MedVision: Benchmarking Quantitative Medical Image Analysis

Yongcheng Yao[♡] **Yongshuo Zong**[♡] **Raman Dutt**[♡]
Yongxin Yang[♠] **Sotirios A Tsaftaris**[♢] **Timothy Hospedales**[♡*]
[♡]School of Informatics, University of Edinburgh
[♠]King's College London [♢]School of Engineering, University of Edinburgh
{yc.yao, yongshuo.zong, raman.dutt, s.tsaftaris, t.hospedales}@ed.ac.uk, yongxin.yang@kcl.ac.uk

https://medvision-vlm.github.io

Abstract

Current vision-language models (VLMs) in medicine are primarily designed for categorical question answering (e.g., "Is this normal or abnormal?") or qualitative descriptive tasks. However, clinical decision-making often relies on quantitative assessments, such as measuring the size of a tumor or the angle of a joint, from which clinicians draw their own diagnostic conclusions. This quantitative reasoning capability remains underexplored and poorly supported in existing VLMs. In this work, we introduce **MedVision**, a large-scale dataset and benchmark specifically designed to evaluate and improve VLMs on quantitative medical image analysis. MedVision spans 22 public datasets covering diverse anatomies and modalities, with **29.0K** 3D images, **11.2M** annotated 2D slices, and **24.3M** single-instance annotations. We focus on three representative quantitative tasks: (1) detection of anatomical structures and abnormalities, (2) tumor/lesion (T/L) size estimation, and (3) angle/distance (A/D) measurement. We show that current off-the-shelf VLMs perform poorly on these tasks. However, supervised and reinforcement fine-tuning (RFT) on MedVision significantly enhances performance across detection, T/L size estimation, and A/D measurement, yielding **MedVision-V0** as a strong open baseline. In the RFT stage, we design and evaluate the efficacy of process rewards, multiplicative reward composition, and multi-task RFT with curriculum learning. MedVision provides a foundation for developing VLMs with robust quantitative reasoning capabilities in medical imaging.

1 Introduction

Recent advances in vision-language models (VLMs) (Li et al., 2024a; Liu et al., 2023; Bai et al., 2023; Zong et al., 2025) have made it possible to pair powerful visual encoders with flexible natural language reasoning, enabling a range of clinical applications such as image-level classification and interactive question answering. In practice, however, most medical VLM usage and evaluation have concentrated on categorical (Lau et al., 2018; Liu et al., 2021; Ben Abacha et al., 2019) or qualitative outputs (Subramanian et al., 2020): distinguishing "normal vs. abnormal" images, labeling a finding, or producing free-text descriptions of image content. These tasks are important, but they do not capture a critical part of routine clinical reasoning: the extraction of precise, *quantitative* measurements from images that clinicians use to stage disease, plan interventions, and monitor therapy.

Quantitative assessment is central to many medical decisions. Examples include detecting metastases and estimating tumor/lesion size for cancer staging (Eisenhauer et al., 2009; Detterbeck et al., 2024), and measuring the Cobb angle for scoliosis diagnosis (Cheng et al., 2015). Despite this, the ability of modern VLMs to produce reliable, reproducible quantitative measurements (e.g., lengths, areas, volumes, angles) has received little systematic attention. Existing datasets (Lau et al., 2018; Ben Abacha et al., 2019, 2021; Liu et al., 2021; He et al., 2020a; Irvin et al., 2019; Chambon et al., 2024; Johnson et al., 2019; Kayser et al., 2022; Boecking et al., 2022; Lin et al., 2023) emphasize classification accuracy or natural language quality rather than measurement precision. Recently, RadGPT (Bassi et al., 2025) presented a dataset of radiology reports with tumor location and size information, but it is limited to abdominal CT scans. Related benchmarks (Chen et al., 2024b; Hu et al., 2024; Wu et al., 2025; Zhou et al., 2025) similarly focus on qualitative tasks such as disease classification, visual question answering, and report generation. A recent work on reinforcement fine-tuning, MedVLM-R1 (Pan et al., 2026), improves medical VLM accuracy on classification tasks, but leaves quantitative measurement largely unaddressed. A detailed comparison of existing datasets and bench-

*Corresponding author.

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 31065–31102
October 24-29, 2026 ©2026 Association for Computational Linguistics

marks is provided in Appendix A.

Quantitative vision–language capability is challenging and underexplored for several reasons: measurement tasks demand precise localization across heterogeneous modalities and an understanding of geometric relationships and physical units rather than semantic categories alone; current architectures and training objectives rarely target calibrated numeric outputs; and public datasets with standardized measurement annotations remain sparse and fragmented across modalities and anatomies.

To address these gaps, we introduce **MedVision**, a large-scale dataset and benchmark specifically created to assess and improve VLM performance on quantitative medical-image tasks. MedVision aggregates 22 public datasets spanning multiple anatomies and imaging modalities, comprising **29.0K** 3D images, **11.2M** annotated 2D slices, and **24.3M** single-instance annotations suitable for measurement tasks.[1] We organize the benchmark around three representative, clinically relevant tasks: (1) detection of anatomical structures and abnormalities (localization and identification), (2) tumor/lesion (T/L) size estimation (bidirectional dimensions), and (3) angle/distance (A/D) measurement (e.g., joint angles, inter-structure distances).

Using MedVision, we systematically evaluate open-weight VLMs in two settings: off-the-shelf inference and fine-tuning on the MedVision training splits. Off-the-shelf VLMs, even those with strong qualitative language abilities, perform poorly on quantitative tasks, failing at precise localization and producing large numeric errors, whereas supervised fine-tuning and reinforcement fine-tuning (SFT+RFT) on MedVision materially improve detection and measurement accuracy. We also analyze common failure modes and identify persistent challenges. Our **main contributions** are as follows:

(i) We highlight a clinically important but underappreciated gap: modern VLMs are not reliably able to produce precise quantitative measurements from medical images.

(ii) We introduce **MedVision**, a large-scale, multi-modality dataset for quantitative medical image analysis.

(iii) We provide the first comprehensive evaluation of contemporary VLMs on detection, tumor/lesion size estimation, and angle/distance measurement.

(iv) We train **MedVision-V0** on our dataset via supervised and reinforcement learning, which significantly outperforms all evaluated VLMs across all three quantitative tasks and generalizes to out-of-distribution imaging planes and anatomical targets.

(v) We design and evaluate the efficacy of process rewards, multiplicative reward composition in RFT, and multi-task RFT with curriculum learning.

(vi) We release the data, model checkpoints, and code to enable further research toward VLMs that support clinically useful quantitative reasoning.

2 MedVision: Dataset and Benchmark

An overview of our MedVision dataset and benchmark design is illustrated in Figure 1. We curated a large-scale multi-anatomy and multi-modality medical image dataset with quantitative annotations (Section 2.1). We designed a benchmark to evaluate VLMs' ability on three quantitative medical image analysis tasks (Section 2.2).

2.1 Dataset

We constructed the MedVision dataset from public medical image data. Quantitative annotations are derived from the original labels. The dataset construction process is automated and reproducible, with a codebase that handles raw data downloading, image preprocessing, and annotation extraction.

Data Source & Preprocessing. Public medical imaging data with segmentation masks or landmark annotations were collected. We restricted the medical image modalities to those with physical spacing information in the image file header, which is essential for generating ground truth quantitative measurements. We collected 22 datasets, summarized in Table 1, covering a diverse range of anatomies and modalities: AbdomenAtlas (Li et al., 2024b), AbdomenCT-1K (Ma et al., 2022), ACDC (Bernard et al., 2018), AMOS22 (Ji et al., 2022), autoPET-III[2], BCV15[3], BraTS24 (de Verdier et al., 2024; LaBella et al., 2025; Kazerooni et al., 2024), CAMUS (Leclerc et al., 2019), Ceph-Bio-400 (Lindner et al., 2016), CrossMoDA (Dorent et al., 2023), FeTA24 (Payette et al., 2023), FLARE22 (Ma et al., 2024), HNTSMRG24[4], ISLES24[5], KiPA22 (He et al., 2021, 2020b; Shao et al., 2011, 2012),

[1] Dataset statistics for v1.0.0; the latest release is available at https://medvision-vlm.github.io/explorer.

[2] https://autopet-iii.grand-challenge.org

[3] https://doi.org/10.7303/syn3193805

[4] https://hntsmrg24.grand-challenge.org

[5] https://isles-24.grand-challenge.org

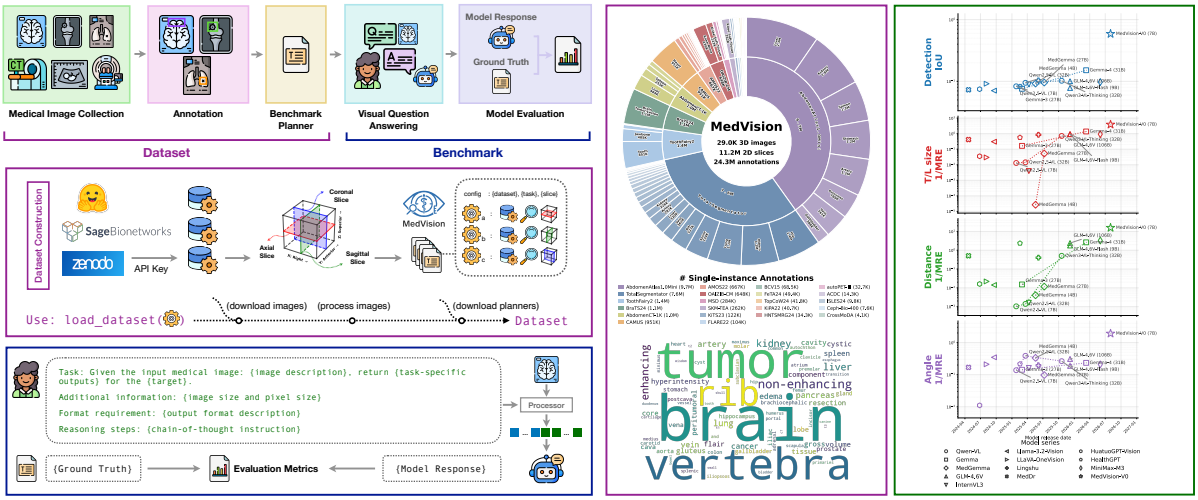


Figure 1: Overview of MedVision dataset construction and benchmark design (left), dataset summary (middle), and benchmark performance against model release date (right, details in Appendix N). The MedVision dataset was constructed for out-of-the-box use: `load_dataset` is all you need. Appendix B shows examples of imaging data and quantitative annotation.

KiTS23 (Heller et al., 2023), MSD (Antonelli et al., 2022), OAIZIB-CM (Ambellan et al., 2019; Yao et al., 2024; Yao and Chen, 2025), SKM-TEA (Desai et al., 2021), ToothFairy2 (Bolelli et al., 2025b,a; Lumetti et al., 2024), TopCoW24 (Yang et al., 2026), and TotalSegmentator (Wasserthal et al., 2023). Images are preprocessed to ensure consistent orientation and standardized format. Images are stored as 3D volumes with orientation corrected to RAS+ convention, where the first, second, and third dimensions correspond to the left-to-Right, posterior-to-Anterior, and inferior-to-Superior directions, respectively. Our data loading codebase supports flexible image slicing along any of the three anatomical planes (sagittal, coronal, and axial).

Quantitative Annotations. We generated bounding box (Box), bidirectional tumor/lesion (T/L) size, and angle/distance (A/D) annotations to support the three benchmark tasks. The new annotations and metadata of images are released as the benchmark planner. ***(i) Bounding Box Annotation:*** For each segmentation label in a 2D slice, a bounding box is fitted to each cluster of the binary mask. Clusters with fewer than 10 pixels in any dimension are excluded. ***(ii) Tumor/Lesion Size Annotation:*** Bidirectional measurements of tumors/lesions from medical images are useful metrics and can provide quantitative information for disease assessment and monitoring. We generated T/L size annotations by calculating the lengths of the major and minor axes of the ellipse fitted to each T/L label in 2D slices (Figure 2a). Ellipse fitting is conducted in the real-world coordinate system by multiplying the image array indices with pixel sizes. ***(iii) Angle/Distance Annotation:*** Angle and distance measurements are calculated from human-annotated landmarks. Figure 2b shows examples of landmarks in the two datasets. All target measurements are in ***real-world units*** (e.g., mm), and models must convert image-space estimates into real-world units using the pixel/image size provided in the prompt.

2.2 Benchmark

Model. We evaluated a range of general and medical open-weight VLMs. General VLMs include Qwen2.5-VL (Bai et al., 2025), Qwen3-VL (Qwen Team, 2025), InternVL3 (Zhu et al., 2025), Gemma3 (Gemma Team et al., 2025), Gemma4 (Gemma Team et al., 2026), GLM-4.6V (GLM-V Team et al., 2026), MiniMax-M3 (Lai et al., 2026), Llama3.2-Vision, and LLaVA-OneVision (Li et al., 2024a). Medical VLMs include Lingshu (LASA Team et al., 2025), MedGemma (Sellergren et al., 2026), MedDr (He et al., 2026), HuatuoGPT-Vision (Chen et al., 2024a), and HealthGPT-L14 (Lin et al., 2025). Details of these models are provided in Appendix D.

Prompt. We focus on clinically relevant open-ended VQA tasks for our benchmark: (1) detection of healthy anatomical structures and abnormalities, (2) tumor/lesion size estimation, and (3) angle/distance measurement. Models are instructed to out-

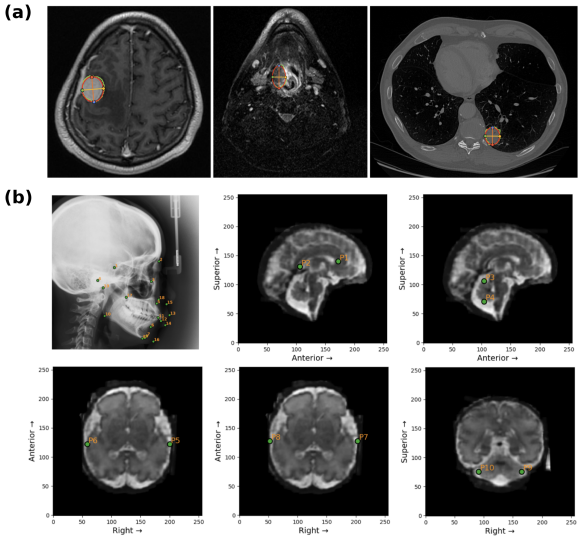


Figure 2: Quantitative annotation examples. (a) Tumor/lesion size annotation. An ellipse is fitted to the tumor/lesion mask. The physical lengths of the major and minor axes are recorded as T/L size annotation. (b) Landmarks in the Ceph-Bio-400 (top-left) and FeTA24 datasets. Ground truth angle and distance measurements are calculated from these landmarks.

Table 1: Details of the MedVision dataset, with single-instance annotation counts at annotation version v1.0.0. MRI: Magnetic Resonance Imaging; CT: Computed Tomography; PET: Positron Emission Tomography; US: Ultrasound; Box: bounding box; T/L: tumor/lesion size; A/D: angle/distance. Full statistics, including multi-instance counts, in Table 10 (Appendix C).

Dataset	Anatomy	Modality	# Single-instance Annotation		
			Box	T/L	A/D
AbdomenAtlas	abdomen	CT	9,748,290	0	0
AbdomenCT-1K	abdomen	CT	1,041,588	0	0
ACDC	heart	MRI	14,271	0	0
AMOS22	abdomen	CT, MRI	666,532	0	0
autoPET-III	whole body	CT, PET	31,794	735	0
BCV15	abdomen	CT	68,543	0	0
BraTS24	brain	MRI	1,115,524	11,071	0
CAMUS	heart	US	951,370	0	0
Ceph-Bio-400	head and neck	X-ray	0	0	7,600
CrossMoDA	brain	MRI	4,076	0	0
FeTA24	fetal brain	MRI	49,087	0	325
FLARE22	abdomen	CT	104,211	0	0
HNTSMRG24	head and neck	MRI	32,029	1,392	0
ISLES24	brain	MRI	9,774	0	0
KiPA22	kidney	CT	37,647	3,095	0
KiTS23	kidney	CT	114,491	8,484	0
MSD	multiple	CT, MRI	277,451	7,472	0
OAIZIB-CM	knee	MRI	648,048	0	0
SKM-TEA	knee	MRI	262,338	0	0
ToothFairy2	tooth	CT	1,413,979	0	0
TopCoW24	brain	CT, MRI	41,829	0	0
TotalSegmentator	multiple	CT, MRI	7,603,455	0	0

put the required estimates in a structured format. Generally, for all tasks, the input prompt includes a task description and a format requirement. Specifically, for T/L and A/D tasks, physical spacing information and chain-of-thought (CoT) instructions are added to provide necessary context and guide the model's reasoning process. ***(i) Physical Spacing Information:*** A key feature of the benchmark is the careful handling of physical spacing information in the text prompt. Since each VLM adopts a different image processing strategy (see Appendix E), we design a dedicated prompt-building pipeline to ensure the physical spacing values provided in the prompt match the actual image input, and the image size and pixel size together always reflect the physical dimensions of the target region. ***(ii) CoT Instruction:*** The CoT instruction for T/L and A/D tasks provides step-by-step guidance for specific measurement tasks. In general, it instructs the model to first estimate landmark coordinates, then calculate the target distance or angle based on the predicted coordinates and the provided physical spacing information. The full CoT instructions are provided in Appendix I.

Metric. Model responses are parsed to extract the required numerical values. Parsing success is determined by whether the required number of numerical values can be extracted from the response: a rule-based parser reads the `<answer></answer>` tags, and an ***LLM-assisted parsing*** pipeline additionally recovers answers stated outside this format, without altering any rule-parsed value (Appendix F). All reported metrics are computed from the LLM-parsed responses. Success rate (SR) is calculated as the proportion of successfully parsed outputs among all samples. Other metrics are computed over successfully parsed outputs only, except the threshold metrics `<metric>`$_{<k}$ and `<metric>`$_{>k}$, whose denominator is the total number of samples so that they reflect both instruction following and measurement accuracy. For detection tasks, we report the region-based recall, precision, F1 score, intersection over union (IoU), $\text{IoU}_{>0.5}$, and $\text{IoU}_{0.5:0.05:0.95}$. For T/L size and A/D measurement tasks, we report the mean absolute error (MAE), mean relative error (MRE), and $\text{MRE}_{<0.1}$. Metric definitions are provided in Appendix G.

Experimental Setting. Details on computational resources, software environment, and code release are provided in Appendix H.

3 MedVision-V0 Training

We established the MedVision benchmark to evaluate VLMs' capabilities on quantitative medical image analysis tasks. Our results reveal that pretrained models struggle with medical image quan-

tification. This performance gap motivates the need for post-training on our dataset. We aim to train a model capable of anatomical structure recognition, precise localization, tumor/lesion size estimation, and measurement via internal perception and reasoning – an end-to-end VLM that does not rely on external tools or specialist models. To achieve this, we adopted a two-stage fine-tuning strategy, which includes supervised fine-tuning (SFT) and reinforcement fine-tuning (RFT). The SFT stage focuses on teaching the model to learn the answer formats and reasoning patterns, while the RFT stage further refines the model's outputs based on task-specific reward functions that encourage accurate intermediate estimates (e.g., landmark coordinates) and final measurements. We denote our model as **MedVision-V0**.

SFT with CoT. At this stage, we curated a multi-task CoT dataset with 121K samples from the MedVision training set. We randomly select 110K detection, 5.5K T/L size estimation, and 5.5K A/D measurement cases from a subset of axial slices. We deliberately leave out coronal and sagittal slices to evaluate the model's generalization ability to out-of-distribution (OOD) data. A weighted random sampler is used during training to allow oversampling of minority-task cases to mitigate the data imbalance issue (Appendix K). Each sample includes an image reshaped to 512×512 and a prompt-answer pair. The prompt follows the same format as the one used in the benchmark. The answer is structured into an internal reasoning part wrapped in `<think></think>` tags and a final answer in `<answer></answer>` tags. The reasoning text is generated by filling in intermediate ground truth values (e.g., landmark coordinates) into the task-specific CoT instruction template. The prompt structure and the CoT instructions and answer templates of all tasks are provided in Appendix I.

RFT via GRPO. For RFT, we used the GRPO algorithm (Shao et al., 2024) implemented in verl (Sheng et al., 2025). We constructed an RFT dataset for each of the three tasks and used them sequentially (A/D $\rightarrow$ T/L $\rightarrow$ detection) for multi-step fine-tuning. An alternative ***multi-task RFT*** approach is discussed in Section 4.3. The same 121K training samples from the SFT stage are used; the image size and prompt remain the same, while the CoT answer is removed in RFT data as models learn from the reward signal. In addition to the standard format and answer rewards in GRPO, we designed ***process rewards*** for T/L and A/D tasks to encourage accurate intermediate estimates. Both the process and answer rewards are calculated from $\exp(-x)$ where x is the error of the model prediction. The final reward is calculated as follows: $r = r_{\text{format}} + r_{\text{process}} \cdot r_{\text{answer}}$. We ablate this ***multiplicative reward composition*** against an additive alternative (Table 17 in Appendix M). Appendix J details the reward computation and the training configuration.

4 Results & Analysis

We first provide an overview of our fine-tuned model, MedVision-V0, and then quantitatively compare the performance of all evaluated models on the three benchmark task families.

4.1 MedVision-V0 Overview

Qualitative Overview. Figure 3 shows example responses from MedVision-V0 on three tasks, each paired with the predicted target and the ground truth overlaid on the input image. The model is able to correctly identify the target region and produce reasonable estimates. The model's internal reasoning process, in the `<think></think>` tags, shows that it first estimates the *relative coordinates* of the target points, then computes the requested measurement from these coordinates together with the provided pixel size and image size. More examples of model responses in the detection, T/L and A/D tasks are provided in Appendix M.

Error Analysis. To understand the challenge posed by MedVision, we analyze the end-to-end measurement error against the error of sub-tasks along the reasoning path: i) landmark localization, and ii) arithmetic mapping of landmark locations to final answers. We quantify the ***localization error*** by the normalized L2 distance, $\|(x_1 - x_2, y_1 - y_2)\|_2/\sqrt{2} \in [0, 1]$, where $(x_1, y_1), (x_2, y_2)$ are relative coordinates. The model learns to use the standard formula to calculate a distance or angle. To analyze the ***arithmetic error***, we calculate the MRE between the model's prediction and the output of a Python execution of the same formula. The arithmetic error reflects the model's ability to perform the calculation, independent of the localization error. The final ***measurement error*** is a coupling of the localization and arithmetic errors. Table 2 shows that the localization of tumor/lesion landmarks is more challenging than that of anatomical landmarks in angle/distance tasks, since the former involves irregular and variable structures. Table 2



Figure 3: Model response text and the corresponding visualization of the predicted target and the ground truth, for the detection, tumor/lesion size, angle measurement, and distance measurement tasks (top to bottom).

also reveals that the arithmetic error is considerably higher in angle measurement than in distance measurement. To address this, we applied additional SFT with function calling to MedVision-V0, yielding a tool-use variant (MedVision-V0-tool). Although this eliminates the arithmetic error entirely, the final measurement error improves only marginally, indicating that ***limited visual perception remains the bottleneck in quantitative medical image measurement tasks***.

Table 2: Error analysis of MedVision-V0 and a tooling variant, MedVision-V0-tool, on A/D and T/L tasks. Sample sizes are given in parentheses. normL2: normalized L2 distance; MRE: mean relative error; reported in %.

Model/Task	Localization (normL2)	Arithmetic (MRE)	Measurement (MRE)
MedVision-V0			
A/D	1.6	12.8	27.7
Distance	1.5	1.4	6.4
Angle	1.6	25.9	52.1
T/L	6.0	3.4	26.0
MedVision-V0-tool			
A/D	1.4	0.0	25.6
Distance	1.4	0.0	6.7
Angle	1.3	0.0	47.4
T/L	6.2	0.0	26.3

Performance. Summary metrics for all evaluated models on all tasks are given in Section 4.2. Here we provide a more detailed breakdown of MedVision-V0's performance on the detection (Table 3) and T/L tasks (Table 4). The model performs better on anatomical structure detection than on tumor/lesion detection. The relative error of lesion size estimation ranges from 13% to 33%. Label-level detection performance and the effect of target size on detection are provided in Appendix O, which shows that ***small-target and tumor/lesion detection are more challenging***.

4.2 Comparative Benchmarking Results

We compare MedVision-V0 with 17 other VLMs on the three task families. The results are visualized in Figure 4 and detailed in Tables 5, 6, and 7; per-model profiles with per-sample metric distributions are given in Figures 14 and 15 in Appendix N. MedVision-V0 shows superior performance across these tasks, demonstrating the effectiveness of our SFT-RFT training strategy. Among the off-the-shelf models, the strongest results come from the most recent general-purpose VLMs (Gemma4,

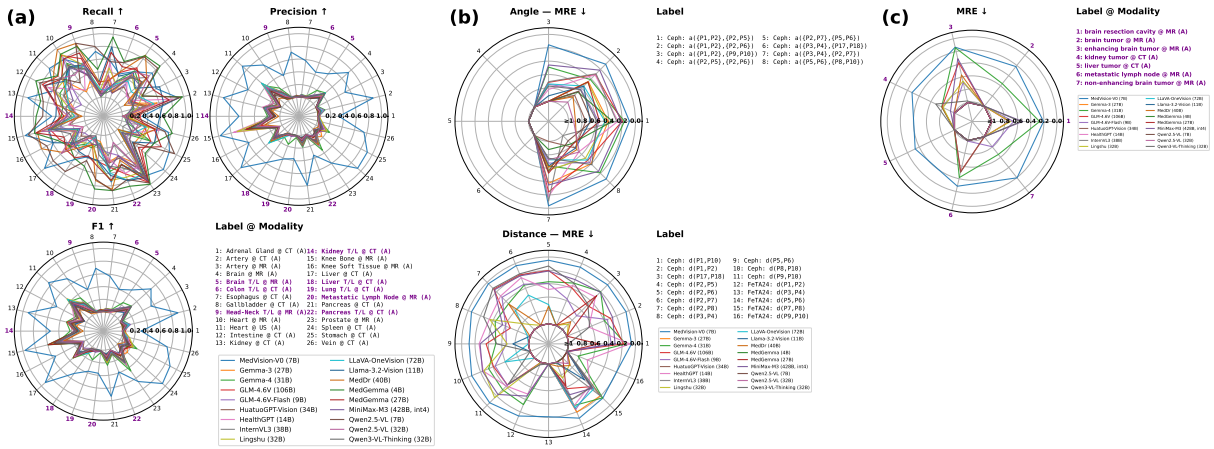


Figure 4: Performance of evaluated models on (a) detection, (b) angle/distance measurement, and (c) tumor/lesion size estimation tasks. The MRE axes are inverted so that the outermost ring corresponds to MRE $= 0$ and the center to MRE ≥ 1, i.e., ***larger polygons always indicate better performance***. In (b), MedVision-V0 attains near-zero angle MRE due to outliers; the distribution of its angle MRE is shown in Figure 10(b), where the median MRE for three of the angle measurements lies between 40% and 60%. Per-model breakdowns are provided in Figures 14 and 15 in Appendix N.

Table 3: Detection performance of MedVision-V0, with the original labels categorized into 26 groups. Metrics are reported in %. R: recall; P: precision; F1: F1 score; IoU: intersection over union; SR: success rate.

Label	R ↑	P ↑	F1 ↑	IoU ↑	SR ↑	$\text{IoU}_{>0.5}$ ↑	$\text{IoU}_{0.5:0.05:0.95}$ ↑
Anatomy (18 regions, 13.4K samples)							
Knee Bone @ MR	96.3	97.2	96.7	94.3	100.0	99.2	94.2
Spleen @ CT	96.1	93.3	94.3	90.4	100.0	98.0	86.8
Liver @ CT	93.3	91.7	91.9	88.0	100.0	94.9	84.1
Kidney @ CT	92.2	92.3	91.8	87.4	100.0	95.8	82.8
Artery @ CT	91.8	90.3	90.8	85.3	100.0	97.3	78.1
Stomach @ CT	87.5	91.4	88.5	82.3	100.0	93.0	72.9
Heart @ US	83.5	88.9	85.0	76.1	100.0	91.6	59.8
Vein @ CT	75.3	75.4	73.3	65.6	100.0	76.2	51.9
Pancreas @ CT	74.0	79.3	73.2	62.6	100.0	74.5	41.2
Brain @ MR	72.6	74.0	70.2	61.0	100.0	67.2	43.4
Heart @ MR	78.7	68.1	70.2	60.7	100.0	67.5	43.0
Gallbladder @ CT	72.8	71.7	69.5	60.4	100.0	68.9	43.7
Prostate @ MR	84.4	59.1	66.3	55.2	100.0	62.7	32.3
Esophagus @ CT	58.1	58.1	56.9	45.1	100.0	44.1	21.6
Knee Soft Tissue @ MR	59.0	58.8	54.8	44.2	100.0	47.9	22.0
Artery @ MR	63.9	51.7	54.8	44.9	100.0	51.5	25.1
Intestine @ CT	57.5	60.3	52.9	39.8	100.0	33.3	11.6
Adrenal Gland @ CT	31.5	33.2	29.8	20.2	100.0	10.2	2.4
Tumor/Lesion (8 regions, 8.5K samples)							
Kidney Tumor/Lesion @ CT	51.2	64.4	53.4	43.9	100.0	47.7	25.5
Liver Tumor/Lesion @ CT	53.3	58.0	50.9	41.6	100.0	46.2	22.7
Brain Tumor/Lesion @ MR	55.4	51.3	48.6	39.5	100.0	42.0	21.5
Colon Tumor/Lesion @ CT	41.3	44.3	40.2	33.5	100.0	35.7	20.7
Head-Neck Tumor/Lesion @ MR	57.6	38.7	40.1	33.0	100.0	34.4	20.4
Pancreas Tumor/Lesion @ CT	51.0	35.4	38.8	27.9	100.0	18.8	7.2
Lung Tumor/Lesion @ CT	37.8	42.0	37.5	30.6	100.0	33.4	16.9
Metastatic Lymph Node @ MR	41.4	37.0	36.7	30.4	100.0	33.5	17.8

GLM-4.6V, Qwen3-VL, and MiniMax-M3) rather than from the specialized medical VLMs (Figure 1, right panel). Even so, every off-the-shelf model remains far behind MedVision-V0: all of them exceed 100% MRE on angle measurement, and only Gemma4 falls below 100% MRE on lesion size estimation. This reveals the limitation of both cutting-edge, general-purpose VLMs and specialized medical VLMs for quantitative medical image analysis, highlighting the value of our benchmark/-dataset and the need for further research in this area.

Table 4: T/L size estimation performance of MedVision-V0 for each label. MRE, SR, and $\text{MRE}_{<0.1}$ are reported in %, while MAE is in millimeters.

Label	MAE ↓	MRE ↓	SR ↑	$\text{MRE}_{<0.1}$ ↑
enhancing brain tumor @ MR	5.2	13.1	100.0	53.6
non-enhancing brain tumor @ MR	6.0	19.6	100.0	38.4
brain resection cavity @ MR	7.8	20.5	100.0	32.5
brain tumor @ MR	6.2	22.3	100.0	38.4
kidney tumor @ CT	12.5	27.5	100.0	18.4
metastatic lymph node @ MR	6.2	28.3	100.0	17.5
liver tumor @ CT	16.4	33.1	100.0	19.5

4.3 Effectiveness of SFT-RFT Training

We study the impact of our training pipeline by comparing our base model with SFT and SFT-RFT post-training on MedVision.

RFT Strategy. MedVision-V0 is trained with RFT on the three tasks sequentially (A/D → T/L → detection). To assess the effect of this schedule, we trained an ablation variant that starts from the same SFT model and undergoes a single ***multi-task RFT*** stage with a curriculum learning scheme (Appendix L). The two variants achieve comparable performance (Table 16 in Appendix M), indicating that multi-task RFT constitutes a viable alternative to the multi-stage strategy. The two strategies nonetheless entail distinct optimization risks: sequential training is susceptible to catastrophic forgetting (Kirkpatrick et al., 2017) of capabilities acquired in earlier stages, whereas multi-task training is subject to data imbalance and joint optimization challenges (Liang et al., 2025). In our

Table 5: Model performance on detection tasks. Labels are categorized into anatomy and tumor/lesion groups. Metrics are reported in %.

Model	R ↑	P ↑	F1 ↑	IoU ↑	SR ↑	$IoU_{>0.5}$ ↑	$IoU_{0.5:0.05:0.95}$ ↑
Anatomy (18 regions, 13.4K samples)							
MedVision-V0 (7B)	81.3	80.4	79.1	72.0	100.0	80.1	59.9
Gemma4 (31B)	34.8	20.7	22.6	16.7	98.5	13.5	5.8
Lingshu (32B)	37.4	20.2	20.2	13.7	100.0	6.7	1.4
GLM-4.6V (106B)	46.5	16.4	20.2	13.5	99.9	6.0	1.8
Qwen3-VL-Thinking (32B)	35.8	17.4	19.5	13.4	96.5	7.5	2.2
MedGemma (27B)	57.2	15.7	19.1	12.9	98.6	6.6	1.8
MedGemma (4B)	68.7	14.7	18.6	12.4	98.9	6.4	1.5
Qwen2.5-VL (32B)	44.8	14.9	18.4	12.5	100.0	6.3	1.8
LLaVA-OneVision (72B)	34.9	19.0	18.1	11.8	100.0	2.4	0.4
MiniMax-M3 (428B)	32.5	15.4	17.5	11.8	99.9	5.6	1.4
InternVL3 (38B)	31.1	17.0	17.2	11.5	100.0	5.3	1.2
Qwen2.5-VL (7B)	69.7	12.2	16.7	11.3	99.4	5.6	1.7
HealthGPT-L14 (14B)	29.6	20.7	16.0	10.2	98.7	1.8	0.4
GLM-4.6V-Flash (9B)	22.6	17.6	15.2	9.8	100.0	2.5	0.5
HuatuoGPT-Vision (34B)	25.6	17.7	15.1	9.8	99.5	2.7	0.6
Gemma3 (27B)	37.1	12.4	14.9	10.1	100.0	4.6	1.3
MedDr (40B)	53.6	11.1	14.6	9.7	96.8	4.2	1.1
Llama3.2-Vision (11B)	52.2	12.7	14.3	9.4	89.4	3.2	0.8
Tumor/Lesion (8 regions, 8.5K samples)							
MedVision-V0 (7B)	52.4	50.5	46.9	38.2	100.0	40.7	21.1
Gemma4 (31B)	43.6	14.6	18.1	12.7	99.5	8.4	2.7
MiniMax-M3 (428B)	35.7	7.8	10.2	6.5	100.0	1.6	0.4
Lingshu (32B)	40.2	6.0	8.6	5.1	100.0	0.2	0.1
HealthGPT-L14 (14B)	29.2	7.0	8.4	5.1	97.8	0.5	0.1
LLaVA-OneVision (72B)	34.1	6.3	8.4	5.0	100.0	0.4	0.1
Qwen3-VL-Thinking (32B)	33.9	6.6	8.2	5.2	98.0	1.3	0.3
InternVL3 (38B)	29.5	6.6	7.9	4.9	100.0	0.8	0.2
Qwen2.5-VL (32B)	38.5	5.7	7.7	4.7	100.0	0.6	0.2
GLM-4.6V (106B)	52.4	5.1	7.7	4.6	99.8	0.5	0.1
GLM-4.6V-Flash (9B)	26.2	6.1	7.5	4.7	100.0	1.0	0.2
MedGemma (27B)	54.1	4.6	7.5	4.3	97.0	0.1	0.0
MedGemma (4B)	77.7	4.4	7.5	4.2	99.3	0.0	0.0
Qwen2.5-VL (7B)	77.4	3.8	6.5	3.6	99.6	0.0	0.0
HuatuoGPT-Vision (34B)	21.6	5.0	6.4	3.8	99.0	0.3	0.1
MedDr (40B)	63.4	3.7	6.2	3.5	98.7	0.1	0.0
Gemma3 (27B)	34.3	4.3	6.1	3.6	100.0	0.3	0.0
Llama3.2-Vision (11B)	49.0	3.6	5.9	3.4	85.4	0.1	0.0

Table 6: Model performance on tumor/lesion size estimation tasks. MRE, SR, and $MRE_{<0.1}$ are reported in %, while MAE is in millimeters.

Model	MAE ↓	MRE ↓	SR ↑	$MRE_{<0.1}$ ↑
Tumor/Lesion (2K samples)				
MedVision-V0 (7B)	10.5	26.0	100.0	23.5
Gemma4 (31B)	21.7	72.6	98.9	16.8
GLM-4.6V (106B)	31.4	107.3	92.2	4.1
MiniMax-M3 (428B)	32.7	108.1	95.8	5.6
GLM-4.6V-Flash (9B)	34.4	109.2	95.2	3.0
Lingshu (32B)	35.7	118.6	99.5	4.5
HealthGPT-L14 (14B)	51.8	176.0	100.0	2.2
Qwen3-VL-Thinking (32B)	53.2	140.6	94.0	6.2
MedDr (40B)	78.0	240.7	87.3	1.9
Llama3.2-Vision (11B)	101.2	329.4	98.3	0.7
Gemma3 (27B)	225.8	611.4	99.0	0.5
MedGemma (27B)	523.4	1905.0	57.0	0.5
HuatuoGPT-Vision (34B)	1018.6	2847.2	98.3	1.5
LLaVA-OneVision (72B)	1089.1	3368.0	100.0	1.2
Qwen2.5-VL (32B)	2054.6	6989.2	99.9	1.3
Qwen2.5-VL (7B)	2897.2	7680.9	95.4	0.7
InternVL3 (38B)	8606.8	25285.1	100.0	0.2
MedGemma (4B)	1109532.7	3733526.8	90.9	0.1

multi-task RFT, we address these challenges with ***temperature-scaled task mixing*** (Appendix K) and epoch-level ***curriculum learning*** (Appendix L). We further ablate the reward design on the multi-task RFT model: the ***multiplicative composition*** of process and answer rewards ($r_{\text{process}} \cdot r_{\text{answer}}$) matches an additive alternative on T/L size estimation and distance measurement, and performs better on angle measurement, whose reasoning trace is more sophisticated (Table 17 in Appendix M).

In-distribution. SFT dramatically improves detection precision, T/L size estimation, and A/D measurement accuracy over the base model; ***the additional RFT step yields further consistent gains across all tasks*** (Table 14 in Appendix M). SFT on MedVision also benefits other base models: Table 15 in Appendix M compares the base → SFT progression for Qwen2.5-VL (7B), Gemma4 (31B), and MedGemma (27B), all of which improve substantially across the three tasks.

Out-of-distribution Generalization. Both SFT and SFT-RFT models generalize to unseen imaging planes (plane-OOD) and unseen anatomical targets (target-OOD), with ***the SFT-RFT model (MedVision-V0) consistently outperforming the SFT-only model*** across detection and T/L size estimation (Table 18 in Appendix M). These settings probe generalization along multiple axes at once: even for an identical subject and target, an unseen imaging plane presents markedly different ***anatomical context, target appearance, and physical extent*** (Figure 11 in Appendix M). Nonetheless, a performance gap relative to the in-distribution setting persists: anatomy detection F1 decreases from 79.1% to 49.5% (plane-OOD) and 36.0% (target-OOD), T/L detection F1 from 46.9% to 36.6% and 45.5%, and T/L size MRE increases from 26.0% to 36.4% and 34.7%.

4.4 Comparison with a Specialist Model

We compare MedVision-V0 with BiomedParse (Zhao et al., 2025), a text-prompted biomedical foundation model for segmentation across diverse anatomical structures and pathologies. Unlike MedVision-V0, which directly predicts bounding box coordinates and lesion axis lengths end-to-end, BiomedParse produces pixel-level segmentation masks and relies on post-processing to derive the final detection and size estimation results. Bounding box and ellipse fitting are applied to convert BiomedParse's segmentation outputs into the target estimates. As shown in Table 8, MedVision-V0 substantially outperforms BiomedParse on both tasks.

We position MedVision-V0 as a generalist model for detection and measurement across anatomies and modalities from language prompts. MedVision-V0 can serve as a measurement tool within larger agentic systems, where an orchestrator (e.g., a radiology report generation model) invokes it on demand. Fine-tuning a segmentation specialist such as BiomedParse on MedVision would primarily

Table 7: Model performance on angle and distance measurement tasks. MRE, $MRE_{<0.1}$, and SR are reported in %. MAEs are given in millimeters (distance) and degrees (angle).

Model	MAE ↓	MRE ↓	SR ↑	$MRE_{<0.1}$ ↑
Distance (Ceph+FeTA, 1,100 samples)				
MedVision-V0 (7B)	3.6	6.4	100.0	81.4
MiniMax-M3 (428B)	18.0	30.4	97.5	22.0
GLM-4.6V (106B)	22.7	41.5	88.4	14.2
HealthGPT-L14 (14B)	23.1	42.4	97.6	18.0
Gemma4 (31B)	23.8	37.9	99.8	10.8
GLM-4.6V-Flash (9B)	30.3	56.0	99.7	12.5
MedDr (40B)	106.3	197.6	91.2	8.5
Qwen3-VL-Thinking (32B)	119.0	197.8	98.4	18.6
Lingshu (32B)	198.8	247.7	100.0	21.4
LLaVA-OneVision (72B)	2541.7	4557.2	100.0	6.1
HuatuoGPT-Vision (34B)	3921.6	6251.5	98.2	5.4
Llama3.2-Vision (11B)	4666.2	7065.8	98.6	2.1
Gemma3 (27B)	5054.2	6606.9	99.0	13.3
MedGemma (27B)	6239.0	8616.4	46.2	6.7
MedGemma (4B)	15222.0	24918.0	95.1	0.1
Qwen2.5-VL (32B)	56045.7	71734.7	99.9	6.3
InternVL3 (38B)	62066.9	62191.8	99.8	8.0
Qwen2.5-VL (7B)	63603.8	96542.0	98.3	0.5
Angle (Ceph, 960 samples)				
MedVision-V0 (7B)	4.7	52.1	99.9	52.0
MiniMax-M3 (428B)	17.9	477.1	48.6	8.5
GLM-4.6V (106B)	19.7	322.7	88.8	20.9
Gemma4 (31B)	23.9	428.0	88.5	11.7
HealthGPT-L14 (14B)	25.9	463.2	99.3	16.7
InternVL3 (38B)	30.3	616.8	100.0	20.4
Qwen3-VL-Thinking (32B)	32.4	378.2	69.6	3.6
LLaVA-OneVision (72B)	33.6	473.5	100.0	2.9
Llama3.2-Vision (11B)	34.2	287.1	100.0	3.2
Lingshu (32B)	35.0	512.5	100.0	6.3
GLM-4.6V-Flash (9B)	35.0	531.7	99.1	8.1
MedGemma (4B)	35.8	298.8	96.7	6.0
Gemma3 (27B)	36.3	702.2	99.9	6.7
Qwen2.5-VL (32B)	41.2	258.3	99.2	1.0
MedDr (40B)	45.7	591.5	94.5	5.3
MedGemma (27B)	46.4	1024.8	93.0	6.8
Qwen2.5-VL (7B)	48.0	724.9	97.6	2.0
HuatuoGPT-Vision (34B)	7070.0	9032.7	95.4	3.7

advance pixel-level segmentation, which deviates from our goal: ***assessing and improving VLMs' perception and reasoning ability for end-to-end (image-text-to-text) quantitative medical image analysis***.

4.5 Clinical Decision Agreement

Absolute error metrics do not reveal whether a measurement error is large enough to alter the resulting clinical decision. We therefore complement the benchmark with a clinical decision agreement (CDA) analysis: each predicted measurement is mapped to a clinical category through a published cutoff table (***maxillary and mandibular position classification*** from cephalometric angles (Steiner, 1953); ***kidney tumor staging*** from AJCC size cutoffs (American Joint Committee on Cancer, 2017)), and the predicted category is compared with the category implied by the ground truth measurement, so that any disagreement is attributable to measurement error alone. Agreement is quantified with chance-corrected, quadratic-weighted Cohen's κ_w. Off-the-shelf VLMs reach at most fair agreement (Landis and Koch, 1977) on every CDA task ($\kappa_w \leq 0.361$), whereas MedVision-V0 attains fair agreement on maxillary position ($\kappa_w = 0.332$), moderate agreement on mandibular position ($\kappa_w = 0.439$), and substantial agreement on kidney tumor staging ($\kappa_w = 0.726$). The measurement gains from post-training thus translate into clinically meaningful decision agreement, a property that absolute error metrics alone do not capture. The complete protocol and per-model results are provided in Appendix P.

Table 8: Comparison of MedVision-V0 with BiomedParse on detection and T/L size estimation; BiomedParse's boxes and sizes are derived from its predicted segmentation masks. R: recall; P: precision. Metrics are reported in %, except MAE (millimeters).

Model	R ↑	P ↑	F1 ↑	IoU ↑	SR ↑	$IoU_{>0.5}$ ↑	$IoU_{0.5:0.05:0.95}$ ↑
Detection: Anatomy							
MedVision-V0 (7B)	**81.3**	**80.4**	**79.1**	**72.0**	**100.0**	**80.1**	**59.9**
BiomedParse	53.9	57.4	53.8	50.2	64.8	55.3	44.5
Detection: Tumor/Lesion							
MedVision-V0 (7B)	**52.4**	**50.5**	**46.9**	**38.2**	**100.0**	**40.7**	**21.1**
BiomedParse	33.7	41.0	32.4	27.7	64.8	30.5	19.0

Model	MAE ↓	MRE ↓	SR ↑	$MRE_{<0.1}$ ↑
T/L Size Estimation				
MedVision-V0 (7B)	**10.5**	**26.0**	**100.0**	23.5
BiomedParse	16.5	52.4	66.4	**36.4**

5 Conclusion

We introduce **MedVision**, a large-scale, multi-modality, multi-anatomy medical imaging dataset with rich quantitative annotations for detection and measurement tasks. Using MedVision, we establish a comprehensive benchmark to evaluate VLMs on detection, tumor/lesion size estimation, and angle/distance measurement. Our results show that off-the-shelf VLMs perform poorly on these quantitative tasks, whereas supervised and reinforcement fine-tuning (SFT-RFT) with MedVision substantially improves accuracy. Moreover, RFT brings additional gains in both performance and generalization to out-of-distribution imaging planes and anatomical targets. Nonetheless, challenges remain, particularly in detecting and measuring small structures and highly variable tumor/lesion targets. We expect MedVision to provide a valuable foundation for advancing quantitative capabilities in medical VLMs.

Acknowledgments

This work was supported by UK Research and Innovation (grant EP/S02431X/1) through the UKRI Centre for Doctoral Training in Biomedical AI at the School of Informatics, University of Edinburgh.

We also acknowledge the support of the Edinburgh International Data Facility (EIDF) and the Data-Driven Innovation Programme at the University of Edinburgh.

Limitations

MedVision is currently generalist in terms of being able to do a wide variety of quantitative tasks with a single VLM, without relying on external tools. However, it is specialist in that it is not simultaneously trained to do all the categorical and qualitative description tasks of mainstream or biomedically focused VLMs. Unifying MedVision-style quantitative analysis datasets and training recipes with categorical and qualitative datasets and training pipelines remains an important open challenge.

Ethical Considerations. The benchmark reveals that current state-of-the-art VLMs are incapable of accurate medical image detection and measurement. The fine-tuned model MedVision-V0, although demonstrating improved performance, remains far from the accuracy and robustness required for clinical application. Such models must not be used to drive any medical diagnosis or clinical decision-making. Details on the intended use of the model and on data privacy are provided in Appendix Q.

References

Felix Ambellan, Alexander Tack, Moritz Ehlke, and Stefan Zachow. 2019. Automated segmentation of knee bone and cartilage combining statistical shape knowledge and convolutional neural networks: Data from the osteoarthritis initiative. *Medical Image Analysis*, 52:109–118.

American Joint Committee on Cancer. 2017. *AJCC Cancer Staging Manual*, eighth edition, corrected at 3rd printing edition. AJCC, American Joint Committee on Cancer, Chicago, IL.

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A. Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M. Summers, Bram van Ginneken, Michel Bilello, Patrick Bilic, Patrick F. Christ, Richard K. G. Do, Marc J. Gollub, Stephan H. Heckers, Henkjan Huisman, William R. Jarnagin, and 39 others. 2022. The medical segmentation decathlon. *Nature Communications*, 13(1):4128.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint*.

Pedro R. A. S. Bassi, Mehmet Can Yavuz, Ibrahim Ethem Hamamci, Sezgin Er, Xiaoxi Chen, Wenxuan Li, Bjoern Menze, Sergio Decherchi, Andrea Cavalli, Kang Wang, Yang Yang, Alan Yuille, and Zongwei Zhou. 2025. Radgpt: Constructing 3d image-text tumor datasets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23720–23730.

Asma Ben Abacha, Sadid A Hasan, Vivek V Datla, Dina Demner-Fushman, and Henning Müller. 2019. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In *Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes*. 9-12 September 2019.

Asma Ben Abacha, Mourad Sarrouti, Dina Demner-Fushman, Sadid A Hasan, and Henning Müller. 2021. Overview of the vqa-med task at imageclef 2021: Visual question answering and generation in the medical domain. In *Proceedings of the CLEF 2021 Conference and Labs of the Evaluation Forum-working notes*. 21-24 September 2021.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 41–48, New York, NY, USA. Association for Computing Machinery.

Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, Gerard Sanroma, Sandy Napel, Steffen Petersen, Georgios Tziritas, Elias Grinias, Mahendra Khened, Varghese Alex Kollerathu, Ganapathy Krishnamurthi, Marc-Michel Rohe, and 18 others. 2018. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE transactions on medical imaging*, 37(11):2514–2525.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. 2022. Making the most of text semantics to improve biomedical vision–language processing. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, volume 13696, pages 1–21. Springer Nature Switzerland, Cham.

Federico Bolelli, Luca Lumetti, Shankeeth Vinayahalingam, Mattia Di Bartolomeo, Arrigo Pellacani, Kevin Marchesini, Niels van Nistelrooij, Pieter van Lierop, Tong Xi, Yusheng Liu, Rui Xin, Tao Yang,

Lisheng Wang, Haoshen Wang, Chenfan Xu, Zhiming Cui, Marek Wodzinski, Henning Müller, Yannick Kirchhoff, and 15 others. 2025a. Segmenting the inferior alveolar canal in cbct volumes: The toothfairy challenge. *IEEE Transactions on Medical Imaging*, 44(4):1890–1906.

Federico Bolelli, Kevin Marchesini, Niels van Nistelrooij, Luca Lumetti, Vittorio Pipoli, Elisa Ficarra, Shankeeth Vinayahalingam, and Costantino Grana. 2025b. Segmenting maxillofacial structures in cbct volumes. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5238–5248.

Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P. Langlotz. 2024. Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats. *arXiv preprint*.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, and Benyou Wang. 2024a. Towards injecting medical visual knowledge into multimodal llms at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7346–7370, Miami, Florida, USA. Association for Computational Linguistics.

Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, Benyou Wang, Shaoting Zhang, Bin Fu, Jianfei Cai, Bohan Zhuang, Eric J. Seibel, Yu Qiao, and Junjun He. 2024b. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. *Advances in Neural Information Processing Systems*, 37:94327–94427.

Jack C. Cheng, René M. Castelein, Winnie C. Chu, Aina J. Danielsson, Matthew B. Dobbs, Theodoros B. Grivas, Christina A. Gurnett, Keith D. Luk, Alain Moreau, Peter O. Newton, Ian A. Stokes, Stuart L. Weinstein, and R. Geoffrey Burwell. 2015. Adolescent idiopathic scoliosis. *Nature Reviews Disease Primers*, 1(1):15030.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. 2025. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 10818–10838. PMLR.

Maria Correia de Verdier, Rachit Saluja, Louis Gagnon, Dominic LaBella, Ujjwall Baid, Nourel Hoda Tahon, Martha Foltyn-Dumitru, Jikai Zhang, Maram Alafif, Saif Baig, Ken Chang, Gennaro D'Anna, Lisa Deptula, Diviya Gupta, Muhammad Ammar Haider, Ali Hussain, Michael Iv, Marinos Kontzialis, Paul Manning, and 66 others. 2024. The 2024 brain tumor segmentation (brats) challenge: Glioma segmentation on post-treatment mri. *arXiv preprint*.

Arjun Desai, Andrew Schmidt, Elka Rubin, Christopher Sandino, Marianne Black, Valentina Mazzoli, Kathryn Stevens, Robert Boutin, Christopher Ré, Garry Gold, Brian Hargreaves, and Akshay Chaudhari. 2021. Skm-tea: A dataset for accelerated mri reconstruction with dense image labels for quantitative clinical evaluation. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 1.

Frank C. Detterbeck, Gavitt A. Woodard, Anna S. Bader, Sanja Dacic, Michael J. Grant, Henry S. Park, and Lynn T. Tanoue. 2024. The proposed ninth edition tnm classification of lung cancer. *CHEST*, 166(4):882–895.

Reuben Dorent, Aaron Kujawa, Marina Ivory, Spyridon Bakas, Nicola Rieke, Samuel Joutard, Ben Glocker, Jorge Cardoso, Marc Modat, Kayhan Batmanghelich, Arseniy Belkov, Maria Baldeon Calisto, Jae Won Choi, Benoit M. Dawant, Hexin Dong, Sergio Escalera, Yubo Fan, Lasse Hansen, Mattias P. Heinrich, and 21 others. 2023. Crossmoda 2021 challenge: Benchmark of cross-modality domain adaptation techniques for vestibular schwannoma and cochlea segmentation. *Medical Image Analysis*, 83:102628.

E. A. Eisenhauer, P. Therasse, J. Bogaerts, L. H. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, L. Rubinstein, L. Shankar, L. Dodd, R. Kaplan, D. Lacombe, and J. Verweij. 2009. New response evaluation criteria in solid tumours: Revised recist guideline (version 1.1). *European Journal of Cancer*, 45(2):228–247.

Asser Abou Elkassem, Brian C. Allen, Kedar G. Sharbidre, Soroush Rais-Bahrami, and Andrew D. Smith. 2021. Update on the role of imaging in clinical staging and restaging of renal cell carcinoma based on the ajcc 8th edition, from the ajr special series on cancer staging. *AJR. American journal of roentgenology*, 217(3):541–555.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, and 304 others. 2026. Gemma 4 technical report. *arXiv preprint*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *arXiv preprint*.

GLM-V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale

Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, and 74 others. 2026. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint*.

Sunan He, Yuxiang Nie, Hongmei Wang, Shu Yang, Yihui Wang, Zhiyuan Cai, Zhixuan Chen, Yingxue Xu, Linshan Wu, Ngai Shing Cheng, Luyang Luo, Huiling Xiang, Xi Lin, Mingxiang Wu, Yifan Peng, George Shih, Ziyang Xu, Xian Wu, Qiong Wang, and 8 others. 2026. Towards generalizable ai in medicine via generalist-specialist collaboration. *Nature Biomedical Engineering*.

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

Yuting He, Guanyu Yang, Jian Yang, Yang Chen, Youyong Kong, Jiasong Wu, Lijun Tang, Xiaomei Zhu, Jean-Louis Dillenseger, Pengfei Shao, Shaobo Zhang, Huazhong Shu, Jean-Louis Coatrieux, and Shuo Li. 2020b. Dense biased networks with deep priori anatomy and hard region adaptation: Semi-supervised learning for fine renal artery segmentation. *Medical Image Analysis*, 63:101722.

Yuting He, Guanyu Yang, Jian Yang, Rongjun Ge, Youyong Kong, Xiaomei Zhu, Shaobo Zhang, Pengfei Shao, Huazhong Shu, Jean-Louis Dillenseger, Jean-Louis Coatrieux, and Shuo Li. 2021. Meta grayscale adaptive network for 3d integrated renal structures segmentation. *Medical Image Analysis*, 71:102055.

Nicholas Heller, Fabian Isensee, Dasha Trofimova, Resha Tejpaul, Zhongchen Zhao, Huai Chen, Lisheng Wang, Alex Golts, Daniel Khapun, Daniel Shats, Yoel Shoshan, Flora Gilboa-Solomon, Yasmeen George, Xi Yang, Jianpeng Zhang, Jing Zhang, Yong Xia, Mengran Wu, Zhiyang Liu, and 26 others. 2023. The kits21 challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase ct. *arXiv preprint*.

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

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, volume 33 of *AAAI'19/IAAI'19/EAAI'19*, pages 590–597, Honolulu, Hawaii, USA. AAAI Press.

Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhang, Wanling Ma, Xiang Wan, and Ping Luo. 2022. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pages 36722–36732, Red Hook, NY, USA. Curran Associates Inc.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317.

Maxime Kayser, Cornelius Emde, Oana-Maria Camburu, Guy Parsons, Bartlomiej Papiez, and Thomas Lukasiewicz. 2022. Explaining chest x-ray pathologies in natural language. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part V*, pages 701–713, Berlin, Heidelberg. Springer-Verlag.

Anahita Fathi Kazerooni, Nastaran Khalili, Xinyang Liu, Deep Gandhi, Zhifan Jiang, Syed Muhammed Anwar, Jake Albrecht, Maruf Adewole, Udunna Anazodo, Hannah Anderson, Ujjwal Baid, Timothy Bergquist, Austin J. Borja, Evan Calabrese, Verena Chung, Gian-Marco Conte, Farouk Dako, James Eddy, Ivan Ezhov, and 52 others. 2024. The brain tumor segmentation in pediatrics (brats-peds) challenge: Focus on pediatrics (cbtn-connect-dipgr-asnr-miccai brats-peds). *arXiv preprint*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Dominic LaBella, Valeriia Abramova, Mehdi Astaraki, Andre Ferreira, Zhifan Jiang, Mason C. Cleveland, Ramandeep Kang, Uma M. Lal-Trehan Estrada, Cansu Yalcin, Rachika E. Hamadache, Clara Lisazo, Adrià Casamitjana, Joaquim Salvi, Arnau Oliver, Xavier Lladó, Iuliana Toma-Dasu, Tiago Jesus, Behrus Puladi, Jens Kleesiek, and 86 others. 2025. Analysis of the 2024 brats meningioma radiotherapy planning automated segmentation challenge. *arXiv preprint*.

Xunhao Lai, Weiqi Xu, Yufeng Yang, Qiaorui Chen, Yang Xu, Lunbin Zeng, Xiaolong Li, Haohai Sun, Haichao Zhu, Vito Zhang, Jinkai Hu, Jiayao Li, Rui Gao, Zekun Li, Songquan Zhu, Jingkai Zhou, and

Pengyu Zhao. 2026. Minimax sparse attention. *arXiv preprint*.

J. R. Landis and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

LASA Team, Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, Yu Sun, Junao Shen, Chaojun Wang, Jie Tan, Deli Zhao, Tingyang Xu, Hao Zhang, and Yu Rong. 2025. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint*.

Jason J. Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5(1):180251.

Sarah Leclerc, Erik Smistad, Joao Pedrosa, Andreas Ostvik, Frederic Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Pierre-Marc Jodoin, Thomas Grenier, Carole Lartizien, Jan Dhooge, Lasse Lovstakken, and Olivier Bernard. 2019. Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. *IEEE transactions on medical imaging*, 38(9):2198–2210.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint*.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pages 28541–28564, Red Hook, NY, USA. Curran Associates Inc.

Wenxuan Li, Chongyu Qu, Xiaoxi Chen, Pedro R.A.S. Bassi, Yijia Shi, Yuxiang Lai, Qian Yu, Huimin Xue, Yixiong Chen, Xiaorui Lin, Yutong Tang, Yining Cao, Haoqi Han, Zheyuan Zhang, Jiawei Liu, Tiezheng Zhang, Yujiu Ma, Jincheng Wang, Guang Zhang, and 2 others. 2024b. Abdomenatlas: A large-scale, detailed-annotated, & multi-center dataset for efficient transfer learning and open algorithmic benchmarking. *Medical Image Analysis*, 97:103285.

Yiqing Liang, Jielin Qiu, Wenhao Ding, Zuxin Liu, James Tompkin, Mengdi Xu, Mengzhou Xia, Zhengzhong Tu, Laixi Shi, and Jiacheng Zhu. 2025. Modomodo: Multi-domain data mixtures for multimodal llm reinforcement learning. *arXiv preprint*.

Tianwei Lin, Wenqiao Zhang, Sijing Li, Yuqian Yuan, Binhe Yu, Haoyuan Li, Wanggui He, Hao Jiang, Mengze Li, Song Xiaohui, Siliang Tang, Jun Xiao, Hui Lin, Yueting Zhuang, and Beng Chin Ooi. 2025. Healthgpt: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. In *Forty-Second International Conference on Machine Learning*.

Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, volume 14227, pages 525–536. Springer Nature Switzerland, Cham.

Claudia Lindner, Ching-Wei Wang, Cheng-Ta Huang, Chung-Hsing Li, Sheng-Wei Chang, and Tim F. Cootes. 2016. Fully automatic system for accurate localisation and analysis of cephalometric landmarks in lateral cephalograms. *Scientific Reports*, 6(1):33581.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654, Nice, France. IEEE.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

Luca Lumetti, Vittorio Pipoli, Federico Bolelli, Elisa Ficarra, and Costantino Grana. 2024. Enhancing patch-based learning for the segmentation of the mandibular canal. *IEEE Access*, 12:79014–79024.

Jun Ma, Yao Zhang, Song Gu, Cheng Ge, Shihao Mae, Adamo Young, Cheng Zhu, Xin Yang, Kangkang Meng, Ziyan Huang, Fan Zhang, Yuanke Pan, Shoujin Huang, Jiacheng Wang, Mingze Sun, Rongguo Zhang, Dengqiang Jia, Jae Won Choi, Natália Alves, and 18 others. 2024. Unleashing the strengths of unlabelled data in deep learning-assisted pan-cancer abdominal organ quantification: The flare22 challenge. *The Lancet Digital Health*, 6(11):e815–e826.

Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyuan Wang, Xin Liu, Shucheng Cao, Qi Zhang, Shangqing Liu, Yunpeng Wang, Yuhui Li, Jian He, and Xiaoping Yang. 2022. Abdomenct-1k: Is abdominal organ segmentation a solved problem? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10_Part_2):6695–6714.

Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. 2026. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, pages 337–347, Cham. Springer Nature Switzerland.

Kelly Payette, Hongwei Bran Li, Priscille De Dumast, Roxane Licandro, Hui Ji, Md Mahfuzur Rahman Siddiquee, Daguang Xu, Andriy Myronenko, Hao Liu, Yuchen Pei, Lisheng Wang, Ying Peng, Juanying Xie, Huiquan Zhang, Guiming Dong, Hao Fu, Guotai Wang, ZunHyan Rieu, Donghyeon Kim, and 39 others. 2023. Fetal brain tissue annotation and segmentation challenge results. *Medical Image Analysis*, 88:102833.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, and 62 others. 2026. Medgemma technical report. *arXiv preprint*.

Pengfei Shao, Chao Qin, Changjun Yin, Xiaoxin Meng, Xiaobing Ju, Jie Li, Qiang Lv, Wei Zhang, and Zhengquan Xu. 2011. Laparoscopic partial nephrectomy with segmental renal artery clamping: Technique and clinical outcomes. *European Urology*, 59(5):849–855.

Pengfei Shao, Lijun Tang, Pu Li, Yi Xu, Chao Qin, Qiang Cao, Xiaobing Ju, Xiaoxin Meng, Qiang Lv, Jie Li, Wei Zhang, and Changjun Yin. 2012. Precise segmental renal artery clamping under the guidance of dual-source computed tomography angiography during laparoscopic partial nephrectomy. *European Urology*, 62(6):1001–1008.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint*.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, EuroSys '25, pages 1279–1297, New York, NY, USA. Association for Computing Machinery.

Cecil C. Steiner. 1953. Cephalometrics for you and me. *American Journal of Orthodontics*, 39(10):729–755.

Sanjay Subramanian, Lucy Lu Wang, Ben Bogin, Sachin Mehta, Madeleine van Zuylen, Sravanthi Parasa, Sameer Singh, Matt Gardner, and Hannaneh Hajishirzi. 2020. Medicat: A dataset of medical images, captions, and textual references. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2112–2120, Online. Association for Computational Linguistics.

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

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. 2025. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Nature Communications*, 16(1):7866.

Kaiyuan Yang, Fabio Musio, Yihui Ma, Norman Juchler, Johannes C. Paetzold, Rami Al-Maskari, Luciano Höher, Hongwei Bran Li, Ibrahim Ethem Hamamci, Anjany Sekuboyina, Suprosanna Shit, Houjing Huang, Chinmay Prabhakar, Ezequiel de la Rosa, Bastian Wittmann, Diana Waldmannstetter, Florian Kofler, Fernando Navarro, Martin J. Menten, and 96 others. 2026. The topcow challenge – topology-aware circle of willis segmentation for ct and mr angiography. *NEJM AI*, 3(8).

Yongcheng Yao and Weitian Chen. 2025. Quantifying knee cartilage shape and lesion: From image to metrics. In *Applications of Medical Artificial Intelligence*, pages 162–172, Cham. Springer Nature Switzerland.

Yongcheng Yao, Junru Zhong, Liping Zhang, Sheheryar Khan, and Weitian Chen. 2024. Cartimorph: A framework for automated knee articular cartilage morphometrics. *Medical Image Analysis*, 91:103035.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, juncai liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, and 17 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 113222–113244. Curran Associates, Inc.

Theodore Zhao, Yu Gu, Jianwei Yang, Naoto Usuyama, Ho Hin Lee, Sid Kiblawi, Tristan Naumann, Jianfeng Gao, Angela Crabtree, Jacob Abel, Christine Moung-Wen, Brian Piening, Carlo Bifulco, Mu Wei, Hoifung Poon, and Sheng Wang. 2025. A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. *Nature Methods*, 22(1):166–176.

Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. 2020. Distance-iou loss: Faster and better learning for bounding box regression. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):12993–13000.

Tianhong Zhou, Yin Xu, Yingtao Zhu, Chuxi Xiao, Haiyang Bian, Lei Wei, and Xuegong Zhang. 2025. Drvd-bench: Do vision-language models reason like human doctors in medical image diagnosis? In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint*.

Yongshuo Zong, Oisin Mac Aodha, and Timothy M. Hospedales. 2025. Self-supervised multimodal learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(07):5299–5318.

A Dataset and Benchmark Comparison

Early medical image–text dataset construction mainly targeted visual question answering (VQA), including VQA-RAD (Lau et al., 2018), VQA-Med (Ben Abacha et al., 2019, 2021), SLAKE (Liu et al., 2021), and PathVQA (He et al., 2020a). These resources support structured evaluation of medical image understanding via question–answer pairs, but the evaluations remain predominantly categorical. Large-scale image-text corpora such as CheXpert (Irvin et al., 2019), CheXpert Plus (Chambon et al., 2024), MIMIC-CXR (Johnson et al., 2019), MIMIC-NLE (Kayser et al., 2022), MS-CXR (Boecking et al., 2022), and PMC-OA (Lin et al., 2023) further support the development of VLMs by aligning medical images with clinical reports or natural language explanations from experts. A comparison of existing medical image–text datasets is summarized in Table 9. Notably, these datasets lack quantitative annotations, limiting their utility for evaluating models on measurement tasks.

Building on these resources, recent VLMs have been adapted to the medical domain, including LLaVA-Med (Li et al., 2023), MedDr (He et al., 2026), HuatuoGPT-Vision (Chen et al., 2024a), HealthGPT-L14 (14B) (Lin et al., 2025), MedGemma (Sellergren et al., 2026), Lingshu (LASA Team et al., 2025), and MedVLM-R1 (Pan et al., 2026). Several recent works have aimed to systematically evaluate medical VLMs, including GMAI-MMBench (Chen et al., 2024b), OmniMedVQA (Hu et al., 2024), RadBench (Wu et al., 2025), and DrVD-Bench (Zhou et al., 2025). These benchmarks emphasize descriptive or categorical tasks such as image understanding and report generation, with limited evaluation of quantitative measurement abilities.

Table 9: Comparison of medical datasets. XR: X-ray; MR: Magnetic Resonance Imaging; CT: Computed Tomography; PET: Positron Emission Tomography; US: Ultrasound.

Dataset	Size	Anatomy	Modality	Quantitative
VQA-RAD (Lau et al., 2018)	3.5K	multiple	XR, CT, MR	✗
VQA-Med (Ben Abacha et al., 2019, 2021)	5K	multiple	XR, CT, MR, US, PET	✗
SLAKE (Liu et al., 2021)	14K	multiple	XR, CT, MR	✗
PathVQA (He et al., 2020a)	33K	N/A	pathology	✗
CheXpert (Irvin et al., 2019)	224K	chest	XR	✗
CheXpert Plus (Chambon et al., 2024)	223K	chest	XR	✗
MIMIC-CXR (Johnson et al., 2019)	377K	chest	XR	✗
MIMIC-NLE (Kayser et al., 2022)	38K	chest	XR	✗
MS-CXR (Boecking et al., 2022)	1.2K	chest	XR	✗
PMC-OA (Lin et al., 2023)	1.6M	multiple	diverse	✗
MedVision	24.3M single-instance 45.3M multi-instance	multiple	XR, CT, MR, US, PET	✓

B Dataset Examples

Figure 5 shows preprocessed images of each dataset and examples of three types of quantitative annotations.

C Full Dataset Statistics

Figure 6 visualizes the composition across modalities, anatomies, and tasks, and Table 10 provides the full statistics of the MedVision dataset v1.0.0, including the number of samples for each annotation type (bounding box, tumor/lesion size, and angle/distance). This detailed breakdown highlights the diversity and scale of the dataset, which supports comprehensive evaluation of VLMs on a wide range of quantitative medical imaging tasks.

D VLMs Version

Table 11 lists the HuggingFace model IDs of the evaluated open-weight VLMs.

E VLM Image Processing Strategy

Different VLMs define their own image processing strategy in the image processor classes (or equivalent modules), where image processing pipelines including resizing, cropping, padding, and rescaling are defined. We summarized the image resizing strategies of the evaluated VLMs in Table 12, upon which we designed appropriate pixel size adjustments for conveying physical spacing information to VLMs.

F LLM-Assisted Response Parsing

The rule-based parser accepts only answers written inside the `<answer></answer>` tags, so a response

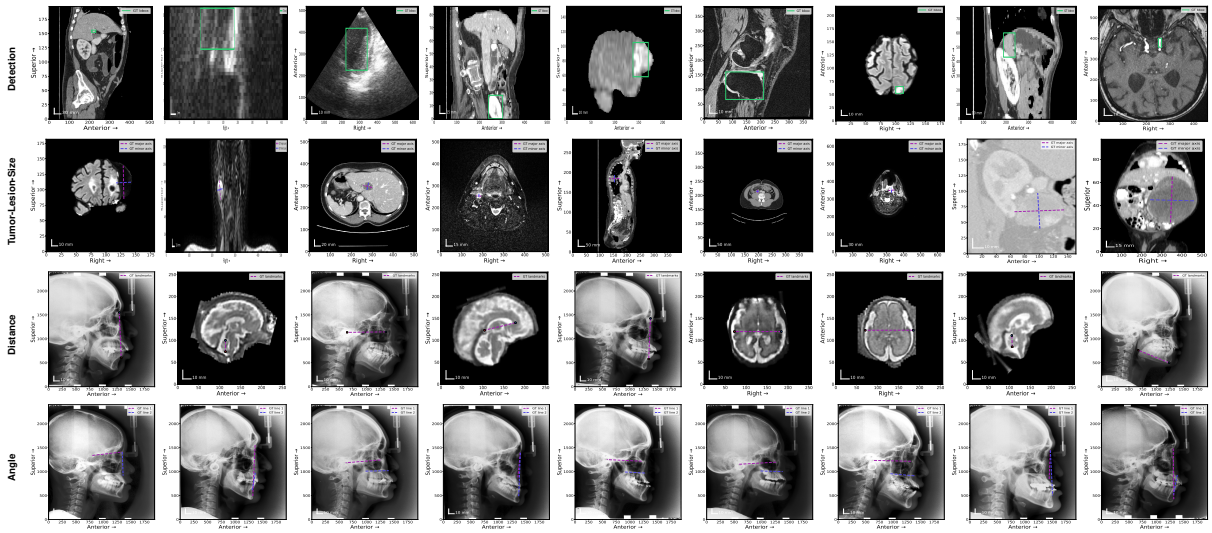


Figure 5: Preprocessed images and annotations from the MedVision dataset.

that states a correct measurement in any other form (e.g., `\boxed{...}`, "Answer: ...", or plain text) is scored as a parsing failure. This conflates two distinct failure modes: the inability to measure and the inability to follow the output format. Across the evaluated models, the rule-based parser rejects 7.8% of detection, 18.3% of T/L, and 21.0% of A/D responses (Table 13), and most rejected responses contain a complete answer in a different wrapper. We therefore re-read every response with an LLM parser (Gemma4, 31B) whose only job is to locate the stated final answer, and compute all reported metrics from the re-parsed records.

Design. The pipeline is built so that the LLM parser cannot corrupt a result. ***(i)*** The parser sees only the response text, never the image or the ground truth, so it cannot favor or penalize any model. ***(ii)*** It is not trusted to copy numbers: it must quote the sentence in which it found the answer, the quote is verified to occur in the response, and the numerical values are then re-extracted from the response text itself, so a value the model never wrote cannot be scored. ***(iii)*** Recovery is strictly additive: a response parsed by the rule-based parser keeps its originally extracted values, and the LLM parser's output is used only for the responses the rule-based parser rejected. Each recovered record additionally stores the quoted sentence, so any recovered value can be audited against the original response. Algorithm 1 summarizes the parsing of one response.

Outcomes and Effect. Each response receives one of four outcomes: ***answer in expected format*** (rule-parsed), ***answer in another format*** (recovered and verified), ***no answer stated***, or ***undetermined*** (the LLM parser output was unusable and the rule-based parser also failed). Only the first two outcomes count as successfully parsed. Pooled over the evaluated models, parsing success rises from 92.2% to 98.6% on detection, from 81.7% to 94.5% on T/L size estimation, and from 79.0% to 93.9% on A/D measurement (Table 13); the gap between the rule-parsed and LLM-parsed reports measures how much of a model's apparent failure was formatting rather than measurement.

G Metric Definition

Success Rate (SR). For all tasks, success rate is defined as the proportion of model responses where the required numeric values can be parsed. SR indicates the model's instruction-following ability.

Detection Metrics. For detection tasks, we calculated precision (P), recall (R), F1 score, and intersection over union (IoU) from the true positive (TP), false positive (FP), and false negative (FN) area. Given a ground truth bounding box B_y and a predicted bounding box B_x, the TP, FP, and FN areas are defined as follows:

$$\text{TP} = \text{area}(B_x \cap B_y), \tag{1}$$
$$\text{FP} = \text{area}(B_x - \text{TP}), \tag{2}$$
$$\text{FN} = \text{area}(B_y - \text{TP}). \tag{3}$$

Precision, recall, F1 score, and IoU are then calcu-

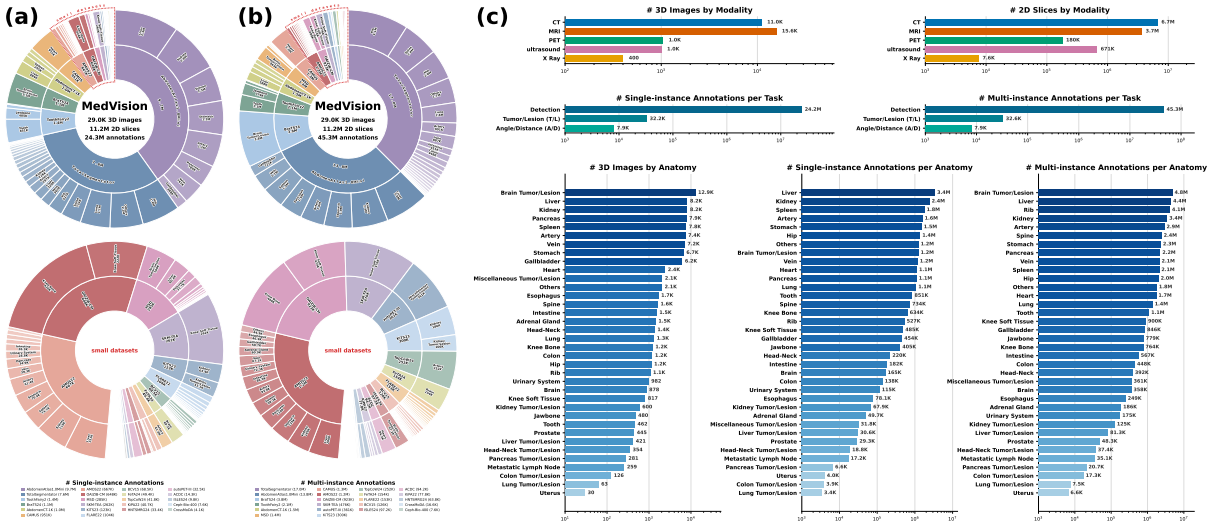


Figure 6: Composition of the MedVision dataset, computed on annotation version v1.0.0. (a) Single-instance and (b) multi-instance annotations broken down by dataset (inner ring) and anatomical subgroup (outer rings); the lower panel zooms into the smallest datasets. (c) Number of 3D images and 2D slices per modality, number of annotations per task, and number of 3D images and annotations per anatomy. The latest data release is available in the data explorer.

Algorithm 1 LLM-assisted parsing of one model response.

Input: response text R; required value count k (detection: $k{=}4$, box coordinates; T/L: $k{=}2$, major and minor axis lengths; A/D: $k{=}1$, the measurement)

Process:

$V_r \leftarrow \textsc{RuleParse}(R)$ ▷ read <answer> tags
if V_r contains k values **then**
  return (V_r, *answer in expected format*)
end if
$(\hat{V}, Q) \leftarrow \textsc{LlmLocate}(R)$ ▷ LLM returns the stated values $\hat{V}$ and a verbatim quote Q of the sentence stating them
if LLM reports no answer **then**
  return ($\varnothing$, *no answer stated*)
end if
if Q occurs in R **then** ▷ quote verification
  $V \leftarrow \textsc{ReExtract}(R, Q, k)$ ▷ k numbers read from R at Q, not from the LLM
  if V matches $\hat{V}$ **then**
    return (V, *answer in another format*)
  end if
end if
return ($\varnothing$, *undetermined*)

Output: the parsed values V, or $\varnothing$ when the response is unparsed, together with one of the four outcome labels: *answer in expected format*, *answer in another format*, *no answer stated*, or *undetermined*

lated as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{4}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{5}$$

$$\text{F1} = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}}, \tag{6}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}, \tag{7}$$

where the region-based F1 score is termed the Dice Similarity Coefficient (DSC) in image segmentation literature. We also reported $\text{IoU}_{>k}$, the proportion of samples with IoU greater than k, defined as:

$$\text{IoU}_{>k} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(\text{IoU}_i - k). \tag{8}$$

Analogous to the threshold-sweep convention in object detection benchmarks, $\text{IoU}_{0.5:0.05:0.95}$ averages $\text{IoU}_{>k}$ over 10 thresholds from 0.5 to 0.95 in steps of 0.05:

$$\text{IoU}_{0.5:0.05:0.95} = \frac{1}{10} \sum_{k \in \{0.5, 0.55, \ldots, 0.95\}} \text{IoU}_{>k}. \tag{9}$$

Measurement Metrics. For the tumor/lesion size and angle/distance measurement tasks, we calculated the mean absolute error (MAE) and mean

Table 10: Full statistics of the MedVision dataset, computed on annotation version v1.0.0. Availability is marked next to the dataset name: ♢ openly available; ♠ registration required; ♣ redistributed by us. Every dataset is partitioned into training and test sets at a 7:3 ratio. A benchmark sample is a (2D slice, target) pair. **Single-instance** annotations retain a target only when it appears as a single, sufficiently large instance on the slice; these are the annotations used for benchmarking and fine-tuning. **Multi-instance** annotations retain every annotated target regardless of how many instances it has. MRI: Magnetic Resonance Imaging; CT: Computed Tomography; PET: Positron Emission Tomography; US: Ultrasound; Box: bounding box; T/L: tumor/lesion size; A/D: angle/distance.

Dataset	Anatomy	Modality	# 3D Image	# Single-instance Annotation			# Multi-instance Annotation		
				Box	T/L	A/D	Box	T/L	A/D
AbdomenAtlas♢	abdomen	CT	5,195	9,748,290	0	0	13,770,398	0	0
AbdomenCT-1K♢	abdomen	CT	1,000	1,041,588	0	0	1,549,325	0	0
ACDC♢♣	heart	MRI	300	14,271	0	0	94,160	0	0
AMOS22♢	abdomen	CT, MRI	360	666,532	0	0	1,215,776	0	0
autoPET-III♢♣	whole body	CT, PET	2,076	31,794	735	0	360,638	749	0
BCV15♢♣	abdomen	CT	60	68,543	0	0	125,870	0	0
BraTS24♢♣	brain	MRI	10,632	1,115,524	11,071	0	3,767,594	11,093	0
CAMUS♢♣	heart	US	1,000	951,370	0	0	1,341,433	0	0
Ceph-Bio-400♢♣	head and neck	X-ray	400	0	0	7,600	0	0	7,600
CrossMoDA♢♣	brain	MRI	105	4,076	0	0	16,623	0	0
FeTA24♠	fetal brain	MRI	80	49,087	0	325	153,599	0	325
FLARE22♢♣	abdomen	CT	50	104,211	0	0	152,954	0	0
HNTSMRG24♢	head and neck	MRI	300	32,029	1,392	0	62,424	1,416	0
ISLES24♢♣	brain	MRI	298	9,774	0	0	97,228	0	0
KiPA22♢♣	kidney	CT	70	37,647	3,095	0	74,690	3,095	0
KiTS23♢♣	kidney	CT	489	114,491	8,484	0	291,550	8,540	0
MSD♢	multiple	CT, MRI	3,225	277,451	7,472	0	1,438,472	7,674	0
OAIZIB-CM♢	knee	MRI	507	648,048	0	0	922,989	0	0
SKM-TEA♠	knee	MRI	310	262,338	0	0	475,828	0	0
ToothFairy2♠	tooth	CT	480	1,413,979	0	0	2,131,223	0	0
TopCoW24♢♣	brain	CT, MRI	250	41,829	0	0	251,901	0	0
TotalSegmentator♢♣	multiple	CT, MRI	1,844	7,603,455	0	0	16,979,575	0	0
Total (22)			**29,031**	**24,236,327**	**32,249**	**7,925**	**45,274,250**	**32,567**	**7,925**

Table 11: HuggingFace model ID of the evaluated open-weight VLMs.

Model	HuggingFace Model ID
MedVision-V0 (7B)	`YongchengYAO/MedVision-V0-7B`
Qwen2.5-VL (7B)	`Qwen/Qwen2.5-VL-7B-Instruct`
Qwen2.5-VL (32B)	`Qwen/Qwen2.5-VL-32B-Instruct`
Qwen3-VL-Thinking (32B)	`Qwen/Qwen3-VL-32B-Thinking`
Lingshu (32B)	`lingshu-medical-mllm/Lingshu-32B`
GLM-4.6V (106B)	`zai-org/GLM-4.6V`
GLM-4.6V-Flash (9B)	`zai-org/GLM-4.6V-Flash`
MiniMax-M3 (428B)	`cyankiwi/MiniMax-M3-AWQ-INT4`
Gemma3 (27B)	`google/gemma-3-27b-it`
Gemma4 (31B)	`google/gemma-4-31B-it`
MedGemma (4B)	`google/medgemma-4b-it`
MedGemma (27B)	`google/medgemma-27b-it`
InternVL3 (38B)	`OpenGVLab/InternVL3-38B`
Llama3.2-Vision (11B)	`meta-llama/Llama-3.2-11B-Vision-Instruct`
LLaVA-OneVision (72B)	`llava-hf/llava-onevision-qwen2-72b-ov-hf`
MedDr (40B)	`Sunanhe/MedDr_0401`
HuatuoGPT-Vision (34B)	`FreedomIntelligence/HuatuoGPT-Vision-34B`
HealthGPT-L14 (14B)	`lintw/HealthGPT-L14`

relative error (MRE) between predictions $\mathbf{x}^{(n)}$ and ground truths $\mathbf{y}^{(n)}$ as follows:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{n}\sum_{j=1}^{n}|x_i^j - y_i^j|, \tag{10}$$

$$\text{MRE} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{n}\sum_{j=1}^{n}(|x_i^j - y_i^j|/y_i^j), \tag{11}$$

where N is the total number of samples, and n is the number of values to be measured in each sample. For tumor/lesion size measurement, $n = 2$ (major and minor axis lengths). For angle/distance measurement, $n = 1$.

Since the detection and measurement metrics are calculated only on the successfully parsed predictions, indicating the model prediction accuracy, we also report metrics that reflect both the instruction-following ability and prediction accuracy:

$$\text{MRE}_{<k} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(k - \frac{1}{n}\sum_{j=1}^{n}|x_i^j - y_i^j|/y_i^j), \tag{12}$$

where $\mathbb{I}(x)$ is the step function that returns 1 for $x \geq 0$. $\text{MRE}_{<k}$ indicates the proportion of samples with relative error less than k.

H Computational Resources & Software

Computational Resources. Model evaluation used 4 NVIDIA H100 (80 GB) GPUs; training of

Table 12: Image resize strategies of the evaluated VLMs, i.e., how each model's own image processor transforms an input slice into the canvas its vision encoder perceives. ‡ aspect ratio preserved.

Model	Perceived Image Size	Image Resize Strategy
MedVision-V0 (7B)	dynamic	round each side to a multiple of 28; no pad
Qwen2.5-VL (7B)	dynamic	round each side to a multiple of 28; no pad
Qwen2.5-VL (32B)	dynamic	round each side to a multiple of 28; no pad
Qwen3-VL-Thinking (32B)	dynamic	round each side to a multiple of 32; no pad
Lingshu (32B)	dynamic	round each side to a multiple of 28; no pad
GLM-4.6V (106B)	dynamic	round each side to a multiple of 28; no pad
GLM-4.6V-Flash (9B)	dynamic	round each side to a multiple of 28; no pad
MiniMax-M3 (428B)	dynamic	round each side to a multiple of 28 under a 451,584-px cap; no pad
Gemma4 (31B)	dynamic	scale to a 2,520-patch budget, then floor each side to a multiple of 48; no pad
InternVL3 (38B)	dynamic	stretch onto a dynamic grid of 448x448 tiles; no pad
Llama3.2-Vision (11B)	dynamic	scale‡ to fit within a canvas of 560x560 tiles, then pad to the tile grid
LLaVA-OneVision (72B)	dynamic	scale‡ to fit within a canvas of 384x384 tiles, then pad to the canvas
Gemma3 (27B)	896x896	stretch to a fixed size
MedGemma (4B)	896x896	stretch to a fixed size
MedGemma (27B)	896x896	stretch to a fixed size
MedDr (40B)	448x448	stretch to a fixed size (single tile, tiling off)
HuatuoGPT-Vision (34B)	336x336	pad to square, then CLIP-336 resize + center-crop
HealthGPT-L14 (14B)	336x336	pad to square, then CLIP-336 resize + center-crop

Table 13: Success rate (SR, in %) of the rule-based parser vs. the LLM-assisted parser (LLM), per model and task family. The last row pools all responses of each task family.

	Detection		T/L		A/D	
Model	Rule	LLM	Rule	LLM	Rule	LLM
MedVision-V0 (7B)	100.0	100.0	100.0	100.0	100.0	100.0
Gemma3 (27B)	100.0	100.0	99.0	99.0	99.2	99.4
Gemma4 (31B)	98.6	98.9	98.9	98.9	94.6	94.6
GLM-4.6V (106B)	60.0	99.9	89.5	91.9	50.8	88.5
GLM-4.6V-Flash (9B)	99.9	100.0	91.2	94.8	96.6	99.4
HealthGPT-L14 (14B)	88.9	98.4	100.0	100.0	92.5	98.3
HuatuoGPT-Vision (34B)	77.7	99.2	97.8	98.4	95.1	96.9
InternVL3 (38B)	100.0	100.0	100.0	100.0	99.9	99.9
Lingshu (32B)	100.0	100.0	99.3	99.3	100.0	100.0
Llama3.2-Vision (11B)	46.5	87.8	14.5	97.9	46.3	99.2
LLaVA-OneVision (72B)	100.0	100.0	100.0	100.0	99.8	100.0
MedDr (40B)	97.2	97.6	53.7	87.3	66.2	92.7
MedGemma (27B)	96.0	98.0	53.5	57.5	44.0	68.0
MedGemma (4B)	98.4	98.9	81.1	90.8	93.2	95.8
MiniMax-M3 (428B)	99.8	100.0	85.9	95.9	54.2	74.7
Qwen2.5-VL (32B)	100.0	100.0	16.4	99.9	9.2	99.6
Qwen2.5-VL (7B)	99.4	99.5	95.9	95.9	98.0	98.0
Qwen3-VL-Thinking (32B)	96.3	97.2	93.2	93.5	83.0	84.9
All responses	92.2	98.6	81.7	94.5	79.0	93.9

MedVision-V0 used 4 NVIDIA H200 (141 GB) GPUs.

Software. To support reproducibility, we publicly release all code used in this work. Dependency requirement files and Docker images are provided for every evaluated model. Reinforcement fine-tuning (RFT) was performed using verl v0.7.0.

I Details of SFT with CoT

This appendix details the construction of the SFT-CoT training data (Section 3): the prompt structure shared by all tasks, and the CoT instruction and answer template of each task.

Prompt Structure. Each training sample pairs a 512×512 image with a prompt assembled from a fixed skeleton (Box 1): a task description naming the measurement target, additional information, a format requirement specifying the final answer format, and a task-specific CoT instruction listing the reasoning steps to follow. For the T/L size and A/D measurement tasks, the additional information provides the image size in pixels and the physical pixel size (pixel spacing rescaled after image resizing), which the reasoning steps use to convert relative coordinates into physical measurements; the detection prompt omits this part because its answer is expressed in relative coordinates. Placeholders are written in angle brackets.

Box 1: Prompt skeleton (all tasks)

```
Task:
Given the input medical image[: <image description>], <task
      description>
Additional information:
The image size is <image_width> pixels (width) x <
     image_height> pixels (height).
The pixel size for this image is <pixel_width> <unit> (
     width) x <pixel_height> <unit> (height).
Format requirement:
<answer format requirement>
Reasoning steps:
<task-specific CoT instruction>
Follow the reasoning steps to get the final answer in the
     required format.
```

CoT Answer Construction. The SFT target is generated from the task's CoT answer template. Placeholders in the template are filled with ground truth values derived from the annotations: the relative coordinates of the bounding-box corners, ellipse-axis endpoints, or landmarks; the image and pixel sizes; and the intermediate and final measurement values. The resulting response walks through the declared reasoning steps, each wrapped in `<step-k-reasoning>` and `<step-k-answer>` tags inside the `<think>` block, and ends with the final answer in `<answer>` tags, so the supervised reasoning trace is consistent with the ground truth at every step. Boxes 2–5 give the task description, format requirement, CoT instruction, and CoT answer

template for the detection, T/L size estimation, distance measurement, and angle measurement tasks, respectively.

Box 2: Detection: prompt components and CoT answer template

Task description

```
return the coordinates of the lower-left and upper-right
      corner of the bounding box for the <label>.
```

Format requirement

```
The final answer must be enclosed within <answer> </answer>
       tags. The answer should be four decimal numbers
      separated by commas without any units or additional
      text. The first two numbers are the coordinates of
      the lower-left corner and the last two numbers are
      the coordinates of the upper-right corner of the
      bounding box. Use relative coordinates in the image
      space, where the origin is at the lower-left corner
      of the image. Relative coordinates should be values
      between 0 and 1, representing the relative positions
      in the image.
```

CoT instruction

```
Step 1: Identify the relative coordinates of the bounding
      box. The relative coordinates must be written as (x,
      y), where x is the relative position in width and y
      is the relative position in height.
Report the reasoning process and final answer within <think
      > </think> and <answer> </answer> tags, respectively.
       Inside <think> </think>, include reasoning and step
      results using <step-k-reasoning> </step-k-reasoning>
      and <step-k-answer> </step-k-answer> tags.
```

CoT answer template

```
<think>
<step-1-reasoning> I need to identify the relative
      coordinates of the bounding box of <label_name>. The
      relative coordinates must be written as (x, y), where
       x is the relative position in width and y is the
      relative position in height. I understand that the
      origin of the image space coordinate system is at the
       lower-left corner of the image. </step-1-reasoning>
<step-1-answer> The relative coordinates of the bounding
      box: (<coor0_w>, <coor0_h>), (<coor1_w>, <coor1_h>).
      </step-1-answer>
</think>
<answer> <coor0_w>,<coor0_h>,<coor1_w>,<coor1_h></answer>
```

Box 3: T/L size estimation: prompt components and CoT answer template

Task description

```
estimate the major and minor axis lengths of the ellipse
      enclosing the <label>, in <unit>.
```

Format requirement

```
The final answer must be enclosed within <answer> </answer>
       tags. The answer should consist of two decimal
      numbers separated by a comma, without units or extra
      text. The first number is the major axis length, and
      the second is the minor axis length.
```

CoT instruction

```
Step 1: Identify the major axis (the longest diameter) of
      the ellipse enclosing the target region. Find its two
       endpoints and record their relative coordinates in
      the format (x, y) = (relative position in width
      direction, relative position in height direction).
      Denote the endpoints as (x1_major, y1_major) and (
      x2_major, y2_major).
Step 2: Identify the minor axis (the shortest diameter) of
      the ellipse. Find its two endpoints and record their
      relative coordinates in the same (x, y) format.
      Denote them as (x1_minor, y1_minor) and (x2_minor,
      y2_minor).
Step 3: Given the pixel dimensions (pixel_width,
      pixel_height) and image size (image_width,
      image_height), compute the physical length of the
      major axis using: major_axis_length = sqrt(((x2_major
       - x1_major) * image_width * pixel_width)^2 + ((
      y2_major - y1_major) * image_height * pixel_height)
      ^2).
Step 4: Similarly, compute the physical length of the minor
       axis using: minor_axis_length = sqrt(((x2_minor -
      x1_minor) * image_width * pixel_width)^2 + ((y2_minor
       - y1_minor) * image_height * pixel_height)^2).
Report the reasoning process and final answer within <think
      > </think> and <answer> </answer> tags, respectively.
       Inside <think> </think>, include reasoning and step
      results using <step-k-reasoning> </step-k-reasoning>
      and <step-k-answer> </step-k-answer> tags.
```

CoT answer template

```
<think>
<step-1-reasoning> I need to identify the major axis of the
       ellipse enclosing the <label> and output its two
      endpoints. The relative coordinates must be written
      as (x1_major, y1_major), (x2_major, y2_major), where
      x is the relative position in width and y is the
      relative position in height. </step-1-reasoning>
<step-1-answer> The endpoints of the major axis: (<x1_major
      >, <y1_major>), (<x2_major>, <y2_major>). </step-1-
      answer>
<step-2-reasoning> Next, I must identify the minor axis of
      the ellipse enclosing the <label> and output its two
      endpoints in the same format: (x1_minor, y1_minor), (
      x2_minor, y2_minor). </step-2-reasoning>
<step-2-answer> The endpoints of the minor axis: (<x1_minor
      >, <y1_minor>), (<x2_minor>, <y2_minor>). </step-2-
      answer>
<step-3-reasoning> I now calculate the major axis length
      using the pixel dimensions (pixel_width, pixel_height
      ) = (<pixel_width>, <pixel_height>), the image size (
      image_width, image_height) = (<image_width>, <
      image_height>), and the distance formula:
      major_axis_length = sqrt(((x2_major - x1_major) *
      image_width * pixel_width)^2 + ((y2_major - y1_major)
       * image_height * pixel_height)^2) = sqrt(((<x2_major
      > - <x1_major>) * <image_width> * <pixel_width>)^2 +
      ((<y2_major> - <y1_major>) * <image_height> * <
      pixel_height>)^2) = <major_axis_length>. </step-3-
      reasoning>
<step-3-answer> The major axis length: <major_axis_length>.
       </step-3-answer>
<step-4-reasoning> I calculate the minor axis length using
      the same distance formula: minor_axis_length = sqrt
      (((x2_minor - x1_minor) * image_width * pixel_width)
      ^2 + ((y2_minor - y1_minor) * image_height *
      pixel_height)^2) = sqrt(((<x2_minor> - <x1_minor>) *
      <image_width> * <pixel_width>)^2 + ((<y2_minor> - <
      y1_minor>) * <image_height> * <pixel_height>)^2) = <
      minor_axis_length>. </step-4-reasoning>
<step-4-answer> The minor axis length: <minor_axis_length>.
       </step-4-answer>
</think>
<answer> (<major_axis_length>, <minor_axis_length>) </
      answer>
```

Box 4: Distance measurement: prompt components and CoT answer template

Task description

```
estimate the distance of <measurement name> in <unit>,
      which is the distance between 2 landmark points: (
      landmark 1) <landmark 1>, (landmark 2) <landmark 2>.
```

Format requirement

```
The final answer must be enclosed within <answer> </answer>
       tags. The answer should be a single decimal number
      without units or extra text.
```

CoT instruction

```
Step 1: Identify the landmark 1 and record its relative
      coordinates in the format (x, y) = (relative position
       in width direction, relative position in height
      direction). Denote the coordinates as (x1, y1).
Step 2: Identify the landmark 2 and record its relative
      coordinates in the same (x, y) format. Denote the
      coordinates as (x2, y2).
Step 3: Given the pixel dimensions (pixel_width,
      pixel_height) and image size (image_width,
      image_height), compute the physical distance between
      the two landmarks using: distance = sqrt(((x2 - x1) *
       image_width * pixel_width)^2 + ((y2 - y1) *
      image_height * pixel_height)^2).
```

```
Report the reasoning process and final answer within <think
      > </think> and <answer> </answer> tags, respectively.
       Inside <think> </think>, include reasoning and step
      results using <step-k-reasoning> </step-k-reasoning>
      and <step-k-answer> </step-k-answer> tags.
```

CoT answer template

```
<think>
<step-1-reasoning> I need to identify <landmark 1> and
      output its relative coordinates. The relative
      coordinates must be written as (x1, y1), where x is
      the relative position in width and y is the relative
      position in height. </step-1-reasoning>
<step-1-answer> The relative coordinates of <landmark 1>:
      (<x1>, <y1>). </step-1-answer>
<step-2-reasoning> Next, I must identify <landmark 2> and
      output its relative coordinates in the same format: (
      x2, y2). </step-2-reasoning>
<step-2-answer> The relative coordinates of <landmark 2>:
      (<x2>, <y2>). </step-2-answer>
<step-3-reasoning> I now calculate the distance between the
       two landmarks using the pixel dimensions (
      pixel_width, pixel_height) = (<pixel_width>, <
      pixel_height>), the image size (image_width,
      image_height) = (<image_width>, <image_height>), and
      the distance formula: distance = sqrt(((x2 - x1) *
      image_width * pixel_width)^2 + ((y2 - y1) *
      image_height * pixel_height)^2) = sqrt(((<x2> - <x1>)
       * <image_width> * <pixel_width>)^2 + ((<y2> - <y1>)
      * <image_height> * <pixel_height>)^2) = <distance>.
      </step-3-reasoning>
<step-3-answer> The distance: <distance>. </step-3-answer>
</think>
<answer> <distance> </answer>
```

Box 5: Angle measurement: prompt components and CoT answer template

Task description

```
estimate the angle of <measurement name> in <unit>, which
      is the angle between 2 lines: (line 1) the line
      connecting <landmark 1> and <landmark 2>, (line 2)
      the line connecting <landmark 3> and <landmark 4>.
```

Format requirement

```
The final answer must be enclosed within <answer> </answer>
       tags. The answer should be a single decimal number
      without units or extra text.
```

CoT instruction

```
Step 1: Identify line 1 and record the relative coordinates
       of its two endpoints in the format (x, y) = (
      relative position in width direction, relative
      position in height direction). Denote the endpoints
      as (x1_line1, y1_line1) and (x2_line1, y2_line1).
Step 2: Identify line 2 and record the relative coordinates
       of its two endpoints in the same (x, y) format.
      Denote them as (x1_line2, y1_line2) and (x2_line2,
      y2_line2).
Step 3: Given the pixel dimensions (pixel_width,
      pixel_height) and image size (image_width,
      image_height), compute the angle between the two
      lines using the formula: angle = arccos(|A · B| / (||
      A|| ||B||)), where A and B are the vectors of the two
       lines computed from the physical coordinates of
      their endpoints. A = ((x2_line1 - x1_line1) *
      image_width * pixel_width, (y2_line1 - y1_line1) *
      image_height * pixel_height) and B = ((x2_line2 -
      x1_line2) * image_width * pixel_width, (y2_line2 -
      y1_line2) * image_height * pixel_height). Denote A=(
      Ax, Ay) and B=(Bx, By). Then, angle = arccos(|Ax*Bx +
       Ay*By| / (sqrt(Ax^2 + Ay^2) * sqrt(Bx^2 + By^2))).
Report the reasoning process and final answer within <think
      > </think> and <answer> </answer> tags, respectively.
       Inside <think> </think>, include reasoning and step
      results using <step-k-reasoning> </step-k-reasoning>
      and <step-k-answer> </step-k-answer> tags.
```

CoT answer template

```
<think>
<step-1-reasoning> I need to identify the relative
      coordinates of <landmark 1> and <landmark 2> that
      define line 1. The relative coordinates must be
      written as (x1_line1, y1_line1), (x2_line1, y2_line1)
      , where x is the relative position in width and y is
      the relative position in height. </step-1-reasoning>
<step-1-answer> The relative coordinates of <landmark 1>
      and <landmark 2>: (<x1_line1>, <y1_line1>), (<
      x2_line1>, <y2_line1>). </step-1-answer>
<step-2-reasoning> Next, I must identify the relative
      coordinates of <landmark 3> and <landmark 4> that
      define line 2, in the same format: (x1_line2,
      y1_line2), (x2_line2, y2_line2). </step-2-reasoning>
<step-2-answer> The relative coordinates of <landmark 3>
      and <landmark 4>: (<x1_line2>, <y1_line2>), (<
      x2_line2>, <y2_line2>). </step-2-answer>
<step-3-reasoning> I now calculate the angle between the
      two lines using the pixel dimensions (pixel_width,
      pixel_height) = (<pixel_width>, <pixel_height>), the
      image size (image_width, image_height) = (<
      image_width>, <image_height>), and the angle formula:
       angle = arccos(|A · B| / (||A|| ||B||)), where A and
       B are the vectors of the two lines computed from the
       physical coordinates of their endpoints. A = ((
      x2_line1 - x1_line1) * image_width * pixel_width, (
      y2_line1 - y1_line1) * image_height * pixel_height) =
       ( (<x2_line1> - <x1_line1>) * <image_width> * <
      pixel_width>, (<y2_line1> - <y1_line1>) * <
      image_height> * <pixel_height>) = (<Ax>, <Ay>). B =
      ((x2_line2 - x1_line2) * image_width * pixel_width, (
      y2_line2 - y1_line2) * image_height * pixel_height) =
       ( (<x2_line2> - <x1_line2>) * <image_width> * <
      pixel_width>, (<y2_line2> - <y1_line2>) * <
      image_height> * <pixel_height>) = (<Bx>, <By>).
      Denote A=(Ax, Ay) and B=(Bx, By). Then, angle =
      arccos(|Ax*Bx + Ay*By| / (sqrt(Ax^2 + Ay^2) * sqrt(Bx
      ^2 + By^2))) = arccos(|<Ax>*<Bx> + <Ay>*<By>| / (sqrt
      (<Ax>^2 + <Ay>^2) * sqrt(<Bx>^2 + <By>^2))) = <angle>
       = <angle_degree> degrees. </step-3-reasoning>
<step-3-answer> The angle: <angle_degree>. </step-3-answer>
</think>
<answer> <angle_degree> </answer>
```

J Details of RFT via GRPO

This appendix details the reward used in the RFT stage (Section 3) and the GRPO training configuration. Algorithm 2 summarizes the reward computation for one rollout.

Reward Components. Each rollout is scored by up to three components, each in $[0, 1]$. The two accuracy-based components share one error-to-reward mapping $\rho(e) = \exp(-e)$: an exact prediction earns a reward of 1, the reward decays smoothly as the error grows, and an unparseable value earns 0. Because predicted coordinates are relative positions in $[0, 1]$ and measurement errors are relative errors, all errors are dimensionless and of comparable scale across tasks. ***(i)*** The ***format reward*** encourages responses that follow the required output structure, so that the reasoning steps and the final answer remain machine-parseable. For the T/L and A/D tasks, it combines a soft reasoning-structure score with a binary check of the final answer format, weighted 0.8 and 0.2, respectively; the reasoning-structure score grants partial credit for the presence of the `<think>` block, the per-step reasoning and answer tags (i.e., `<step-k-reasoning>` and `<step-k-answer>`), well-formed step values, and the correct step order. For detection, only the bi-

nary answer-format check is used. ***(ii)*** The ***process reward*** supervises the reasoning chain itself, granting credit for accurate intermediate estimates rather than only for the final answer. Each CoT step declares one intermediate result: localization steps predict relative coordinates (e.g., landmarks or axis endpoints), and measurement steps predict an intermediate measurement value. The value in each `<step-k-answer>` block is parsed and compared with its ground truth: localization steps are scored by the point-displacement error, and measurement steps by the relative error. Concretely, the localization error of a point is its Euclidean displacement from the ground truth point in relative coordinates, divided by $\sqrt{2}$, the diagonal of the unit square, so that it lies in $[0, 1]$; a step declaring multiple points is scored by its worst-localized point. Each step error is mapped through ρ, and the process reward is the mean of the per-step rewards (4 steps for T/L, 3 for A/D). ***(iii)*** The ***answer reward*** scores the final answer: the mean relative error of the parsed values for T/L and A/D, and the overlap error $e = (1 - \text{CIoU})/2$ for detection. CIoU (Zheng et al., 2020) augments IoU with a normalized center-distance penalty and an aspect-ratio consistency penalty:

$$\text{CIoU} = \text{IoU} - \frac{d^2}{c^2} - \alpha v, \tag{13}$$

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w}{h} - \arctan\frac{\hat{w}}{\hat{h}}\right)^2, \tag{14}$$

where d is the distance between the centers of the predicted and ground truth boxes, c is the diagonal length of the smallest box enclosing both, (w, h) and $(\hat{w}, \hat{h})$ are the ground truth and predicted box dimensions, and $\alpha = v/\big((1 - \text{IoU}) + v\big)$ down-weights the aspect-ratio penalty when the overlap is poor. Since $\text{CIoU} \in [-1, 1]$, the overlap error lies in $[0, 1]$, and unlike plain IoU, which is uniformly zero for any non-overlapping prediction, it provides a graded learning signal everywhere.

Reward Combination. For the T/L and A/D tasks, the final reward is $r = r_{\text{format}} + r_{\text{process}} \cdot r_{\text{answer}}$. The multiplicative form makes the final-answer credit conditional on a verifiable chain of intermediate estimates: a rollout cannot collect the full answer reward from a final number that is inconsistent with its own reasoning trace, and accurate intermediate localizations pay off only through an accurate final measurement. The detection answer is a single localization step, so no process reward is defined and the reward reduces to $r = r_{\text{format}} + r_{\text{answer}}$.

Training Configuration. RFT uses the GRPO implementation in verl (Sheng et al., 2025) with full-parameter fine-tuning, and the three sequential stages (A/D $\rightarrow$ T/L $\rightarrow$ detection) share one configuration, each stage initialized from the checkpoint of the previous one: 8 rollouts per prompt (the GRPO group size), a training batch of 256 prompts per step, with policy updates performed on mini-batches of 128 prompts, a constant learning rate of 3×10^{-6}, KL regularization toward the reference policy in the loss with coefficient 0.01 (low-variance estimator), and maximum prompt and response lengths of 4,096 tokens. Rollouts are generated with vLLM on 4 H200 GPUs (Appendix H). The multi-task RFT ablation (Table 16) keeps this configuration in a single stage over the task mixture, with two additions: temperature-scaled task mixing (Appendix K) that flattens the task ratio toward balance (approximately 42%, 29%, and 29% of draws per epoch for detection, T/L, and A/D, respectively), and an epoch-level curriculum (Appendix L) that progressively filters out solved samples.

K Temperature-Scaled Task Mixing

The multi-task training mixture is highly imbalanced: 110K detection samples against 5.5K T/L and 5.5K A/D samples (Section 3). Sampling proportionally to these counts would devote about 91% of all draws to detection, whereas uniform task probabilities would ignore the size of the detection set entirely. The SFT stage and the multi-task RFT ablation therefore rebalance the mixture with temperature-scaled task mixing, which interpolates between these two extremes.

Mechanism. Samples are grouped into three tasks (angle and distance measurements are merged into A/D). For a task t with N_t samples, the task-level sampling probability is

$$P(t) = \frac{N_t^{1/T}}{\sum_u N_u^{1/T}}, \tag{15}$$

where $T > 0$ is the mixing temperature: $T{=}1$ recovers the natural task proportions, and increasing T flattens the probabilities toward uniform. Each sample of task t receives the weight $P(t)/N_t$, which is uniform within a task and reproduces the task-level probabilities exactly under weighted random sampling. One epoch draws as many samples as the training set contains, *with replacement*, so a minority-task sample can be drawn more than once

Algorithm 2 Reward computation for one RFT rollout.

Input: rollout text R; task t; ground truth answer $\mathbf{y}$; ground truth step values $\mathbf{s}_1, \ldots, \mathbf{s}_n$ (T/L: n=4; A/D: n=3; detection: n=0); error-to-reward map $\rho(e) = \exp(-e)$

Process:

$r_{\text{format}} \leftarrow \textsc{FormatScore}(R)$ ▷ tag-structure adherence, in $[0, 1]$
for $k = 1, \ldots, n$ **do** ▷ process reward from CoT steps
  $\hat{\mathbf{s}}_k \leftarrow \textsc{ParseStep}(R, k)$ ▷ value in `<step-k-answer>`
  if $\hat{\mathbf{s}}_k = \varnothing$ **then**
    $r_k \leftarrow 0$
  else
    $e_k \leftarrow \textsc{StepError}(\hat{\mathbf{s}}_k, \mathbf{s}_k)$ ▷ localization step: worst-case point displacement; measurement step: relative error
    $r_k \leftarrow \rho(e_k)$
  end if
end for
$r_{\text{process}} \leftarrow \frac{1}{n} \sum_{k=1}^{n} r_k$
$\hat{\mathbf{y}} \leftarrow \textsc{ParseAnswer}(R)$ ▷ values in `<answer>`
if $\hat{\mathbf{y}} = \varnothing$ **then**
  $r_{\text{answer}} \leftarrow 0$
else
  if t is detection **then**
    $e_a \leftarrow (1 - \text{CIoU}(\hat{\mathbf{y}}, \mathbf{y}))/2$ ▷ overlap error
  else
    $e_a \leftarrow \frac{1}{d} \sum_{j=1}^{d} |\hat{y}_j - y_j|/y_j$ ▷ mean relative error; d: number of answer values
  end if
  $r_{\text{answer}} \leftarrow \rho(e_a)$
end if
if t is detection **then**
  $r \leftarrow r_{\text{format}} + r_{\text{answer}}$ ▷ no intermediate steps
else
  $r \leftarrow r_{\text{format}} + r_{\text{process}} \cdot r_{\text{answer}}$
end if

Output: scalar reward r

per epoch while a majority-task sample may be skipped; the sampler is seeded for reproducibility.

Instantiations in SFT and RFT. The SFT stage uses T=5, giving task probabilities of approximately 47.7% for detection and 26.2% each for T/L and A/D (natural proportions: 90.9%, 4.5%, and 4.5%). The multi-task RFT ablation uses T=8, giving approximately 42.1% for detection and 29.0% each for T/L and A/D. When composed with the epoch-level curriculum (Appendix L), the task counts and sample weights are recomputed over the active set (S_t in Algorithm 3) at every epoch, and the sampler falls back to unweighted sampling if fewer than two distinct tasks remain active.

L Epoch-Level Curriculum Learning for Multi-Task RFT

In GRPO, the advantage is normalized within the group of rollouts of one prompt. Once the policy reliably solves a prompt, all rollouts in its group receive nearly identical rewards, the within-group variance collapses, and the advantage approaches zero: the sample contributes almost no gradient while still consuming a full group of rollouts each time it is drawn. The multi-task RFT run therefore adopts an epoch-level curriculum that progressively removes solved samples from training, concentrating rollout compute on the samples the policy still fails. Unlike the classic easy-to-hard curriculum (Bengio et al., 2009), the schedule emphasizes hard examples, in line with dynamic-sampling practice in reinforcement learning of large language models (Yu et al., 2025). Algorithm 3 summarizes the procedure.

Sample Pools and Evidence. For each task, samples are partitioned into two pools: a ***hard pool*** H_t of samples under active training and an ***easy pool*** E_t of samples classified as solved (notation as in Algorithm 3). The first epoch trains on the full dataset. Within each epoch, the mean reward and the mean final-answer error of every drawn sample are computed over all of its rollouts. Averaging over rollouts estimates the policy's expected performance on the sample: a sample counts as solved only when the policy solves it reliably, not when a single rollout succeeds by chance. At the end of the epoch, these epoch-level means update a persistent per-sample record comprising exponential moving averages (EMA, weight α=0.4) of the reward and of the answer error ($\bar{r}_i$ and $\bar{e}_i$), together with the numbers of consecutive observed epochs in which the sample's error remained below, or above, the task-specific error threshold g_t (c_i^+ and c_i^-): the former is required for promotion to the easy pool, and the latter for demotion back to the hard pool. The threshold g_t matches the benchmark's own success criteria: MRE < 0.1 for the T/L and A/D tasks,

and overlap error $(1 - \text{CIoU})/2 < 0.25$ for detection (implying IoU > 0.5). This record makes pool assignment robust to the stochasticity of a single epoch's rollouts: although one qualifying epoch is formally sufficient for promotion in our configuration (ν^+=1), an unfavorable epoch lowers the EMA, and several favorable epochs are then required before the promotion criterion is met.

Pool Updates. At the end of each epoch, the following updates are applied to each task. ***(i) Promotion.*** The hard-pool samples drawn during the epoch are ranked by EMA reward $\bar{r}_i$, and at most the top 20% (f=0.2) of the task's current training set is eligible for promotion; of these, only samples whose EMA answer error $\bar{e}_i$ falls below the threshold g_t and whose passing count c_i^+ reaches the required number of consecutive epochs ν^+ are moved to the easy pool. ***(ii) Retention mix-in.*** To mitigate catastrophic forgetting of solved samples, the *most recently promoted easy samples* are mixed back into training (the mix-in set M_t); their share s_t increases with the task's solved fraction p_t, up to a maximum composition of 30% easy (m=0.3) and 70% hard samples once half of the task is solved (p^*=0.5). The promotion fraction f and the maximum mix-in share m, like the other constants in this section, are tunable hyperparameters; we report the values used in our experiments. ***(iii) Auditing and demotion.*** A rotating audit subset A_t (5% of the hard pool per epoch, a=0.05) re-evaluates the *least recently audited easy samples*; a mixed-in or audited sample (in $M_t \cup A_t$) whose EMA error $\bar{e}_i$ exceeds λ=1.5 times the task's error threshold g_t and whose failing count c_i^- reaches the required number of consecutive epochs ν^- is returned to the hard pool, which corrects erroneous promotions and detected forgetting. The margin λ introduces hysteresis, preventing borderline samples from oscillating between the two pools. Mix-in and audit are complementary but distinct: the mix-in keeps the *most recently promoted samples* under continued training to prevent forgetting, whereas the audit rotates through the *least recently audited samples* so that every easy sample is eventually re-evaluated. Neither changes pool membership by itself: mixed-in and audited samples remain in the easy pool E_t while temporarily rejoining the active set S_t, and since they are the only easy samples with fresh rollouts, they are the only demotion candidates, demoted only when the above criterion is met. ***(iv) Per-task minimum.*** At least 10% (ϕ=0.1) of each task's original samples remain in training, so that a nearly solved task is never excluded entirely: while the task remains solved, its rollout groups yield near-zero advantages and thus contribute little gradient, and if performance on the task degrades, the demotion rule returns its samples to training.

Composition with Task Balancing. The temperature-scaled task-mixing distribution (Appendix K) is recomputed over the active set S_t at every epoch, so task-level rebalancing adapts as the large detection pool shrinks relative to the T/L and A/D pools. Epochs shorten as the pools shrink.

M MedVision-V0 Overview

Qualitative Results. Figures 7, 8, and 9 show MedVision-V0 responses for the detection, T/L size and A/D tasks, respectively. In each figure, the left column gives the full prompt, the model response, and the ground truth, prediction and metrics for one example; the middle column shows the input image and the same image with the ground truth and prediction overlaid; and the right panel shows random samples of model predictions, grouped by dataset.

Quantitative Results. Figure 10 summarizes the performance gains from each training stage (base $\rightarrow$ SFT $\rightarrow$ SFT-RFT): the average performance metric in each label group is presented for the base, SFT, and SFT-RFT models, and the distribution of metric values in each group for the SFT-RFT model is visualized as a violin plot. Table 14 reports the full quantitative ablation across all three tasks.

Base Model Ablation. Table 15 compares the base $\rightarrow$ SFT progression across three base models (Qwen2.5-VL, Gemma4, and MedGemma), all of which improve substantially across the three tasks.

Multi-Task RFT. The RFT stage of MedVision-V0 trains on the three tasks sequentially (A/D $\rightarrow$ T/L $\rightarrow$ detection; Section 3), and this task order is a design choice. To examine its impact, we train an ablation variant that starts from the same SFT model and replaces the three sequential stages with a single RFT stage on a mixture of all task data, using a task-balancing sampler and an epoch-level curriculum (Appendix L). Table 16 compares the two variants on the detection, T/L size estimation, and A/D measurement benchmarks. The multi-task variant is nearly identical to the sequential model on detection (F1 78.2% vs. 79.1% on anatomy; 47.4% vs. 46.9% on T/L), T/L size estimation (MRE 26.7% vs. 26.0%), and distance measurement (MRE 6.6% vs. 6.4%), and performs better

Algorithm 3 Epoch-level curriculum for multi-task RFT.

Input: hard pool H_t (all N_t samples of task t); easy pool $E_t \leftarrow \varnothing$; mix-in and audit subsets $M_t, A_t \leftarrow \varnothing$; per-sample EMAs $\bar{r}_i, \bar{e}_i$ and counters $c_i^+, c_i^- \leftarrow 0$; error threshold g_t (T/L, A/D: MRE 0.1; detection: overlap error 0.25); promotion fraction f=0.2; maximum mix-in share m=0.3, reached at solved fraction p^*=0.5; hysteresis margin λ=1.5; audit fraction a=0.05; per-task minimum fraction ϕ=0.1; EMA weight α=0.4; required consecutive passing (failing) epochs ν^+=1 (ν^-=1)

Process:

$S_t \leftarrow H_t$ for every task t
for epoch $k = 1, 2, \ldots$ **do**
  run GRPO on $\bigcup_t S_t$; compute the mean reward $r_i^{(k)}$ and mean answer error $e_i^{(k)}$ of every drawn sample i over its rollouts, where $D_t^{(k)}$ denotes the samples of task t drawn in epoch k
  for each task t **do**
    for $i \in D_t^{(k)}$ **do**
      $\bar{r}_i \leftarrow \alpha\, r_i^{(k)} + (1-\alpha)\, \bar{r}_i; \quad \bar{e}_i \leftarrow \alpha\, e_i^{(k)} + (1-\alpha)\, \bar{e}_i$ ▷ update the reward and error EMAs
      $c_i^+ \leftarrow \mathbb{I}[e_i^{(k)} < g_t] \cdot (c_i^+ + 1); \quad c_i^- \leftarrow \mathbb{I}[e_i^{(k)} \geq g_t] \cdot (c_i^- + 1)$ ▷ count consecutive passing/failing epochs
    end for
    $T_t \leftarrow$ the top $f \cdot |S_t|$ samples of $D_t^{(k)} \cap H_t$ ranked by EMA reward $\bar{r}_i$ ▷ promotion candidates: at most a fraction f of the active set
    $P_t \leftarrow \{\, i \in T_t : \bar{e}_i < g_t,\ c_i^+ \geq \nu^+ \,\}$ ▷ promote: error below threshold for ν^+ consecutive epochs
    $Q_t \leftarrow \{\, i \in M_t \cup A_t : \bar{e}_i \geq \lambda\, g_t,\ c_i^- \geq \nu^- \,\}$ ▷ demote: error regressed beyond the hysteresis margin
    $E_t \leftarrow (E_t \cup P_t) \setminus Q_t; \quad H_t \leftarrow (H_t \setminus P_t) \cup Q_t$ ▷ apply promotions and demotions to the pools
    $p_t \leftarrow |E_t|/N_t; \quad s_t \leftarrow m \cdot \min(1,\ p_t/p^*)$ ▷ solved fraction; mix-in share
    $M_t \leftarrow$ the $\mathrm{round}\big(s_t\,|H_t|/(1-s_t)\big)$ most recently promoted samples of E_t ▷ mix solved samples back in against forgetting
    $A_t \leftarrow$ the $\mathrm{round}(a\,|H_t|)$ least recently audited samples of E_t ▷ rotating audit of the easy pool
    $S_t \leftarrow H_t \cup M_t \cup A_t$, extended with the most recently promoted samples of E_t until $|S_t| \geq \mathrm{round}(\phi\, N_t)$ ▷ enforce the per-task minimum active size
  end for
  recompute the temperature-scaled task-mixing distribution over $\bigcup_t S_t$
end for

Output: per-epoch active sets S_t that concentrate rollout compute on unsolved samples while retaining solved samples for re-evaluation

on angle measurement (MAE 3.7 vs. 4.7 degrees; MRE 41.9% vs. 52.1%). As A/D is optimized in the first stage of the sequential schedule, the sequential model's weaker angle measurement may be attributable to catastrophic forgetting during the subsequent T/L and detection stages. These results indicate that multi-task RFT is a viable alternative to the multi-stage strategy.

RFT Reward Design. For the T/L and A/D tasks, the final reward combines the process and answer rewards multiplicatively, $r = r_{\text{format}} + r_{\text{process}} \cdot r_{\text{answer}}$, making the final-answer credit conditional on accurate intermediate estimates (Appendix J). To assess this design choice, we train an ablation variant of the multi-task RFT model that instead combines the two rewards additively ($r = r_{\text{format}} + r_{\text{process}} + r_{\text{answer}}$), with all other training settings unchanged. As the detection task defines no process reward, the comparison concerns the T/L size estimation and A/D measurement benchmarks (Table 17). The two designs perform comparably on T/L size estimation (MRE 26.7% vs. 26.9%) and distance measurement (MRE 6.6% vs. 6.7%), while the multiplicative design performs better on angle measurement (MRE 41.9% vs. 48.1%). These results support the multiplicative composition adopted in MedVision-V0.

OOD Ablation. Table 18 presents out-of-distribution generalization results. We evaluate two OOD settings: ***(i) Plane-OOD***, where test samples involve imaging planes not seen during training for a given anatomical target (e.g., a structure trained on axial CT slices evaluated on coronal or sagittal views); and ***(ii) Target-OOD***, where test samples involve anatomical targets (structures or tumors) entirely absent from the training label set; Figure 12 lists the in-distribution and Target-OOD targets of the detection and T/L size estimation tasks. Figure 11 illustrates the Plane-OOD setting in the T/L size estimation task: each grid position shows the same subject and target sliced along the axial (in-distribution) plane and the coronal and sagittal (OOD) planes, revealing the markedly different ***anatomical context, target appearance, and physical extent*** that the model must generalize across. Across both OOD settings, SFT-RFT consistently outperforms the SFT model, indicating that reinforcement learning fosters stronger generalization beyond memorized training distributions. This observation is consistent with the finding that RL generalizes better than SFT (Chu et al., 2025). MedVision-V0 nonetheless trails its in-distribution performance: anatomy detection F1 drops from 79.1% (Table 5) to 49.5% (plane-OOD) and 36.0% (target-OOD), T/L detection F1 from 46.9% to 36.6% and 45.5%, and T/L size MRE rises from 26.0% (Table 6) to 36.4% and 34.7%.

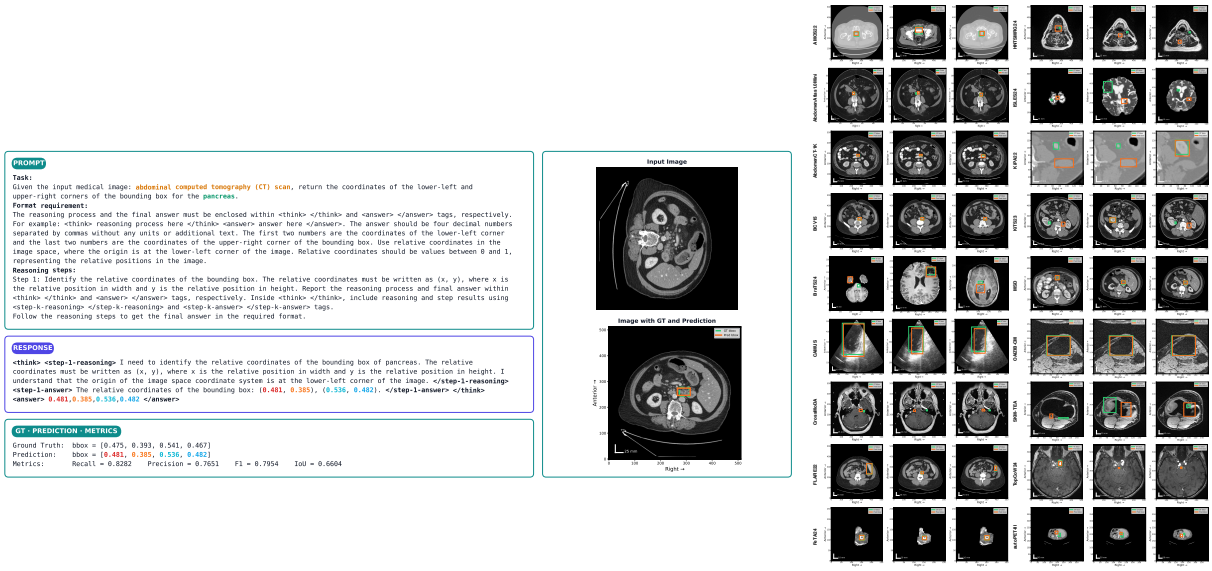


Figure 7: Model responses in the detection task. Left: prompt, model response, and the ground truth, prediction and metrics of one example. Middle: the input image, and the image with the ground truth and predicted bounding boxes. Right: random samples of model predictions, grouped by dataset.

N Per-Model Performance Profiles

Figure 13 provides a general overview of model performance against model release date on the three benchmark task families. Figures 14 and 15 present per-model performance profiles for the 17 evaluated VLMs: each row shows one model's results as six radar charts (detection recall, precision, and F1; angle and distance MRE; T/L size MRE), with the model's per-sample metric distributions overlaid as violin and box plots on each spoke.

O Details of Detection Performance

Label-Level Detection Performance. Figure 16 shows the label-level detection performance of VLMs at each box-image-ratio group. Box-image-ratio is defined as the area ratio of the bounding box to the image. Labels are grouped into different box-image-ratio ranges. We visualized the composition of box-image-ratio groups for each label at the bottom of Figure 16. For each box-image-ratio group and label combination, we visualized recall, precision, and F1 score. Figure 16 is a detailed map of VLM performance at various target sizes and labels.

Effect of Target Size. Figure 16 shows a trend of increasing detection performance with larger target sizes (box-image-ratio). To further illustrate this effect, we grouped all detection targets into different box-image-ratio groups, and calculated the overall detection performance for each group. As shown in Figure 17, precision and recall are positively correlated with relative target size (box-image-ratio). As a reference, the figure includes a random-guessing baseline (***Random***): for each test sample, we simulate 100 random boxes whose width and height are drawn uniformly at random (from a 3-pixel minimum up to the full image extent) and whose position is drawn uniformly among all placements fully inside the image; per-sample metrics are averaged over the 100 simulations and then aggregated per box-image-ratio group in the same way as for the evaluated models. The baseline's rising precision and F1 reflect that larger targets are easier to overlap by chance. MedVision-V0 consistently outperforms all off-the-shelf models across all target sizes.

P Clinical Decision Agreement Analysis

This appendix details the clinical decision agreement (CDA) analysis summarized in Section 4.5. CDA re-scores the parsed benchmark outputs of all evaluated models: each predicted measurement is converted into a clinical category through a published cutoff table, and the predicted category is compared with the category derived from the ground truth measurement. Because both sides pass through the same cutoff table, any disagreement is attributable to measurement error alone.

CDA Tasks. We evaluate three CDA tasks. ***(i–ii) Maxillary and mandibular position:*** the antero-posterior position of the maxilla and the mandible

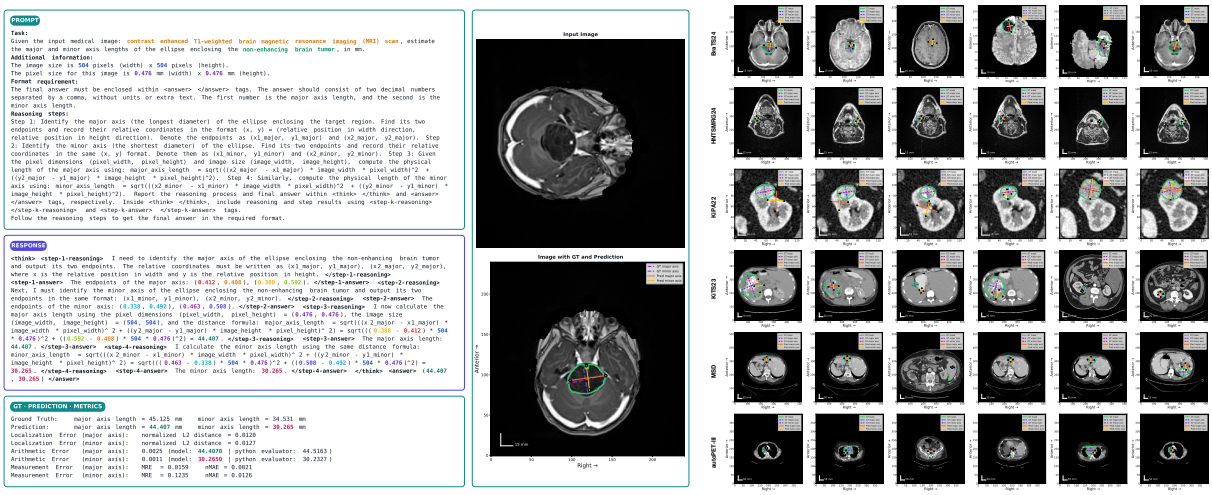


Figure 8: Model responses in the T/L size task. Left: prompt, model response, and the ground truth, prediction and metrics of one example. Middle: the input image, and the image with the ground truth and predicted major and minor axes. Right: random samples of model predictions, grouped by dataset.

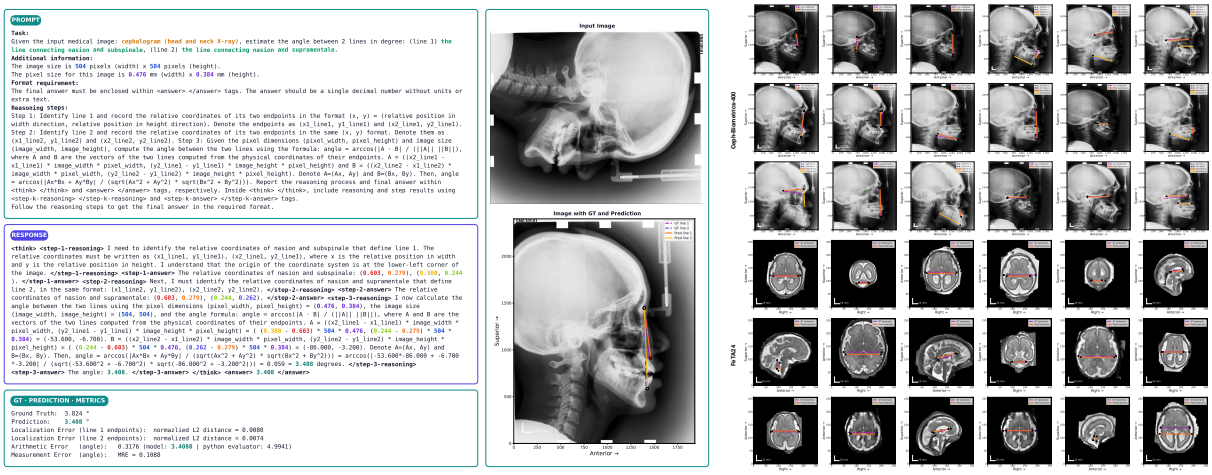


Figure 9: Model responses in the A/D task. Left: prompt, model response, and the ground truth, prediction and metrics of one example. Middle: the input image, and the image with the ground truth and predicted landmarks and measurement. Right: random samples of model predictions, grouped by dataset.

of the Ceph-Bio-400 samples, revealed by the SNA (sella–nasion–A point) and SNB (sella–nasion–B point) angles, respectively. Each is a three-class task: an angle within the Steiner norm band (82 ± 2 degrees for SNA, 80 ± 2 degrees for SNB) (Steiner, 1953) indicates a normal position, an angle below the band a retrusive position, and an angle above the band a protrusive position. ***(iii) AJCC kidney tumor staging:*** the kidney tumor category from tumor size, a four-class task (T staging: T1a $\leq$ 40 mm, T1b >40–70 mm, T2a >70–100 mm, T2b $>$ 100 mm) (American Joint Committee on Cancer, 2017), applied to the per-slice T/L size predictions on the KiPA22 and KiTS23 datasets. Tumor size measured on cross-sectional imaging is the primary criterion for the clinical T staging of organ-confined kidney tumors (American Joint Committee on Cancer, 2017; Elkassem et al., 2021), so these cutoffs constitute clinically established decision boundaries for size measurement. The benchmark measures the tumor's major axis on each annotated 2D slice, a standard clinical practice consistent with the RECIST guideline (v1.1) for tumor response assessment (Eisenhauer et al., 2009). The kidney tumor staging task therefore compares each slice-level prediction with the category of the ground truth measurement, evaluating whether the model's measurement error is large enough to change the staging category.

Metrics. Accuracy (Acc) is the fraction of parsed

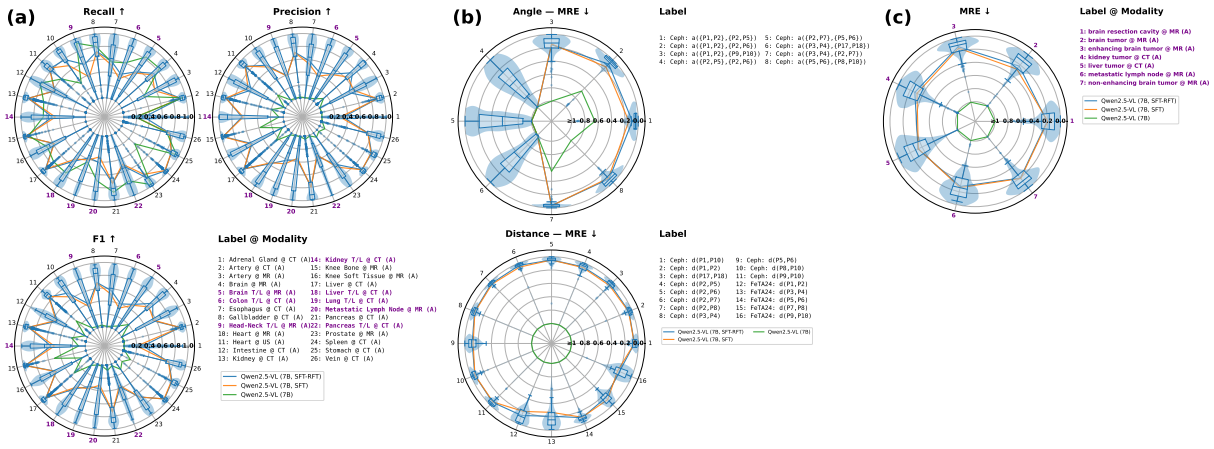


Figure 10: Effectiveness of the SFT-RFT training strategy on the three benchmark tasks. Each panel shows the average performance metric of each label group for the base (Qwen2.5-VL-7B), SFT, and SFT-RFT (MedVision-V0) models; violin plots show the distribution of metric values within each group for the SFT-RFT model.

Table 14: Training progression of MedVision-V0: base (Qwen2.5-VL-7B) → SFT → SFT-RFT (MedVision-V0) across all three tasks. Parentheses give the change from the preceding stage (SFT vs. base; SFT-RFT vs. SFT; green: improvement; red: decline; gray: no change).

Model	R ↑	P ↑	F1 ↑	IoU ↑	SR ↑	$\text{IoU}_{>0.5}$ ↑	$\text{IoU}_{0.5:0.05:0.95}$ ↑
Detection: Anatomy (18 regions, 13.4K samples)							
base	69.7	12.2	16.7	11.3	99.4	5.6	1.7
SFT	79.8 (+10.1)	80.5 (+68.3)	78.3 (+61.6)	71.2 (+59.9)	100.0 (+0.6)	79.3 (+73.7)	59.0 (+57.3)
SFT-RFT	81.3 (+1.5)	80.4 (-0.1)	79.1 (+0.8)	72.0 (+0.8)	100.0 (0.0)	80.1 (+0.8)	59.9 (+0.9)
Detection: Tumor/Lesion (8 regions, 8.5K samples)							
base	77.4	3.8	6.5	3.6	99.6	0.0	0.0
SFT	47.9 (-29.5)	49.3 (+45.5)	44.1 (+37.6)	35.4 (+31.8)	100.0 (+0.4)	37.0 (+37.0)	17.9 (+17.9)
SFT-RFT	52.4 (+4.5)	50.5 (+1.2)	46.9 (+2.8)	38.2 (+2.8)	100.0 (0.0)	40.7 (+3.7)	21.1 (+3.2)

Model	MAE ↓	MRE ↓	SR ↑	$\text{MRE}_{<0.1}$ ↑
Tumor/Lesion Size (2K samples)				
base	2897.2	7680.9	95.4	0.7
SFT	11.3 (-2885.9)	28.6 (-7652.3)	100.0 (+4.6)	20.6 (+19.9)
SFT-RFT	10.5 (-0.8)	26.0 (-2.6)	100.0 (0.0)	23.5 (+2.9)

Model	MAE ↓	MRE ↓	SR ↑	$\text{MRE}_{<0.1}$ ↑
Distance (Ceph+FeTA, 1,100 samples)				
base	63603.8	96542.0	98.3	0.5
SFT	4.2 (-63599.6)	7.5 (-96534.5)	100.0 (+1.7)	77.7 (+77.2)
SFT-RFT	3.6 (-0.6)	6.4 (-1.1)	100.0 (0.0)	81.4 (+3.7)
Angle (Ceph, 960 samples)				
base	48.0	724.9	97.6	2.0
SFT	5.8 (-42.2)	53.3 (-671.6)	100.0 (+2.4)	48.8 (+46.8)
SFT-RFT	4.7 (-1.1)	52.1 (-1.2)	99.9 (-0.1)	52.0 (+3.2)

predictions whose category matches the reference, and success rate (SR) is the fraction of predictions from which a measurement could be parsed. Chance-corrected agreement is measured with Cohen's κ and quadratic-weighted κ_w. Cohen's κ measures exact-category agreement beyond chance:

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{16}$$

where p_o is the observed fraction of samples whose predicted and reference categories match, and p_e is the matching fraction expected by chance from the marginal distributions. All three CDA tasks have ordered categories, and κ_w, which gives partial credit to near-miss categories, is the primary statistic:

$$\kappa_w = 1 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij} O_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij} E_{ij}}, \quad w_{ij} = \frac{(i-j)^2}{(k-1)^2}, \tag{17}$$

where k is the number of categories, O_{ij} is the observed fraction of samples with predicted category i and reference category j, E_{ij} is the fraction expected by chance from the marginal distributions, and w_{ij} is the quadratic disagreement weight; the unweighted κ corresponds to the special case with $w_{ij} = 0$ for $i = j$ and $w_{ij} = 1$ otherwise. We interpret values with the standard bands (0.21–0.40 fair,

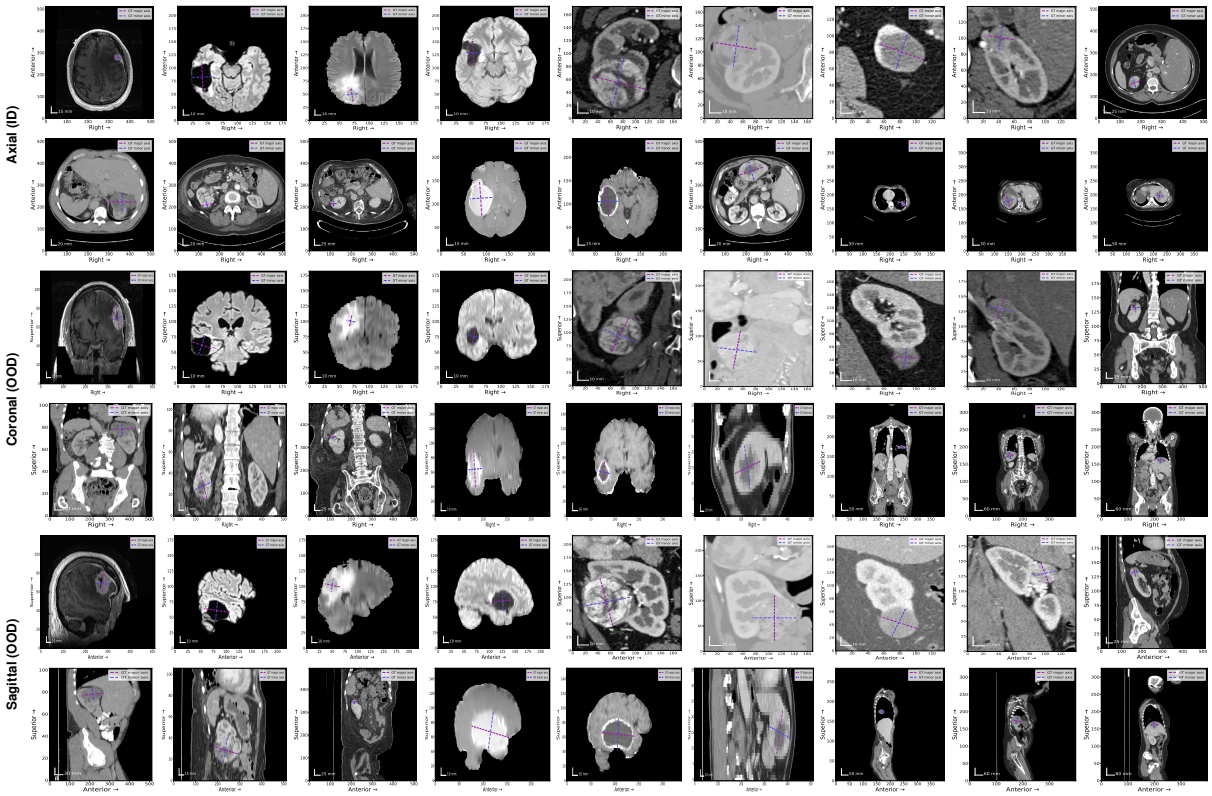


Figure 11: Examples of the Plane-OOD setting in the T/L size estimation task. Panels are grouped by imaging plane: axial (in-distribution; the only plane used during training) and coronal and sagittal (OOD; held out from training). Panels at the same grid position within each plane group show the same subject and target, with the ground truth major and minor axes overlaid and an L-shaped bar indicating the physical scale. Although the subject and target are identical, the OOD planes present markedly different ***anatomical context, target appearance, and physical extent***, which the model must generalize across.

0.41–0.60 moderate, 0.61–0.80 substantial) (Landis and Koch, 1977). Uncertainty is estimated with a percentile bootstrap over whole imaging volumes (4,000 resamples), from which we also derive a one-sided p-value for agreement above chance. Because κ depends on how close the cohort's values lie to the cutoffs, it is comparable across models within a CDA task but not across tasks.

Results. Table 19 reports the decision agreement of all 18 models on the three CDA tasks. On the cephalometric tasks, no off-the-shelf model except LLaVA-OneVision on SNB ($\kappa_w = 0.243$) achieves agreement significantly above chance. On the kidney tumor staging task, off-the-shelf models reach at most fair ordinal agreement ($\kappa_w \leq 0.361$), and the weakest collapse toward a single category: InternVL3 assigns 99.1% of its parsed predictions to T2b and MedGemma (4B) assigns 81.4%, although only 6.8% of the ground truth measurements fall in that category. This collapse reflects a failure of physical scale: these models emit tumor sizes far beyond the plausible anatomical range regardless of the input image (e.g., InternVL3's median predicted major axis is $5{,}590$ mm against a ground truth median of 38 mm), so nearly every prediction exceeds the highest staging cutoff of 100 mm and folds into the top category. MedVision-V0 is the only model whose agreement is significantly above chance on all three CDA tasks, attaining fair agreement on SNA ($\kappa_w = 0.332$), moderate agreement on SNB ($\kappa_w = 0.439$), and substantial agreement on kidney tumor staging ($\kappa_w = 0.726$) at full parsing coverage. Although κ_w values are not strictly comparable across CDA tasks, the lower agreement on the two angle-derived tasks than on the size-derived staging task mirrors the benchmark results: angle measurement remains harder for MedVision-V0 than length measurements such as landmark distance and tumor size (Tables 7 and 6).

Q Intended Use, License & Data Privacy

Intended Use. The MedVision dataset, benchmark, and the fine-tuned model MedVision-V0 are released for research purposes only. They are in-



Figure 12: In-distribution and Target-OOD targets of the detection (top) and T/L size estimation (bottom) tasks. Each entry is one benchmark target, written as label @ modality, with the source dataset or MR sequence given in parentheses where needed to distinguish targets. The Target-OOD targets are entirely absent from the training label set and constitute the Target-OOD test sets in Table 18.

tended to support the study of quantitative reasoning in vision-language models, reproducible benchmarking, and the development of stronger models. None of these artifacts is a medical device, none has been validated in a clinical setting, and none may be used for diagnosis, treatment planning, patient management, or any other clinical decision-making. As reported in Section 4, even the best model evaluated here retains measurement errors above clinical tolerances. Any use beyond research would require independent clinical validation and the appropriate regulatory approval.

License. MedVision is released under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Users are permitted to utilize, adapt, and build upon this dataset for both academic and commercial purposes, provided that appropriate credit is given. Users are reminded that MedVision acts as a meta-dataset built upon various publicly available source datasets. While the annotations provided by MedVision are covered by the CC-BY 4.0 license, any downstream application must continue to comply with the specific usage terms and licensing requirements stipulated by the curators of the original raw imaging data. It is the responsibility of the user to ensure that their application of this data aligns with the license agreements of all constituent source datasets.

Data Privacy. All source imaging datasets incorporated in MedVision were publicly released in anonymized form by their respective curators. The annotations contributed by MedVision (bounding boxes, tumor/lesion size measurements, and angle/distance measurements) are purely geometric descriptors of image content and contain no information that could identify individual subjects.

Table 15: Training progression base → SFT for three base models: Qwen2.5-VL (7B), Gemma4 (31B), and MedGemma (27B), across all three tasks. For SFT rows, parentheses give the change from the corresponding base model (green: improvement; red: decline).

Model	R ↑	P ↑	F1 ↑	IoU ↑	SR ↑	$IoU_{>0.5}$ ↑	$IoU_{0.5:0.05:0.95}$ ↑
Detection: Anatomy (18 regions, 13.4K samples)							
Qwen2.5-VL	69.7	12.2	16.7	11.3	99.4	5.6	1.7
+ SFT	79.8 (+10.1)	80.5 (+68.3)	78.3 (+61.6)	71.2 (+59.9)	100.0 (+0.6)	79.3 (+73.7)	59.0 (+57.3)
Gemma4	34.8	20.7	22.6	16.7	98.5	13.5	5.8
+ SFT	72.1 (+37.3)	72.2 (+51.5)	70.1 (+47.5)	61.8 (+45.1)	100.0 (+1.5)	68.4 (+54.9)	46.6 (+40.8)
MedGemma	57.2	15.7	19.1	12.9	98.6	6.6	1.8
+ SFT	48.6 (-8.6)	49.5 (+33.8)	44.7 (+25.6)	33.8 (+20.9)	100.0 (+1.4)	29.8 (+23.2)	11.8 (+10.0)
Detection: Tumor/Lesion (8 regions, 8.5K samples)							
Qwen2.5-VL	77.4	3.8	6.5	3.6	99.6	0.0	0.0
+ SFT	47.9 (-29.5)	49.3 (+45.5)	44.1 (+37.6)	35.4 (+31.8)	100.0 (+0.4)	37.0 (+37.0)	17.9 (+17.9)
Gemma4	43.6	14.6	18.1	12.7	99.5	8.4	2.7
+ SFT	52.8 (+9.2)	54.1 (+39.5)	48.8 (+30.7)	39.9 (+27.2)	100.0 (+0.5)	44.4 (+36.0)	21.8 (+19.1)
MedGemma	54.1	4.6	7.5	4.3	97.0	0.1	0.0
+ SFT	12.4 (-41.7)	15.4 (+10.8)	10.6 (+3.1)	7.0 (+2.7)	100.0 (+3.0)	2.5 (+2.4)	0.7 (+0.7)

Model	MAE ↓	MRE ↓	SR ↑	$MRE_{<0.1}$ ↑
Tumor/Lesion Size (2K samples)				
Qwen2.5-VL	2897.2	7680.9	95.4	0.7
+ SFT	11.3 (-2885.9)	28.6 (-7652.3)	100.0 (+4.6)	20.6 (+19.9)
Gemma4	21.7	72.6	98.9	16.8
+ SFT	10.5 (-11.2)	28.1 (-44.5)	100.0 (+1.1)	25.0 (+8.2)
MedGemma	523.4	1905.0	57.0	0.5
+ SFT	17.0 (-506.4)	42.1 (-1862.9)	100.0 (+43.0)	10.3 (+9.8)

Model	MAE ↓	MRE ↓	SR ↑	$MRE_{<0.1}$ ↑
Distance (Ceph+FeTA, 1,100 samples)				
Qwen2.5-VL	63603.8	96542.0	98.3	0.5
+ SFT	4.2 (-63599.6)	7.5 (-96534.5)	100.0 (+1.7)	77.7 (+77.2)
Gemma4	23.8	37.9	99.8	10.8
+ SFT	6.2 (-17.6)	10.1 (-27.8)	100.0 (+0.2)	60.2 (+49.4)
MedGemma	6239.0	8616.4	46.2	6.7
+ SFT	5.2 (-6233.8)	9.1 (-8607.3)	100.0 (+53.8)	68.6 (+61.9)
Angle (Ceph, 960 samples)				
Qwen2.5-VL	48.0	724.9	97.6	2.0
+ SFT	5.8 (-42.2)	53.3 (-671.6)	100.0 (+2.4)	48.8 (+46.8)
Gemma4	23.9	428.0	88.5	11.7
+ SFT	4.3 (-19.6)	57.3 (-370.7)	100.0 (+11.5)	49.8 (+38.1)
MedGemma	46.4	1024.8	93.0	6.8
+ SFT	5.2 (-41.2)	63.4 (-961.4)	100.0 (+7.0)	44.4 (+37.6)

Table 16: RFT data-mixture ablation: sequential three-stage RFT (A/D → T/L → detection; MedVision-V0) vs. single-stage RFT on a mixture of all task data, starting from the same SFT model. Detection metrics, MRE, SR, and $MRE_{<0.1}$ are reported in %. MAEs are given in millimeters (T/L size, distance) and degrees (angle).

Model	R ↑	P ↑	F1 ↑	IoU ↑	SR ↑	$IoU_{>0.5}$ ↑	$IoU_{0.5:0.05:0.95}$ ↑
Detection: Anatomy (18 regions, 13.4K samples)							
Sequential RFT	81.3	80.4	79.1	72.0	100.0	80.1	59.9
Multi-task RFT	80.4	79.3	78.2	71.2	100.0	79.6	59.4
Detection: Tumor/Lesion (8 regions, 8.5K samples)							
Sequential RFT	52.4	50.5	46.9	38.2	100.0	40.7	21.1
Multi-task RFT	50.2	52.9	47.4	38.5	100.0	40.3	20.5

Model	MAE ↓	MRE ↓	SR ↑	$MRE_{<0.1}$ ↑
Tumor/Lesion Size (2K samples)				
Sequential RFT	10.5	26.0	100.0	23.5
Multi-task RFT	10.5	26.7	100.0	23.4
Distance (Ceph+FeTA, 1,100 samples)				
Sequential RFT	3.6	6.4	100.0	81.4
Multi-task RFT	3.6	6.6	100.0	82.0
Angle (Ceph, 960 samples)				
Sequential RFT	4.7	52.1	99.9	52.0
Multi-task RFT	3.7	41.9	100.0	52.4

Table 17: RFT reward-design ablation: multiplicative composition of the process and answer rewards ($r = r_{\text{format}} + r_{\text{process}} \cdot r_{\text{answer}}$; default) vs. additive composition ($r = r_{\text{format}} + r_{\text{process}} + r_{\text{answer}}$), both trained as a single-stage multi-task RFT from the same SFT model. The detection task defines no process reward and is therefore not affected by the composition. MRE, SR, and $\text{MRE}_{<0.1}$ are reported in %. MAEs are given in millimeters (T/L size, distance) and degrees (angle).

Reward	MAE ↓	MRE ↓	SR ↑	$\text{MRE}_{<0.1}$ ↑
Tumor/Lesion Size (2K samples)				
$r_{\text{process}} \cdot r_{\text{answer}}$ (default)	10.5	26.7	100.0	23.4
$r_{\text{process}} + r_{\text{answer}}$	10.1	26.9	100.0	25.2
Distance (Ceph+FeTA, 1,100 samples)				
$r_{\text{process}} \cdot r_{\text{answer}}$ (default)	3.6	6.6	100.0	82.0
$r_{\text{process}} + r_{\text{answer}}$	3.6	6.7	100.0	80.9
Angle (Ceph, 960 samples)				
$r_{\text{process}} \cdot r_{\text{answer}}$ (default)	3.7	41.9	100.0	52.4
$r_{\text{process}} + r_{\text{answer}}$	3.8	48.1	100.0	50.3

Table 18: OOD generalization of the base model (Qwen2.5-VL-7B), SFT, and SFT-RFT (MedVision-V0) on detection and T/L size estimation tasks. Parentheses give the change from the preceding stage (SFT vs. base; SFT-RFT vs. SFT; green: improvement; red: decline; gray: no change).

Model	R ↑	P ↑	F1 ↑	IoU ↑	SR ↑	$\text{IoU}_{>0.5}$ ↑	$\text{IoU}_{0.5:0.05:0.95}$ ↑
Plane-OOD Detection: Anatomy (25 regions, 25.6K samples)							
base	32.6	9.1	11.6	7.5	72.9	2.7	0.9
SFT	48.5 (+15.9)	55.7 (+46.6)	49.0 (+37.4)	43.7 (+36.2)	100.0 (+27.1)	47.4 (+44.7)	34.6 (+33.7)
SFT-RFT	49.3 (+0.8)	54.7 (-1.0)	49.5 (+0.5)	44.3 (+0.6)	100.0 (0.0)	47.5 (+0.1)	35.6 (+1.0)
Plane-OOD Detection: Tumor/Lesion (16 regions, 15.2K samples)							
base	43.8	5.2	8.3	4.9	73.2	0.5	0.1
SFT	30.1 (-13.7)	43.8 (+38.6)	30.7 (+22.4)	23.9 (+19.0)	100.0 (+26.8)	22.9 (+22.4)	10.7 (+10.6)
SFT-RFT	35.7 (+5.6)	48.8 (+5.0)	36.6 (+5.9)	28.9 (+5.0)	100.0 (0.0)	28.1 (+5.2)	13.8 (+3.1)

Model	R ↑	P ↑	F1 ↑	IoU ↑	SR ↑	$\text{IoU}_{>0.5}$ ↑	$\text{IoU}_{0.5:0.05:0.95}$ ↑
Target-OOD Detection: Anatomy (27 regions, 5.6K samples)							
base	21.5	5.6	7.3	4.7	62.3	1.2	0.2
SFT	31.3 (+9.8)	44.7 (+39.1)	32.9 (+25.6)	28.5 (+23.8)	98.6 (+36.3)	29.5 (+28.3)	21.2 (+21.0)
SFT-RFT	34.8 (+3.5)	47.0 (+2.3)	36.0 (+3.1)	31.2 (+2.7)	98.8 (+0.2)	31.7 (+2.2)	23.3 (+2.1)
Target-OOD Detection: Tumor/Lesion (3 regions, 7.3K samples)							
base	34.1	2.6	4.4	2.5	44.7	0.1	0.0
SFT	40.6 (+6.5)	50.0 (+47.4)	40.2 (+35.8)	32.7 (+30.2)	99.8 (+55.1)	35.4 (+35.3)	17.9 (+17.9)
SFT-RFT	47.4 (+6.8)	51.5 (+1.5)	45.5 (+5.3)	38.0 (+5.3)	99.9 (+0.1)	41.7 (+6.3)	23.5 (+5.6)

Model	MAE ↓	MRE ↓	SR ↑	$\text{MRE}_{<0.1}$ ↑
Plane-OOD Tumor/Lesion Size (16 labels, 4.5K samples)				
base	16240.6	44054.2	96.9	0.4
SFT	20.3 (-16220.3)	42.1 (-44012.1)	100.0 (+3.1)	6.8 (+6.4)
SFT-RFT	18.2 (-2.1)	36.4 (-5.7)	100.0 (0.0)	8.1 (+1.3)

Model	MAE ↓	MRE ↓	SR ↑	$\text{MRE}_{<0.1}$ ↑
Target-OOD Tumor/Lesion Size (5 labels, 764 samples)				
base	25989.2	132518.7	99.2	0.1
SFT	10.9 (-25978.3)	41.6 (-132477.1)	100.0 (+0.8)	14.1 (+14.0)
SFT-RFT	9.5 (-1.4)	34.7 (-6.9)	100.0 (0.0)	17.7 (+3.6)

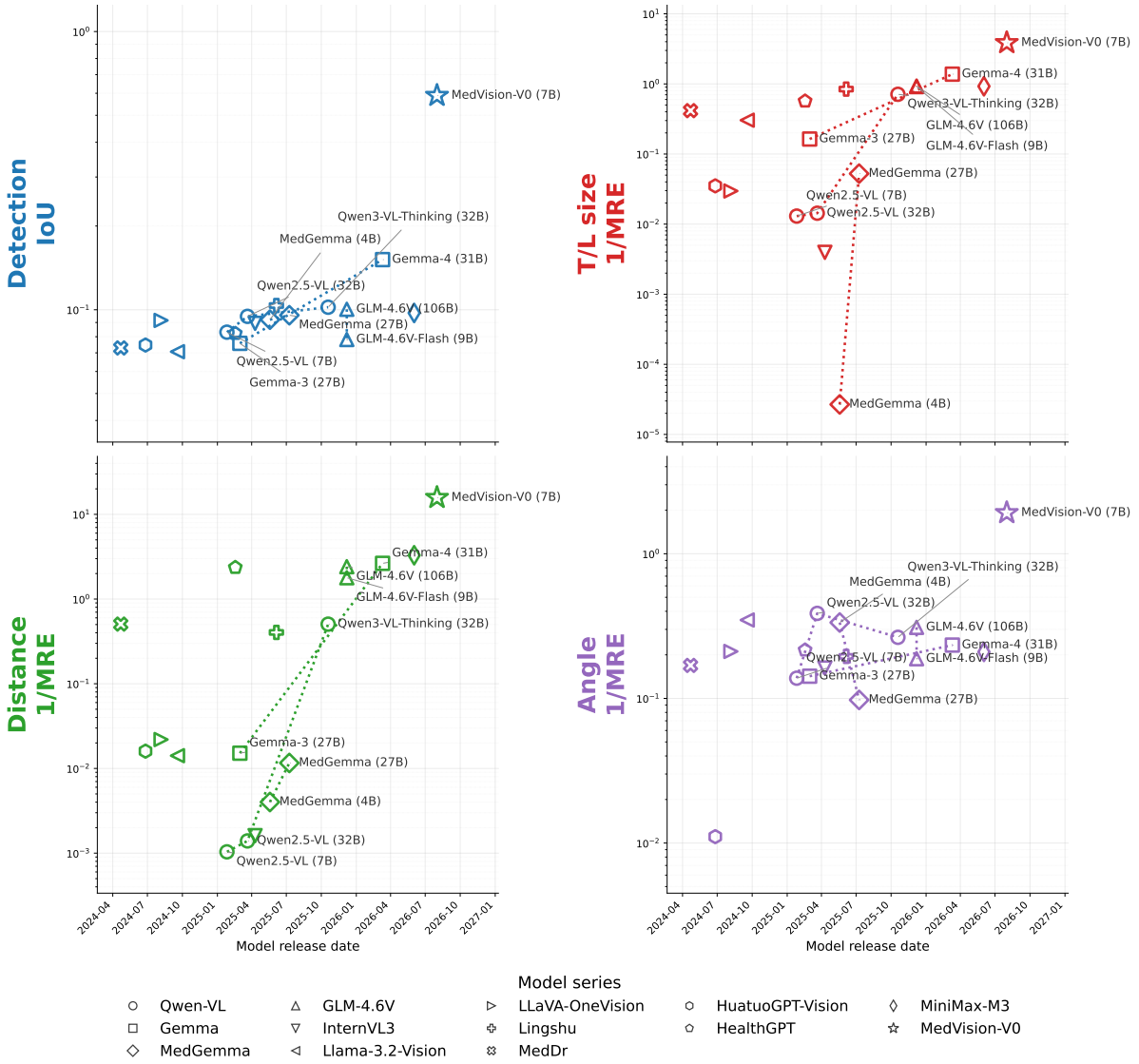


Figure 13: Overview of model performance against model release date. From left to right and top to bottom: detection IoU, and the inverse mean relative error (1/MRE) of T/L size estimation, distance measurement, and angle measurement, on a logarithmic scale; ***higher is better in all panels***. Markers denote model series, dotted lines connect successive releases within a series, and the star marks MedVision-V0.

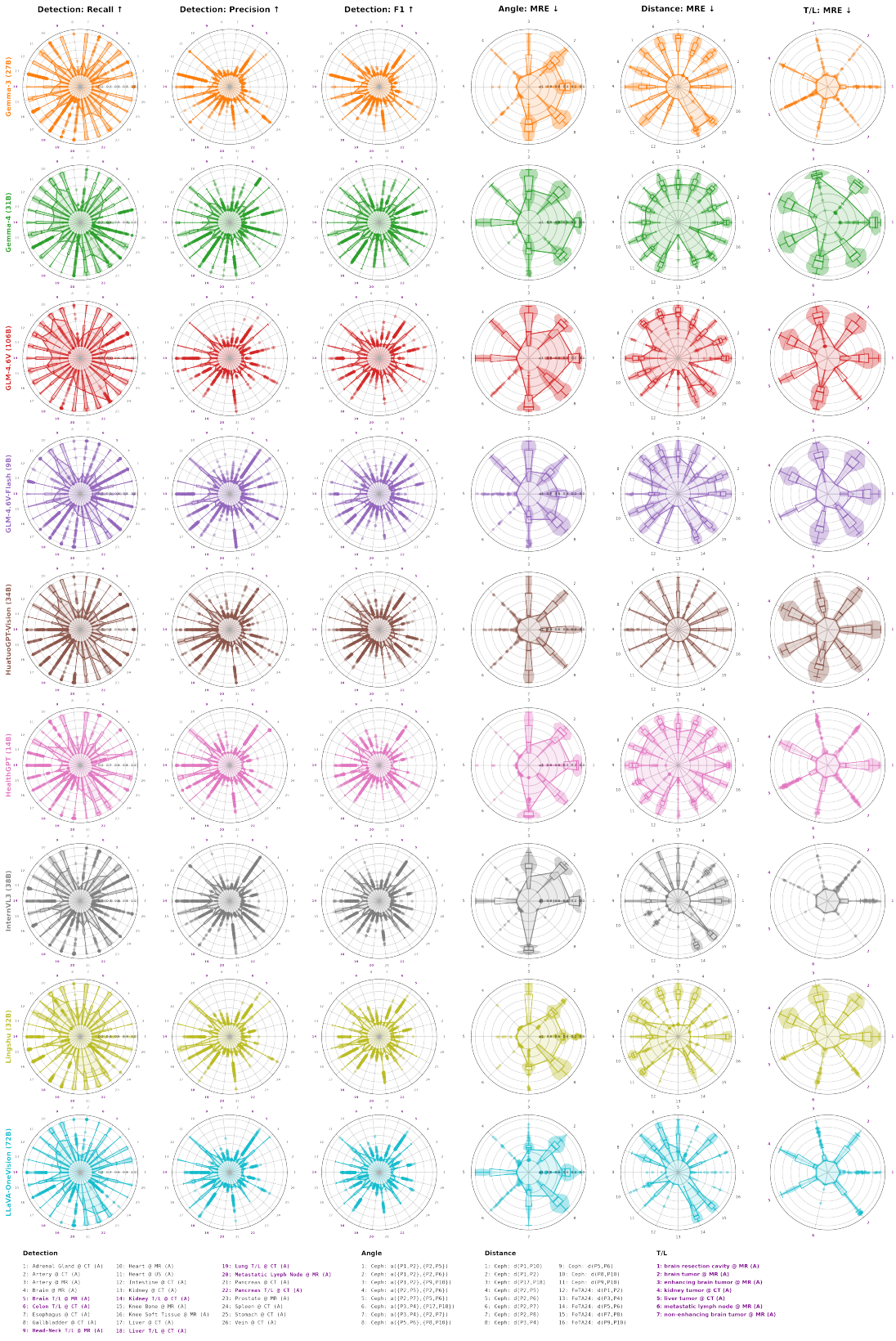


Figure 14: Per-model performance profiles of the evaluated VLMs (part 1 of 2; the first nine models). Each row shows one model's results as six radar charts: recall, precision, and F1 in the detection task; MRE in the angle and distance measurement tasks; and MRE in the T/L size task. The MRE axes are inverted so that the outermost ring corresponds to MRE $= 0$ and the center to MRE ≥ 1, i.e., ***larger polygons always indicate better performance***. Violin and box plots on each spoke show the model's per-sample metric distribution, with dots marking outliers beyond the $1.5\times$IQR whiskers. MedVision-V0 is shown in Figure 10.

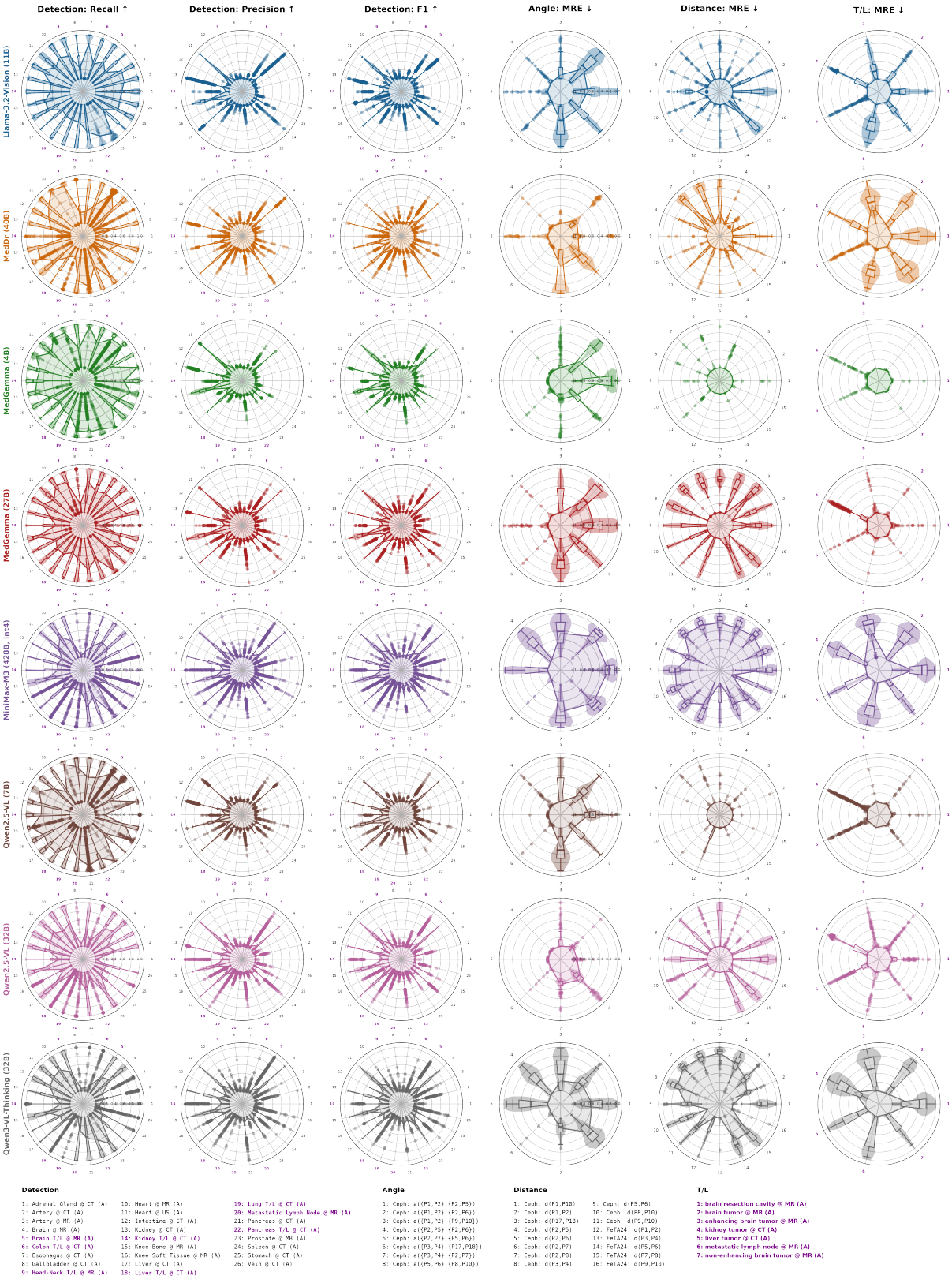


Figure 15: Per-model performance profiles of the evaluated VLMs (part 2 of 2; the remaining eight models).

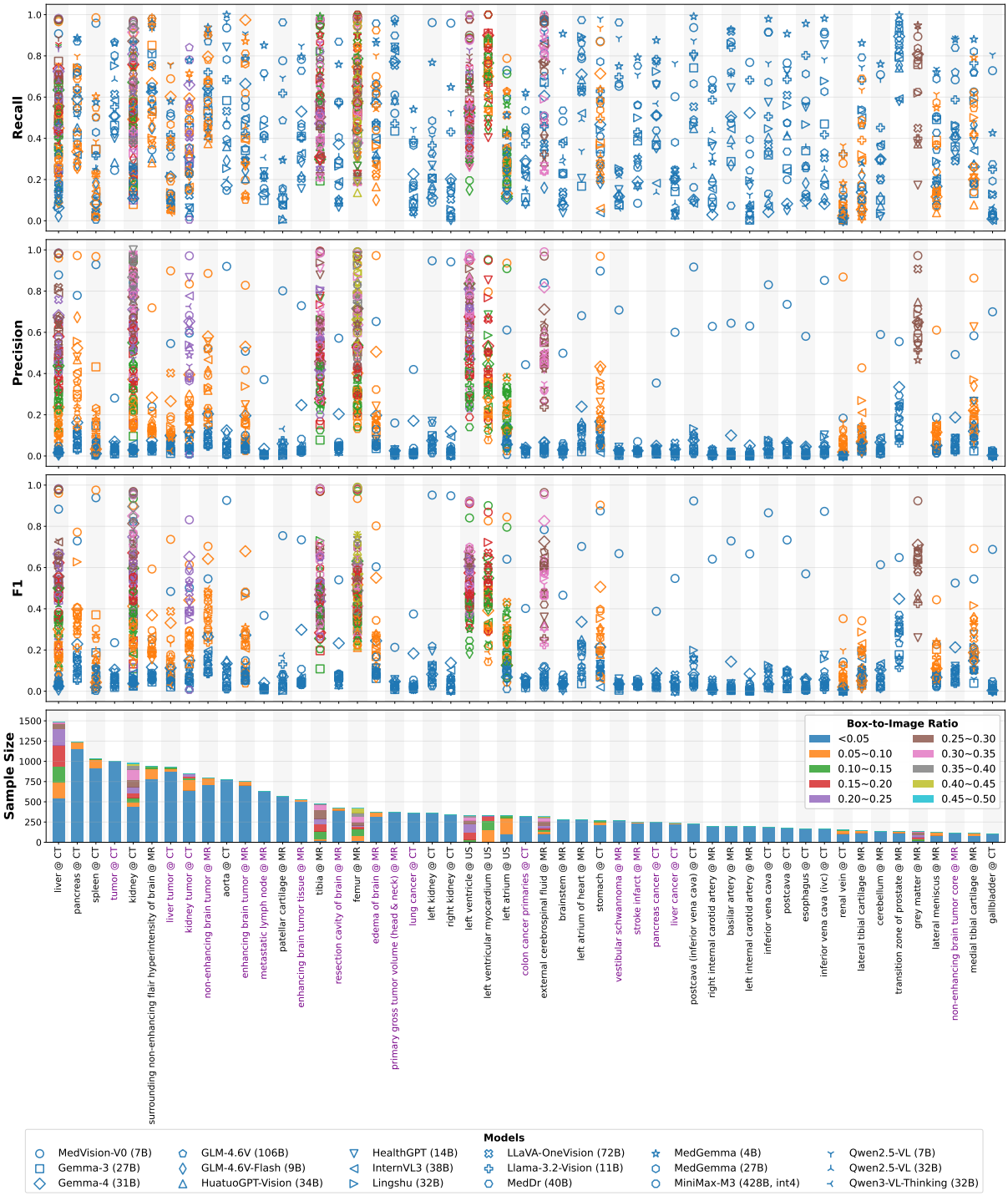


Figure 16: Label-level detection performance of the evaluated VLMs at each box-image-ratio group, where box-image-ratio is the area ratio of the bounding box to the image. Top: recall, precision, and F1 score for each label and box-image-ratio group. Bottom: the composition of box-image-ratio groups for each label.

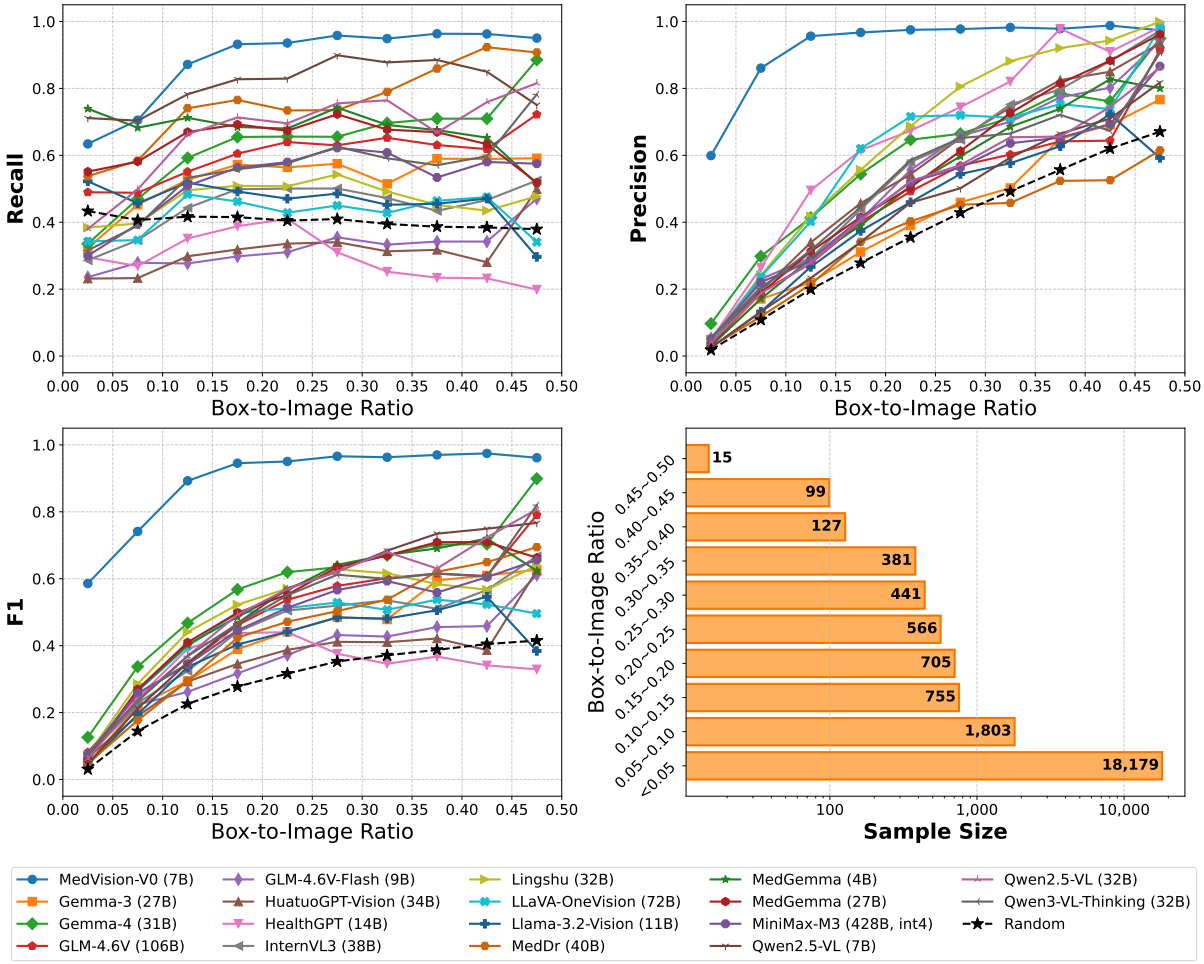


Figure 17: Effect of target size, quantified by box-image-ratio (the area ratio of the bounding box to the image), on detection performance. All detection targets are grouped by box-image-ratio, and each model's overall detection performance is computed within each group. ***Random*** denotes a random-guessing baseline in which each sample's metrics are averaged over 100 uniformly sampled boxes.

Table 19: Clinical decision agreement of all evaluated models: the category of the predicted measurement vs. the category of the ground truth measurement, both derived through the same published cutoff table. Acc: category agreement on parsed samples; κ: Cohen's kappa; κ_w: quadratic-weighted kappa; CI and p: 95% percentile bootstrap confidence interval and one-sided p-value for agreement above chance, computed for κ_w by resampling whole imaging volumes (4,000 resamples); SR: success rate, the fraction of predictions from which a measurement could be parsed, in %. κ values are not comparable across CDA tasks.

Model	**Acc** ↑	κ ↑	κ_w ↑	**95% CI**	**p**	**SR** ↑
SNA maxillary position (Ceph-Bio-400; 120 samples)						
MedVision-V0 (7B)	0.458	0.183	0.332	[0.163, 0.484]	<0.001	100.0
InternVL3 (38B)	0.358	0.078	0.119	[-0.048, 0.283]	0.080	100.0
Gemma4 (31B)	0.393	0.066	-0.033	[-0.294, 0.223]	0.609	46.7
Lingshu (32B)	0.325	0.052	0.090	[-0.029, 0.206]	0.071	100.0
Qwen2.5-VL (7B)	0.303	0.022	0.037	[-0.039, 0.115]	0.176	99.2
GLM-4.6V-Flash (9B)	0.322	0.021	0.004	[-0.073, 0.090]	0.457	98.3
Llama3.2-Vision (11B)	0.300	0.017	-0.002	[-0.031, 0.029]	0.556	100.0
LLaVA-OneVision (72B)	0.300	0.000	0.001	[-0.169, 0.174]	0.499	100.0
Qwen2.5-VL (32B)	0.283	0.000	0.000	[0.000, 0.000]	1.000	100.0
Gemma3 (27B)	0.283	-0.006	0.014	[-0.093, 0.116]	0.404	100.0
HuatuoGPT-Vision (34B)	0.293	-0.014	-0.044	[-0.190, 0.112]	0.712	96.7
MedDr (40B)	0.278	-0.014	-0.025	[-0.123, 0.070]	0.709	90.0
Qwen3-VL-Thinking (32B)	0.295	-0.018	0.001	[-0.056, 0.070]	0.539	79.2
MedGemma (4B)	0.291	-0.020	-0.023	[-0.194, 0.148]	0.606	97.5
MiniMax-M3 (428B)	0.250	-0.066	-0.027	[-0.269, 0.190]	0.590	30.0
HealthGPT-L14 (14B)	0.250	-0.082	-0.132	[-0.289, 0.027]	0.951	100.0
MedGemma (27B)	0.233	-0.082	-0.153	[-0.332, 0.030]	0.948	75.0
GLM-4.6V (106B)	0.272	-0.121	-0.018	[-0.164, 0.132]	0.595	85.8
SNB mandibular position (Ceph-Bio-400; 120 samples)						
MedVision-V0 (7B)	0.483	0.227	0.439	[0.292, 0.575]	<0.001	100.0
LLaVA-OneVision (72B)	0.425	0.153	0.243	[0.097, 0.392]	<0.001	100.0
HealthGPT-L14 (14B)	0.375	0.055	0.090	[-0.088, 0.258]	0.155	100.0
MedGemma (27B)	0.367	0.055	0.122	[-0.066, 0.305]	0.100	90.8
MedDr (40B)	0.345	0.050	0.080	[-0.057, 0.211]	0.110	96.7
MiniMax-M3 (428B)	0.341	0.029	0.050	[-0.209, 0.315]	0.357	36.7
GLM-4.6V-Flash (9B)	0.308	0.020	0.012	[-0.045, 0.067]	0.331	100.0
Llama3.2-Vision (11B)	0.308	0.018	0.019	[-0.024, 0.079]	0.256	100.0
Lingshu (32B)	0.300	-0.001	-0.002	[-0.103, 0.102]	0.526	100.0
Qwen2.5-VL (7B)	0.294	-0.007	-0.011	[-0.067, 0.041]	0.667	99.2
Gemma3 (27B)	0.283	-0.021	-0.032	[-0.133, 0.060]	0.744	100.0
Qwen2.5-VL (32B)	0.277	-0.028	-0.034	[-0.107, 0.026]	0.852	99.2
GLM-4.6V (106B)	0.281	-0.034	-0.107	[-0.242, 0.020]	0.948	95.0
Qwen3-VL-Thinking (32B)	0.302	-0.035	-0.043	[-0.129, 0.025]	0.882	71.7
Gemma4 (31B)	0.273	-0.037	-0.017	[-0.159, 0.119]	0.583	91.7
InternVL3 (38B)	0.283	-0.049	-0.077	[-0.231, 0.074]	0.842	100.0
HuatuoGPT-Vision (34B)	0.263	-0.082	-0.120	[-0.277, 0.038]	0.927	95.0
MedGemma (4B)	0.250	-0.092	-0.154	[-0.283, -0.026]	0.993	96.7
AJCC kidney tumor staging (KiPA22 + KiTS23; 1,025 samples)						
MedVision-V0 (7B)	0.657	0.379	0.726	[0.428, 0.858]	<0.001	100.0
Gemma4 (31B)	0.402	0.203	0.361	[0.166, 0.522]	<0.001	97.9
Qwen3-VL-Thinking (32B)	0.405	0.159	0.342	[0.106, 0.509]	<0.001	96.7
GLM-4.6V-Flash (9B)	0.356	0.153	0.238	[0.063, 0.421]	<0.001	95.9
Lingshu (32B)	0.364	0.140	0.266	[0.079, 0.439]	<0.001	99.9
GLM-4.6V (106B)	0.352	0.127	0.235	[0.050, 0.408]	0.002	93.2
HuatuoGPT-Vision (34B)	0.459	0.120	0.185	[0.051, 0.277]	<0.001	98.0
MiniMax-M3 (428B)	0.323	0.113	0.266	[0.123, 0.387]	<0.001	95.0
HealthGPT-L14 (14B)	0.267	0.101	0.182	[0.052, 0.313]	0.001	100.0
MedGemma (27B)	0.180	0.051	0.086	[0.017, 0.170]	0.004	51.0
Gemma3 (27B)	0.198	0.050	0.088	[0.020, 0.173]	0.003	98.6
MedDr (40B)	0.326	0.026	0.095	[-0.031, 0.213]	0.072	83.6
LLaVA-OneVision (72B)	0.220	0.025	0.032	[-0.038, 0.100]	0.173	100.0
MedGemma (4B)	0.149	0.019	0.039	[0.009, 0.075]	0.005	90.4
Qwen2.5-VL (32B)	0.142	0.014	0.060	[0.005, 0.128]	0.015	99.9
InternVL3 (38B)	0.073	0.004	-0.000	[-0.002, 0.002]	0.550	100.0
Llama3.2-Vision (11B)	0.286	0.002	0.046	[-0.049, 0.147]	0.207	99.4
Qwen2.5-VL (7B)	0.092	-0.059	-0.051	[-0.215, 0.049]	0.658	92.3