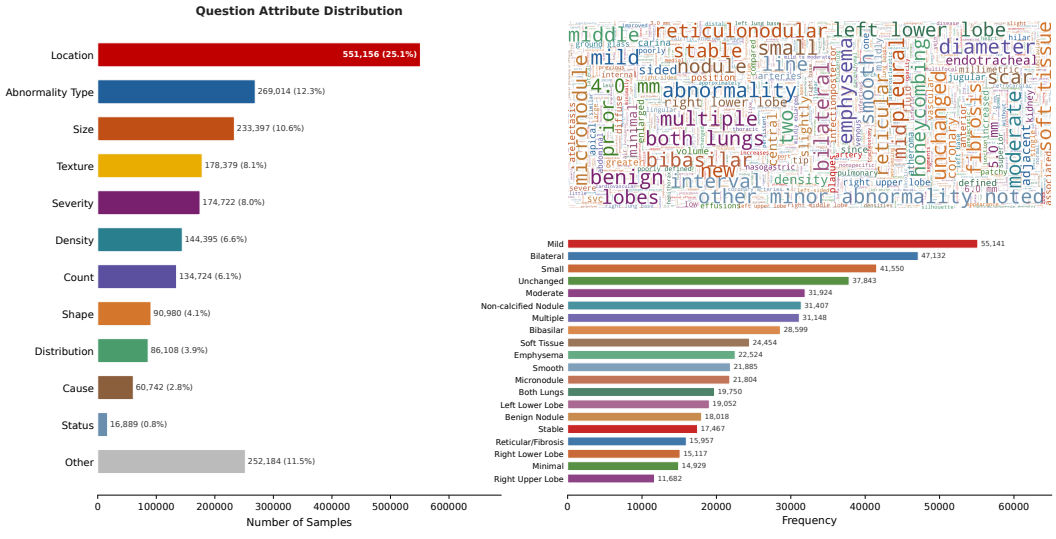


Figure 9: Stage 2 attribute judgement VQA data analysis. **Left:** Distribution of question attributes across 12 categories, with Location (25.1%) being the most frequent. **Right top:** Word cloud of medical entities extracted from QA pairs. **Right bottom:** Top-20 most frequent medical entities and their occurrence counts.

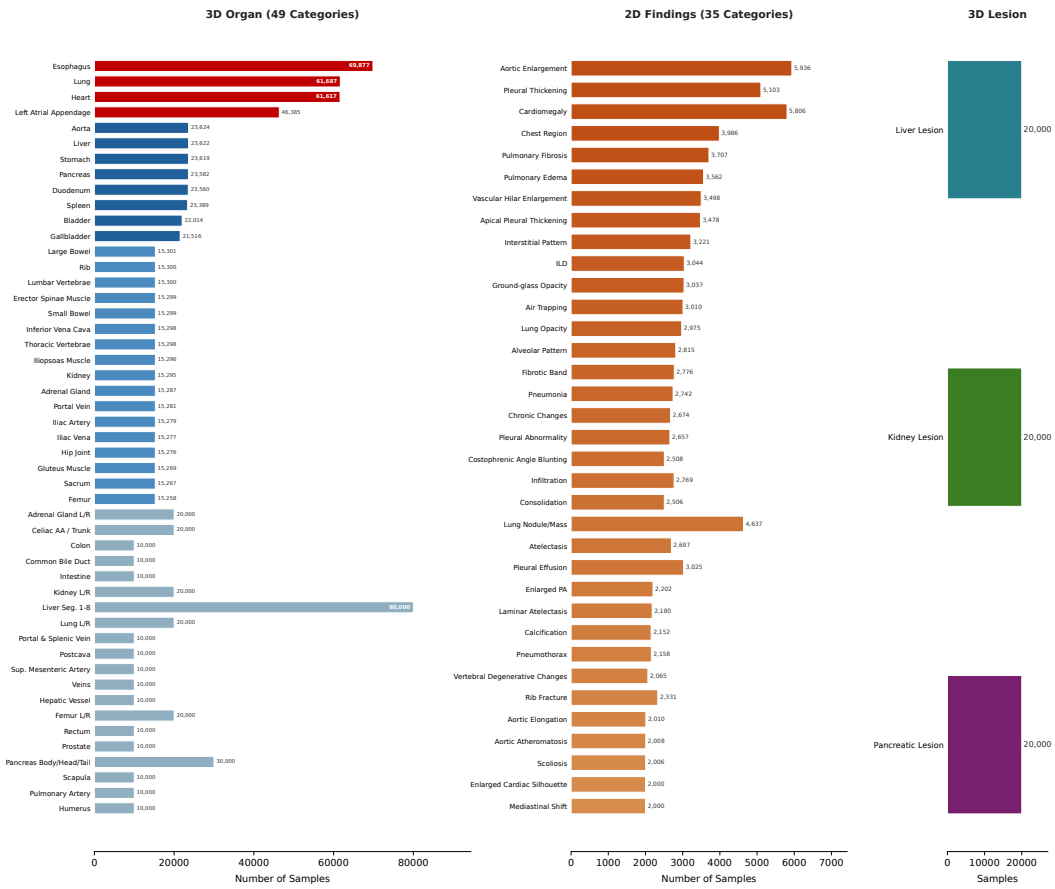


Figure 10: Distribution of Stage 2 spatial grounding training data. **Left:** 3D organ categories (49 types) extracted from CT volumes, covering major anatomical structures across the thorax and abdomen. **Middle:** 2D radiographic findings (35 types) from chest X-rays, including common abnormalities such as cardiomegaly, pleural effusion, and consolidation. **Right:** 3D lesion categories (3 types) with 20K samples each for liver, kidney, and pancreatic lesions.

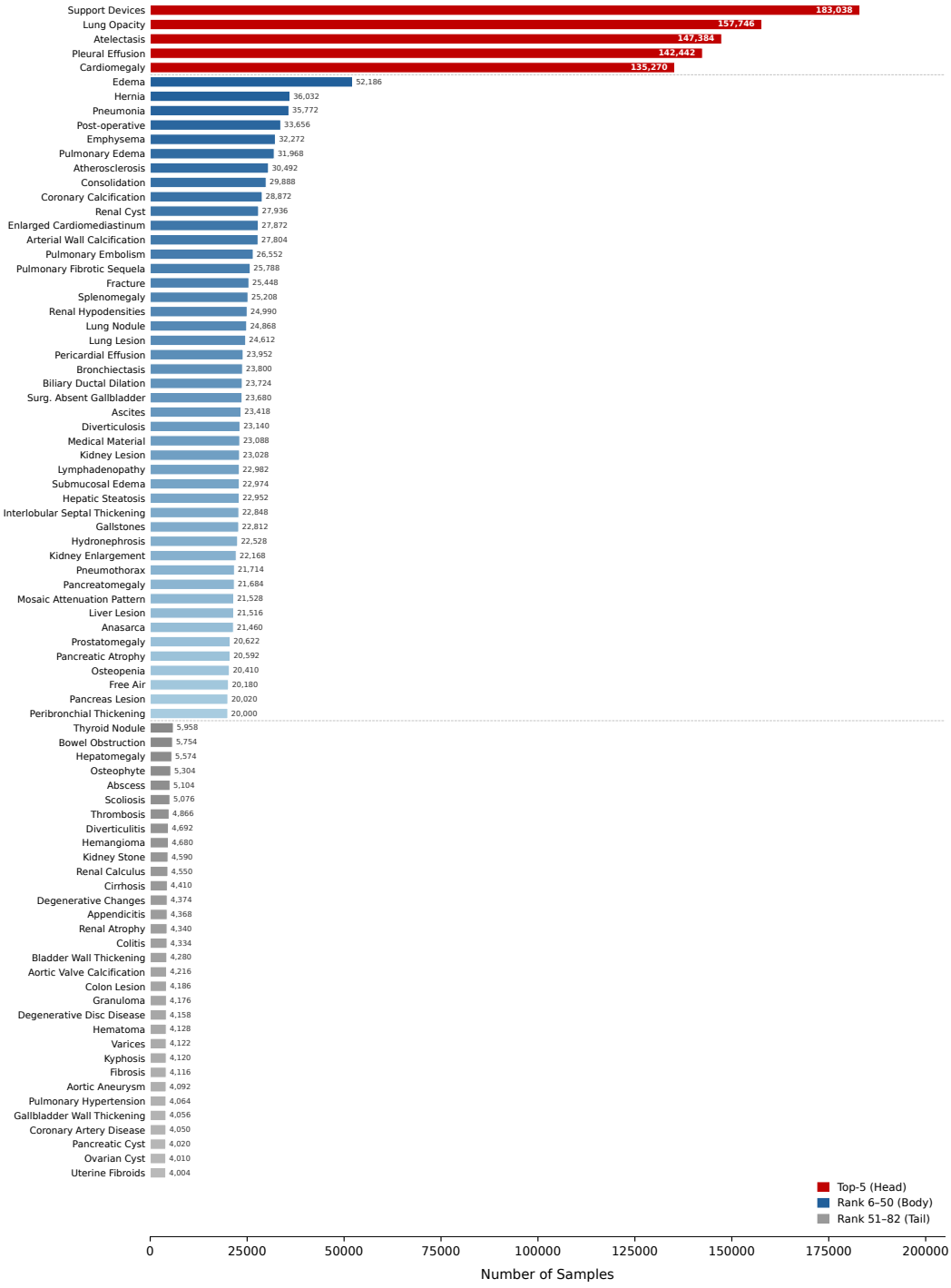


Figure 11: Disease distribution of the Stage 3 anomaly detection training data across 82 categories. The top-5 categories (red) are predominantly chest X-ray findings and account for over 37% of all samples. The body (blue, Rank 6–50) covers a broad spectrum of thoracic and abdominal pathologies, while the tail (gray, Rank 51–82) comprises 32 less common categories.

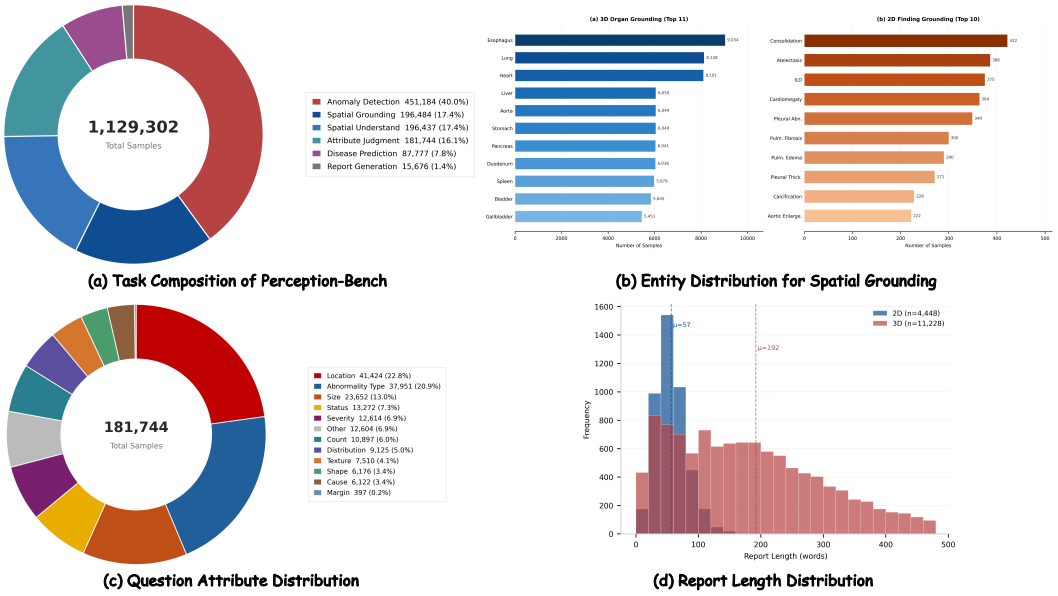


Figure 12: **Composition of Perception-Bench.** **(a)** Distribution of the 1,129,302 evaluation instances across six tasks. **(b)** Most frequent 3D anatomical structures and 2D radiographic findings in the spatial-grounding subset. **(c)** Distribution of the 181,744 attribute-judgment instances across 12 attribute types. **(d)** Report-length distributions for the 2D and 3D report-generation subsets; dashed lines indicate the corresponding mean lengths.

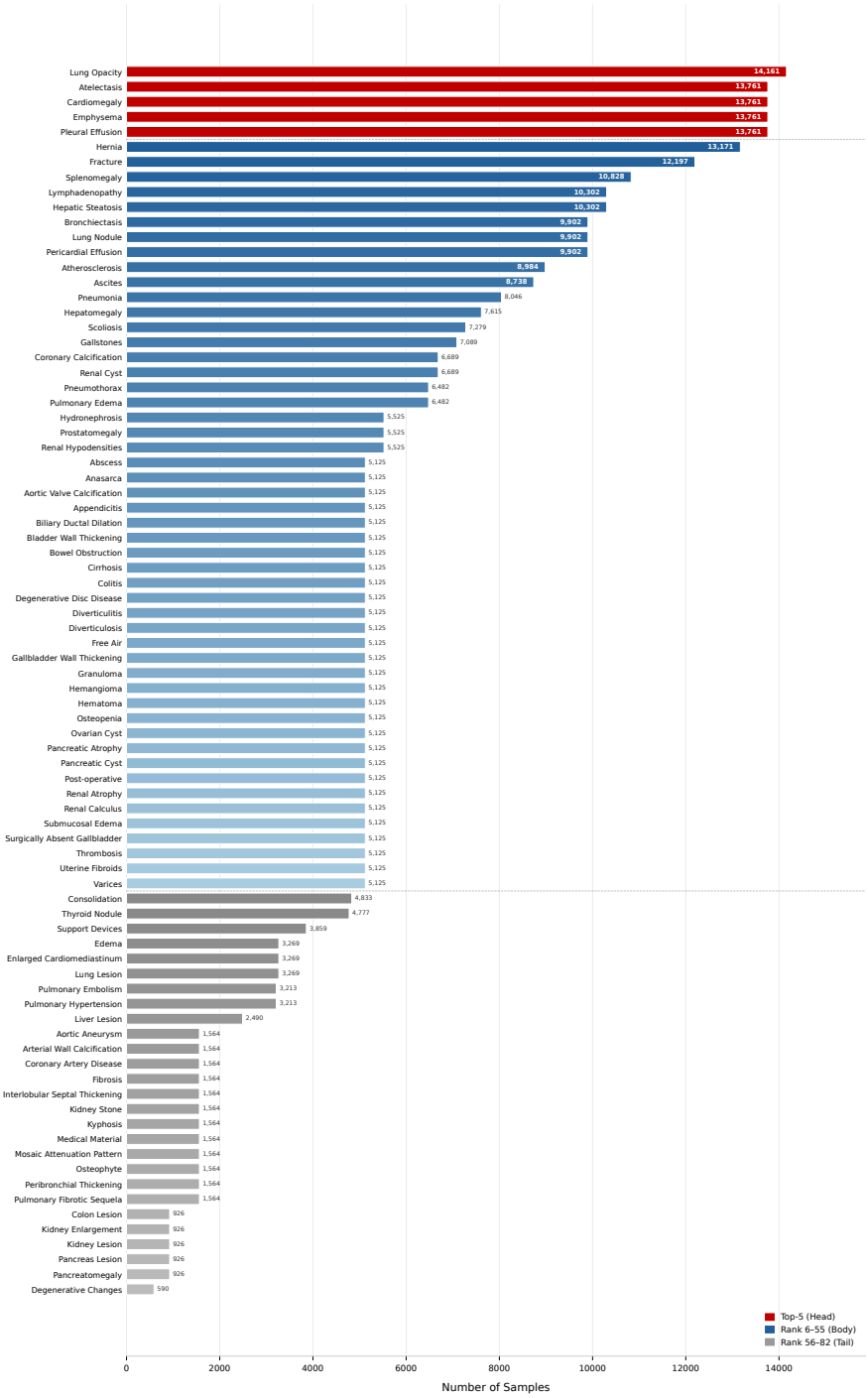


Figure 13: **Category distribution of the Perception-Bench anomaly-detection subset.** The subset contains 451,184 instances across 82 abnormality and imaging-finding categories. Categories are ordered by sample count and divided into the five most frequent categories (red), body categories ranked 6–55 (blue), and tail categories ranked 56–82 (gray).

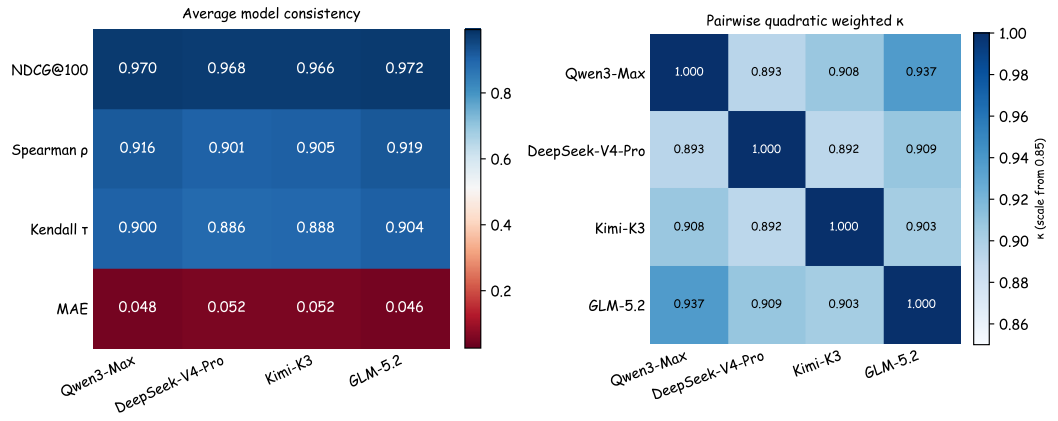


Figure 14: Cross-judge agreement on 200 RadSight attribute-judgment predictions using the shared rubric in Section D.2. Left: each judge's mean pairwise NDCG@100, Spearman ρ, Kendall τ, and MAE against the other three judges; higher indicates stronger agreement, except for MAE. Right: pairwise quadratic weighted κ for scores in $\{0, 0.5, 1\}$. Qwen3-Max exhibits agreement comparable to the alternative judges, supporting its use as the primary evaluator.

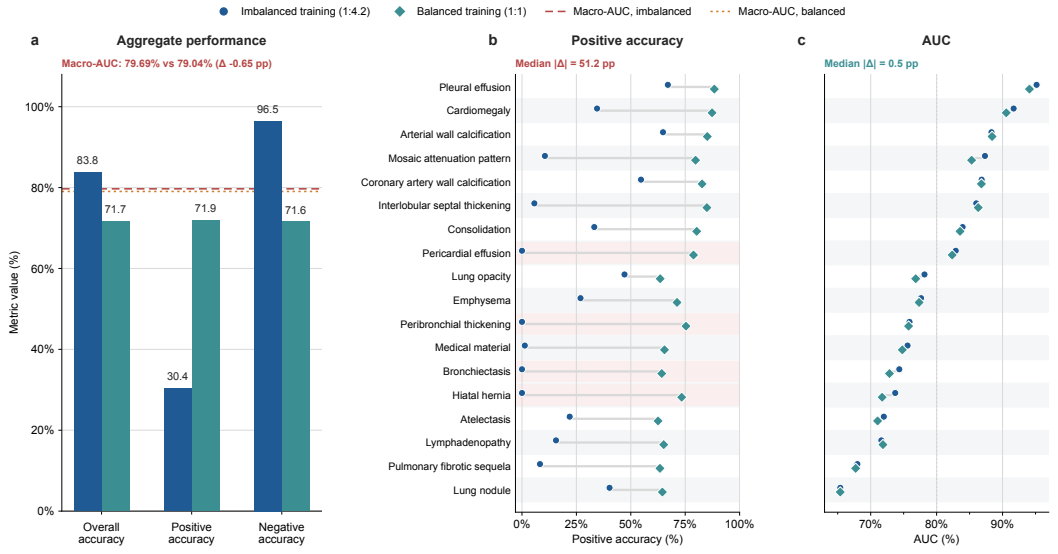


Figure 15: **AUC provides a more reliable evaluation of anomaly-detection ability under class imbalance. a**, Aggregate performance under imbalanced (1:4.2) and balanced (1:1) training distributions. Although balancing substantially improves positive accuracy, the overall ACC decreases because fewer samples belong to the majority negative class. The two horizontal lines indicate the corresponding macro-average AUC values. **b**, Disease-level positive accuracy changes markedly after class balancing, with a median absolute difference of 51.2 percentage points (pp). Light-red rows denote diseases for which the model produces no positive predictions under imbalanced training. **c**, Disease-level AUC remains comparatively stable across the two training distributions, with a median absolute difference of only 0.5 pp. Circles and diamonds represent models trained with the imbalanced and balanced distributions, respectively.

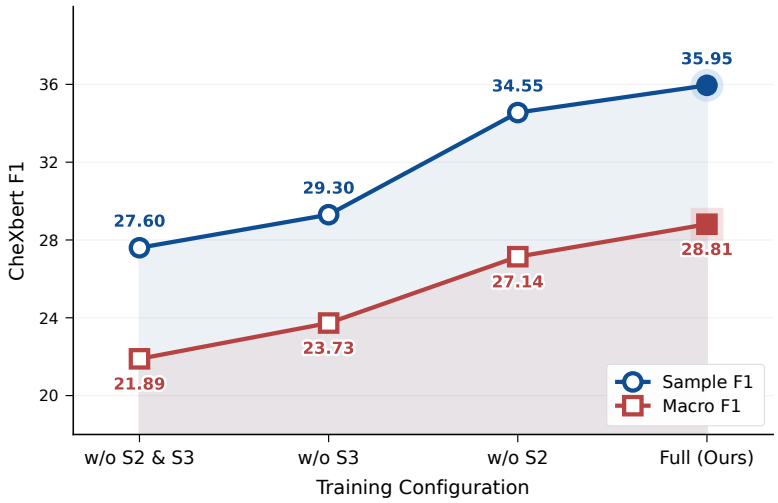


Figure 16: Ablation study on the impact of intermediate training stages on 2D chest X-ray report generation performance (MIMIC-CXR). CheXbert F1 metrics are reported.

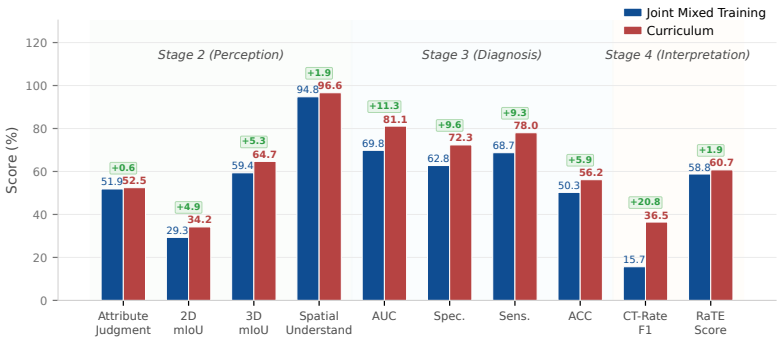


Figure 17: Ablation study on the impact of curriculum learning. Performance comparison across ten metrics spanning perception, diagnosis, and interpretation. Curriculum learning consistently outperforms joint mixed training.

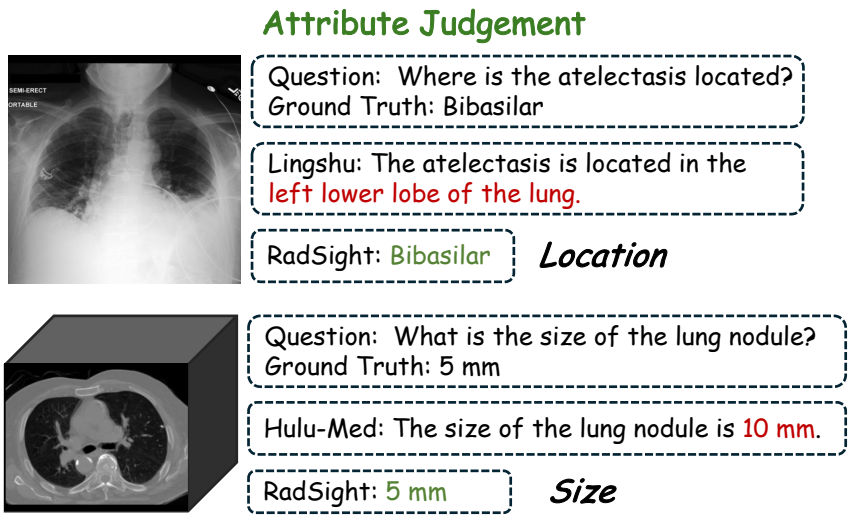


Figure 18: Case studies on fine-grained attribute perception, including lesion location and size estimation.

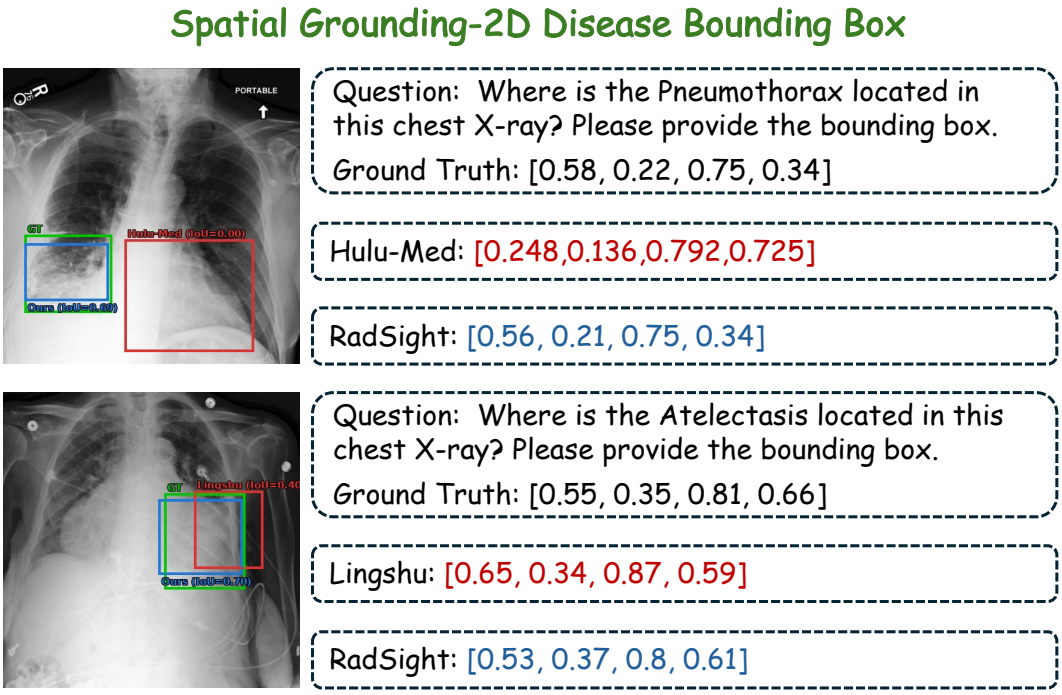
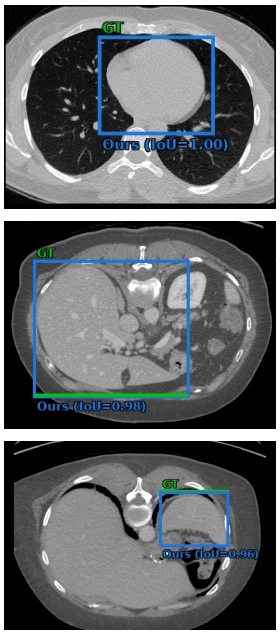


Figure 19: Case studies on 2D spatial grounding for disease localization in chest X-rays.

Spatial Grounding-3D Organ Bounding Box



Question: Where is the heart located in this CT scan? Please provide the 3D bounding box coordinates.
Ground Truth: [0.39, 0.22, 0.34, 0.7, 0.61, 0.58]

RadSight: [0.39, 0.22, 0.34, 0.7, 0.61, 0.58]

Question: Where is the liver located in this CT scan? Please provide the 3D bounding box coordinates.
Ground Truth: [0.17, 0.2, 0.57, 0.66, 0.83, 0.9]

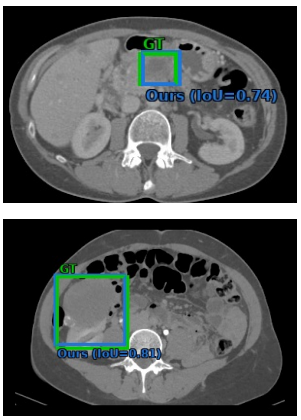
RadSight: [0.17, 0.2, 0.57, 0.66, 0.84, 0.9]

Question: Where is the spleen lesion located in this CT scan? Please provide the 3D bounding box coordinates.
Ground Truth: [0.58, 0.27, 0.75, 0.82, 0.55, 0.88]

RadSight: [0.58, 0.28, 0.75, 0.82, 0.55, 0.88]

Figure 20: Case studies on 3D spatial grounding for organ localization in CT scans.

Spatial Grounding-3D Lesion Bounding Box



Question: Where is the pancreatic lesion located in this CT scan? Please provide the 3D bounding box coordinates.
Ground Truth: [0.5, 0.3, 0.52, 0.59, 0.41, 0.56]

RadSight: [0.51, 0.29, 0.52, 0.6, 0.41, 0.56]

Question: Where is the kidney lesion located in this CT scan? Please provide the 3D bounding box coordinates.
Ground Truth: [0.19, 0.3, 0.49, 0.42, 0.65, 0.7]

RadSight: [0.18, 0.29, 0.49, 0.41, 0.63, 0.71]

Figure 21: Case studies on 3D spatial grounding for lesion localization in CT scans.

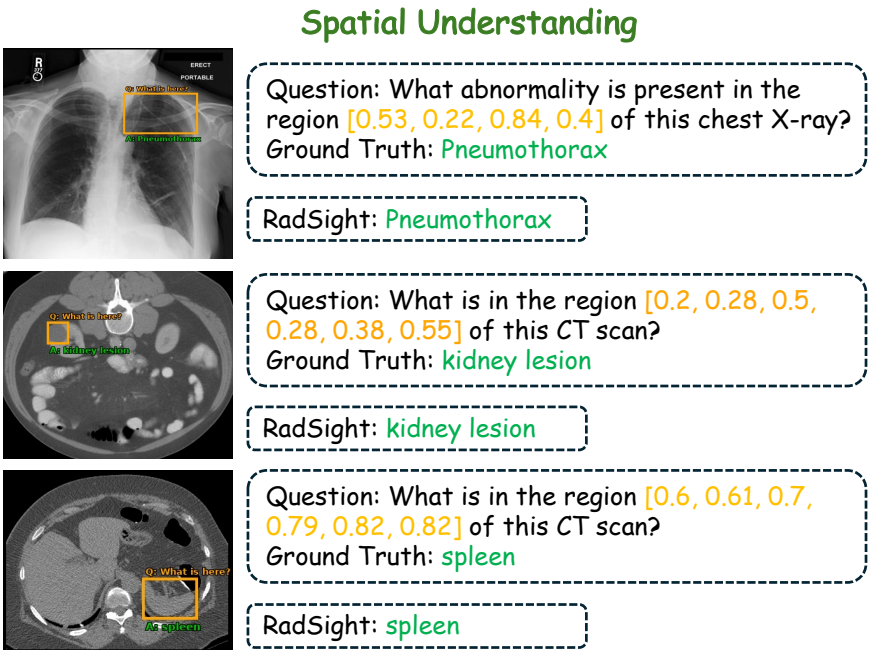


Figure 22: Case studies on spatial understanding.

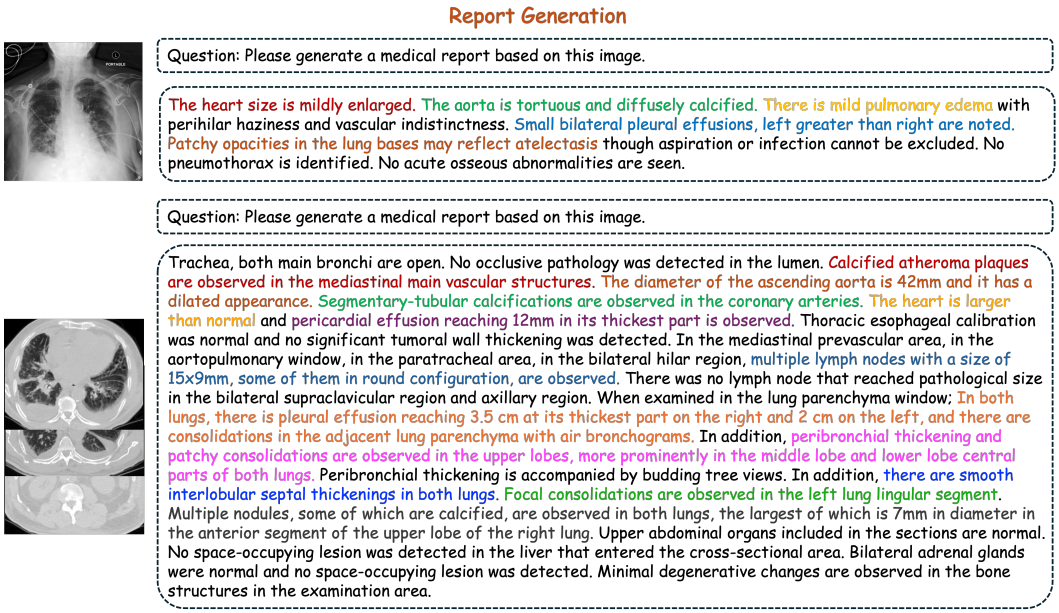
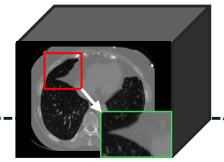


Figure 23: Case studies on report generation for X-Ray and CT scans.

(c) Case Study on Perception-Induced Failures



Q: Please generate a medical report based on this image.
Ground Truth: Trachea and both main bronchi are open... There are pleuroparenchymal sequelae changes in both lung apex. Atelectasis were observed in the middle lobe of the right lung and the lingular segment of the left lung upper lobe. Minimal emphysematous changes were observed in both lungs. There are several millimetric nonspecific nodules in both lungs. There is no mass or infiltrative lesion in both lungs...There are atheromatous plaques in the aorta and coronary arteries...There is a sliding type hiatal hernia at the lower end of the esophagus...No pathological increase in wall thickness was detected...

w/o perception: Trachea, both main bronchi are open. Mediastinal main vascular structures, heart contour,.....,Thoracic esophagus calibration was normal and no significant tumoral wall thickening was detected. No enlarged lymph nodes in prevascular, pre-paratracheal, subcarinal or bilateral hilar-axillary pathological dimensions were detected...Pleural effusion-thickening was not detected. Upper abdominal organs included in the sections are normal. No space-occupying lesion was detected in the liver that entered the cross-sectional area. Bilateral adrenal glands were normal and no space-occupying lesion was detected...

w/ perception: Trachea and both main bronchi are open... There are minimal emphysematous changes in both lungs. Linear atelectasis was observed in the right lung middle lobe and left lung upper lobe lingular segment. There are millimetric nonspecific nodules in both lungs. No mass or infiltrative lesion was detected in both lungs...There are atheromatous plaques in the aorta and coronary arteries. No pathologically enlarged lymph nodes were detected in the mediastinum and hilar regions. No pathological wall thickness increase was observed in the esophagus within the sections. There is a sliding type hiatal hernia at the lower end of the esophagus. No upper abdominal free fluid-collection was detected in the sections.

Figure 24: Perceptual learning supports high-level clinical interpretation. For the same chest CT study, the report excerpt generated with perceptual supervision covers more reference abnormalities and their anatomical locations. Matching colors indicate corresponding findings across report excerpts.