

MEDVISION: DATASET AND BENCHMARK FOR QUANTITATIVE MEDICAL IMAGE ANALYSIS

Anonymous authors

Paper under double-blind review

ABSTRACT

Current vision-language models (VLMs) in medicine are primarily designed for categorical question answering (e.g., “Is this normal or abnormal?”) or qualitative tasks (e.g., “Describe the image”). 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 physicians 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 **30.8 million** image-annotation pairs. 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. Our benchmarks show that current off-the-shelf VLMs perform poorly on these tasks. However, with supervised fine-tuning on MedVision, we significantly enhance their performance across detection, T/L estimation, and A/D measurement, demonstrating reduced error rates and improved precision. This work 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., 2024) 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 has concentrated on categorical (Lau et al., 2018; Liu et al., 2021; Ben Abacha et al., 2019) or qualitative outputs (Subramanian et al., 2020): answering whether an image is “normal or abnormal,” 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 estimating tumor diameter to determine cancer stage, measuring joint angles for orthopedic planning, and computing lesion volume. Clinicians seldom make a diagnosis based solely on a binary label; instead they combine measured quantities with clinical context to reach a conclusion. 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 benchmarks (Wu et al., 2025; Zhou et al., 2025) and 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, and off-the-shelf VLMs are rarely evaluated on tasks that require exact spatial localization and numeric regression.

There are several reasons why quantitative vision–language capability is both challenging and underexplored. First, measurement tasks demand precise localization (often at sub-pixel or small-structure scales) and robust handling of modality-specific image characteristics (CT, MRI, X-ray, ultrasound).

Second, producing correct measurements requires the model to understand geometric relationships and physical units, not just semantic categories. Third, public datasets with standardized, high-quality measurement annotations are sparse or fragmented across modalities and anatomies, making large-scale evaluation and supervised fine-tuning difficult. Finally, many VLM architectures and training objectives are not designed to produce calibrated numeric outputs or to combine localized visual evidence with explicit numeric reasoning.

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 (e.g., radiography, CT, MRI, ultrasound), comprising **30.8 million** images with rich, structured annotations suitable for measurement tasks. 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). These tasks capture the range of quantitative demands found in routine radiology and specialty workflows.

Using MedVision, we perform a systematic evaluation of both open-weight and commercial VLMs in two settings: off-the-shelf inference and supervised fine-tuning on the MedVision training splits. Our benchmarks show that 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. We then demonstrate that targeted supervised fine-tuning on MedVision materially improves model performance across detection metrics (recall, precision, F1, IoU) and measurement accuracy (reduced absolute and relative errors for size and angle/distance). Alongside quantitative results, we analyze common failure modes and identify persistent challenges that future research must address.

To summarize, our main contributions are as follows:

- We highlight a clinically important but underappreciated gap: modern VLMs are not reliably able to produce precise quantitative measurements from medical images.
- We introduce MedVision, a large-scale, multi-modality dataset and benchmark for quantitative medical image analysis, covering 22 public datasets and 30.8M images with structured measurement annotations.
- We provide the first comprehensive evaluation of contemporary VLMs on detection, tumor/lesion size estimation, and angle/distance measurement, demonstrating the limitations of off-the-shelf models.
- We release the benchmark (data splits, evaluation code, and baselines) 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). The benchmark is conducted in the form of open-ended VQA. We convey physical spacing information (i.e., pixel size) in the text prompt to provide VLMs with necessary context for quantitative analysis tasks.

2.1 DATASET CONSTRUCTION

We constructed the MedVision dataset with public medical imaging data and annotations. Details on data inclusion criteria, image preprocessing, data annotation, and codebase are described below.

Inclusion Criteria: Public medical imaging data with pixel-level annotations (such as mask or landmark coordinate) 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 from images.

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.,

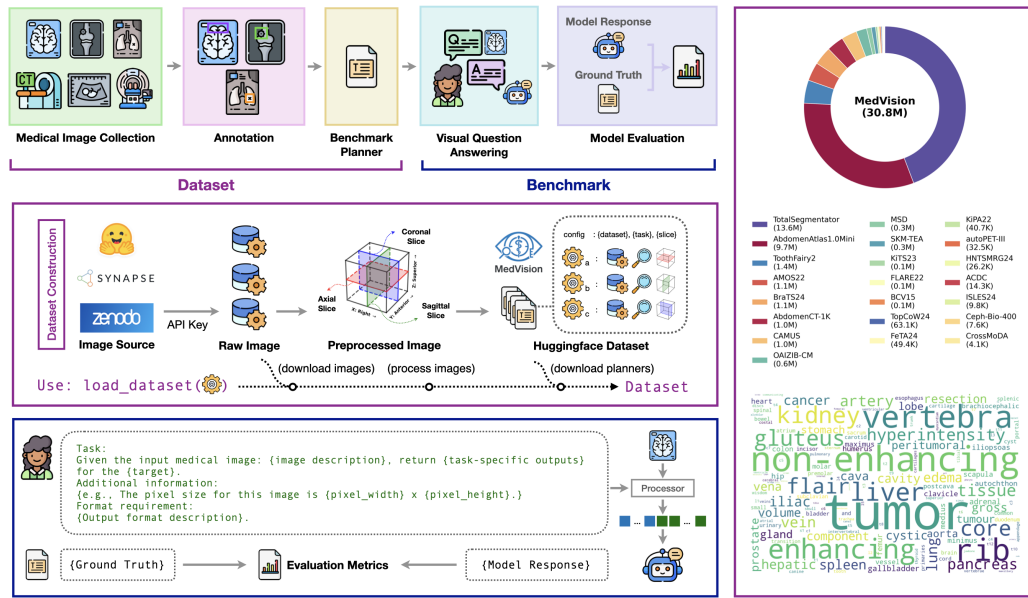


Figure 1: Overview of MedVision dataset construction and benchmark design (left), and dataset summary (right). The MedVision dataset was constructed for out-of-the-box use: `load_dataset` is all you need. Appendix A.1 shows examples of imaging data and quantitative annotation.

2018), AMOS22 (Ji et al., 2022), autoPET-II ¹, BCV15 ², 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), CrossModDA (Dorent et al., 2023), FeTA24 (Payette et al., 2023), FLARE22 (Ma et al., 2024), HNTSMRG24 ³, ISLES24 ⁴, KiPA22 (He et al., 2021; 2020b; Shao et al., 2011; 2012), KiTS23 (Heller et al., 2023), MSD (Antonelli et al., 2022), OAIZIB-CM (Ambellan et al., 2019; Yao et al., 2024; Yao & Chen, 2025), SKM-TEA (Desai et al., 2021), ToothFairy2 (Boelli et al., 2025; 2024; Lumetti et al., 2024), TopCoW24 (Yang et al., 2025), and TotalSegmentator (Wasserthal et al., 2023).

Image Preprocessing: We converted the raw imaging data to the standard format and applied standardization as follows. Data are stored in 3D volumes with image 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. Images are neither resampled, cropped, padded, nor rescaled. 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 (b-box), bidirectional tumor/lesion (T/L) size, and angle/distance (A/D) measurements for detection, T/L size and A/D estimate tasks, respectively.

(i) Bounding Box Annotation: For each segmentation label in a 2D slice, a bounding box is fitted to the mask. Metadata of b-box, including the coordinates of the bottom-left and upper-right corners, pixel and physical dimensions are recorded in the benchmark planner. For labels with scattered clusters, b-box for each cluster is recorded in the released MedVision dataset. However, we excluded 2D slices with multiple b-boxes in our benchmark since we are focusing on single instance detection in this study. Targets with less than 10 pixels in any dimension are excluded from our benchmark.

¹<https://autopet-iii.grand-challenge.org>

²<https://doi.org/10.7303/syn3193805>

³<https://hntsmrg24.grand-challenge.org>

⁴<https://isles-24.grand-challenge.org>

(ii) **Tumor/Lesion Size Annotation:** Bidirectional measurements of tumor/lesion 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 fitted ellipse to each T/L label in 2D slices. Ellipse fitting is conducted in real-world coordinate system, so that the fitting process takes into account the physical spacing and results in accurate measuring directions (i.e., major and minor axes). Fitted ellipses whose endpoints are outside the buffer zone between the 10% shrunk and enlarged bounding boxes of the target region are excluded to avoid unreliable annotations. The endpoint coordinates of the two axes (in pixel indices) and physical lengths are recorded in the benchmark planner.

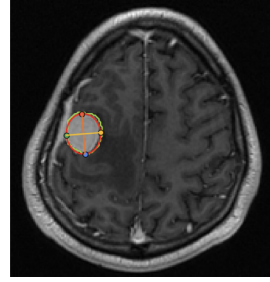


Figure 2: Tumor/lesion size annotation.

(iii) **Angle/Distance Annotation:** Angle and distance measurements are calculated from human annotated landmarks. We collected data from the Ceph-Bio-400 and FeTA24 datasets and generated standardized annotations in the format of landmark coordinates and landmark/angle/distance maps, where each measurement is well-defined and can be calculated from landmark coordinates. Figure 3 shows examples of landmarks in the two datasets. We used the original landmark labels in Ceph-Bio-400 (Lindner et al., 2016). In FeTA24, we assigned labels (P0-P10) to 10 landmarks, where each pair of landmarks defines a distance. Note that each label unambiguously refers to a specific anatomical points, e.g., P0 and P1 denote the most anterior and posterior points of the corpus callosum, respectively.

Dataset Splits: MedVision dataset contains **30.8 million** images-annotation pairs across 22 data sources and 3 tasks. For each task, we randomly split the data into training (70%) and test sets (30%) at the patient level. We use the training set for model finetuning where applicable and test set for evaluation.

Dataset Codebase: MedVision is built for transparency, version control, easy expansion, and simple use (Figure 1). Its dataset codebase supports reproducible construction and automated loading so users can get started with a single call: `load_dataset()`. A small data configuration (dataset + task + plane) selects the subset to load, while the codebase handles downloading, preprocessing, and preparing benchmark splits.

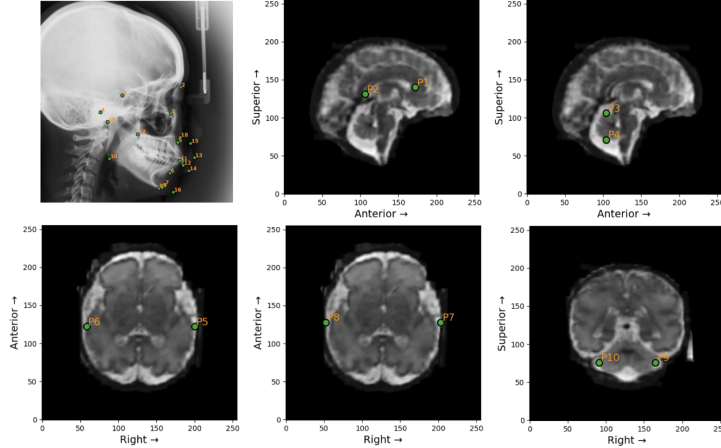


Figure 3: Landmarks in the Ceph-Bio-400 (top-left) and FeTA24 datasets.

2.2 BENCHMARK DESIGN

We focus on clinically related open-ended VQA tasks for our benchmark: (1) healthy anatomical structures and abnormalities detection, (2) T/L size estimation, and (3) A/D measurement. Model outputs are parsed to extract the required estimates and compared with ground truths.

Baseline: We benchmarked 15 state-of-the-art open-weight and commercial VLMs as baselines. Open-weights models include Qwen2.5-VL (Bai et al., 2025), Lingshu (Team et al., 2025b), InternVL3 (Zhu et al., 2025), Gemma3 (Team et al., 2025a), MedGemma (Sjellergren et al., 2025), Llama3.2-Vision, LLaVA-OneVision (Li et al., 2024a), LLaVA-Med (Li et al., 2023), MedDr (He et al., 2024), HuatuoGPT-Vision (Chen et al., 2024), and HealthGPT-L14 (Lin et al., 2025). We

Table 1: The MedVision dataset consists of public medical images and quantitative annotations from this study. MRI: Magnetic Resonance Imaging; CT: Computed Tomography; PET: positron emission tomography; US: Ultrasound; b-box: bounding box; T/L: tumor/lesion size; A/D: angle/distance; HF: HuggingFace; GC: Grand-Challenge; † redistributed.

Dataset	Anatomy	Modality	Annotation	Availability	Source	# Sample (Train / Test)		
						b-box	T/L	A/D
AbdomenAtlas	abdomen	CT	b-box	open	HF	6.8 / 2.9M	0	0
AbdomenCT-1K	abdomen	CT	b-box	open	Zenodo	0.7 / 0.3M	0	0
ACDC	heart	MRI	b-box	open	HF†, others	9.5 / 4.8K	0	0
AMOS22	abdomen	CT, MRI	b-box	open	Zenodo	0.8 / 0.3M	0	0
autoPEI-III	whole body	CT, PET	b-box, T/L	open	HF†, others	22 / 9.7K	0.5 / 0.2K	0
BCV15	abdomen	CT	b-box	open	HF†, Synapse	71 / 30K	0	0
BraTS24	brain	MRI	b-box, T/L	open	HF†, Synapse	0.8 / 0.3M	7.9 / 3.1K	0
CAMUS	heart	US	b-box	open	HF†, others	0.7 / 0.3M	0	0
Ceph-Bio-400	head and neck	X-ray	b-box, A/D	open	HF†, others	0	0	5.3 / 2.3K
CrossModDA	brain	MRI	b-box	open	HF†, Zenodo	3.0 / 1.0K	0	0
FeTA24	fetal brain	MRI	b-box, A/D	registration	Synapse	34 / 15K	0	0.2 / 0.1K
FLARE22	abdomen	CT	b-box	open	HF†, others	72 / 33K	0	0
HNTSMRG24	head and neck	MRI	b-box, T/L	open	Zenodo	18 / 6.6K	1.0 / 0.4K	0
ISLES24	brain	MRI	b-box	open	HF†, GC	7.3 / 2.5K	0	0
KiPA22	kidney	CT	b-box, T/L	open	HF†, GC	26 / 11K	2.1 / 1.0K	0
KITS23	kidney	CT	b-box, T/L	open	HF†, GC	80 / 35K	5.9 / 2.6K	0
MSD	multiple	CT, MRI	b-box, T/L	open	others	0.2 / 0.1M	5.3 / 2.2K	0
OAIZIB-CM	knee	MRI	b-box	open	HF	0.5 / 0.2M	0	0
SKM-TEA	knee	MRI	b-box	registration	others	0.2 / 0.1M	0	0
ToothFairy2	tooth	CT	b-box	registration	others	1.0 / 0.4M	0	0
TopCoW24	brain	CT, MRI	b-box	open	HF†, Zenodo	43 / 20K	0	0
TotalSegmentator	multiple	CT, MRI	b-box	open	HF†, Zenodo	9.6 / 4.0M	0	0
Total						22 / 9.2M	23 / 9.6K	5.6 / 2.4K

evaluated Gemini2.5-Flash under the setting of with and without tool use. Details of these models are provided in Appendix A.2.

Supervised Finetuning (SFT): We define task-specific data splits for training and testing. Detection and T/L size tasks use axial slices, while A/D measurement uses all three planes due to limited data. Fine-tuning employed LoRA (Hu et al., 2021) with $r = 16$, $\alpha = 16$, dropout 0.05, applied to all linear projections, with the embedding layer and LM head trainable.

Out-of-Distribution (OOD) Performance: We evaluated the OOD performance of baselines and SFT models on 3 tasks using 2 types of OOD data: (1) **plane-OOD**, where models are evaluated on coronal and sagittal slices after finetuned only on axial slices; (2) **target-OOD**, where models are evaluated on holdout datasets with unseen targets.

Prompt Design: We designed text prompts to convey the image description (anatomy and modality, such as “brain MRI”), task instruction (task type and target, such as “bounding box of brain tumor”), format requirement, and optionally physical spacing information. Physical spacing is not provided in detection tasks since the expected output is relative position of b-box in $[0, 1]$. Prompt templates for the detection, T/L size, and A/D measurement tasks are provided in Appendix A.3.

Physical Spacing Information: A key feature of our dataset is the inclusion of quantitative annotations, which requires careful handling of physical spacing in text prompts. Since each VLM applies its own image preprocessing, we provide spacing values that match the actual model input. For fixed-size pipelines, adjusted pixel size is derived from input-output dimensions, while for dynamic pipelines, we use the model’s image processor to compute the resize ratio. Appendix A.4 details these strategies.

Evaluation Metrics: Model performance is evaluated with task-specific metrics. For detection tasks, we report the region-based recall, precision, F1 score, and intersection over union (IoU) for assessment of location accuracy. These metrics are calculated only on the successfully parsed outputs. To reflect the model’s ability to follow instructions, we also report the success rate (SR) of generating valid numerical outputs, such as coordinates. We also report $\text{IoU}_{>0.5}$, the proportion of samples with IoU greater than 0.5 among all samples. Therefore, $\text{IoU}_{>0.5}$ reflects both instruction

Table 2: VLM performance on detection tasks. Targets are grouped into health anatomy and tumor/lesion detection tasks. Mean metrics weighted by sample sizes are reported in %. R: recall; P: precision; F1: F1 score; IoU: intersection over union; SR: success rate.

Model	Anatomy						Tumor/Lesion					
	R ↑	P ↑	F1 ↑	IoU ↑	SR ↑	IoU _{>0.5} ↑	R ↑	P ↑	F1 ↑	IoU ↑	SR ↑	IoU _{>0.5} ↑
Qwen2.5-VL (7B)	50.4	9.5	12.4	7.6	100	1.4	54.5	3.4	5.6	3.1	100	0.0
Qwen2.5-VL (32B)	35.2	13.6	16.1	11.0	99.7	5.7	38.0	6.2	8.4	5.1	99.9	0.7
Lingshu (32B)	26.4	19.1	17.3	11.6	100	4.0	23.1	5.9	7.2	4.4	100	0.6
InternVL3 (38B)	24.5	14.5	14.5	9.8	100	5.3	39.2	5.9	7.6	4.5	100	0.4
Gemma3 (27B)	33.3	17.4	19.4	13.7	100	9.9	33.4	5.1	6.6	4.0	100	0.5
MedGemma (4B)	66.2	16.3	19.0	12.8	100	7.2	72.1	5.2	8.6	4.9	99.6	0.2
MedGemma (27B)	65.6	17.3	19.3	12.9	100	6.5	65.8	5.7	8.8	5.1	100	0.2
Llama3.2-Vision (11B)	47.0	8.1	10.4	6.8	73.8	2.7	45.6	2.1	3.6	2.0	70.1	0.0
LLava-OneVision (72B)	36.4	17.8	18.5	12.3	100	4.4	38.1	6.0	8.5	5.1	100	0.4
LLaVA-Med-v1.5 (7B)	60.7	15.7	18.6	12.6	99.1	7.0	50.6	4.7	7.3	4.3	89.0	0.8
MedDr (40B)	64.4	11.9	17.3	11.3	99.8	4.1	74.9	4.4	7.5	4.3	97.8	0.1
HuatuoGPT-Vision (34B)	25.3	20.5	16.3	10.7	100	3.6	20.8	5.9	7.0	4.3	100	0.6
HealthGPT-L14	21.0	15.2	13.5	8.3	100	0.5	22.7	6.2	6.8	4.2	100	0.5
Gemini2.5-Flash (w/o tool)	35.4	16.1	18.7	12.8	99.5	6.8	41.4	7.1	10.1	6.3	98.3	1.2
Gemini2.5-Flash (w tool)	29.9	13.0	15.2	10.6	82.0	5.8	38.4	9.1	10.2	6.9	77.8	3.6
Qwen2.5-VL (7B, SFT _{500K})	79.8	76.1	76.0	68.9	100	76.4	52.7	44.2	43.3	35.0	100	38.1
Qwen2.5-VL (32B, SFT _{500K})	76.9	79.8	76.3	68.9	100	76.9	50.8	48.2	44.7	36.6	100	39.7

following and localization ability. For T/L size and A/D measurement tasks, we report the (mean) relative error (M)RE, absolute error (AE), SR and (M)RE_{<k}. (M)RE_{<k} refers to the proportion of samples with (M)RE less than k among all samples.

3 RESULTS & ANALYSIS

In this section, we are interested in answering the following questions:

- **Q1:** How accurately can VLMs detect *anatomical structures* and *tumors/lesions*?
- **Q2:** Can VLMs estimate the *size* of tumors/lesions by predicting the lengths of their major and minor axes?
- **Q3:** Can VLMs measure *angles and distances* in medical images?
- **Q4:** Can supervised fine-tuning improve the performance of VLMs?
- **Q5:** How do VLMs perform these tasks in out-of-distribution settings?

3.1 DETECTION PERFORMANCE

Baselines vs SFT Models: Table 2 summarizes detection performance, with targets grouped into healthy anatomical structures and tumors/lesions. Detailed evaluation metrics are provided in Appendix A.5. Most VLMs can produce valid numerical outputs (i.e., bounding-box corner coordinates) with high success rates (SR > 90%), except Llama-3.2-Vision, which achieves only ~ 70%. However, baseline localization accuracy is low, with IoU below 15% for healthy anatomical structures and below 10% for tumors/lesions. After finetuning with 500K samples from our dataset, the SFT models achieved substantially higher recall, precision, F1, IoU, and IoU_{>0.5} compared to baselines (Figure 4).

OOD Performance: Figure 5 summarizes the OOD performance of pretrained and finetuned Qwen2.5-VL (32B). The pretrained model serves as a baseline for OOD performance. We observed better generalization ability of the SFT model, indicated by higher precision and F1 score on both plane-OOD and target-OOD data.

Failure Mode Analysis: Despite overall improved detection ability of SFT models, they still struggle in some targets. To further understand the failure modes, we analyzed anatomy-level (Appendix A.5.1) and label-level (Appendix A.5.2) performance, and investigated the impact of target size on detection (Appendix A.5.3). We found VLM localization ability increases positively with relative target size and small objects (< 5% in relative size) detection is the bottleneck in current VLMs.

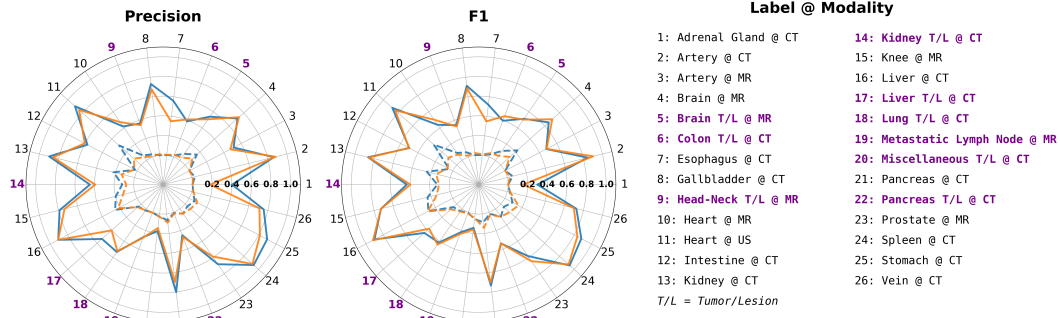


Figure 4: SFT improves VLM performance on healthy anatomical structures and tumor/lesion detection tasks.

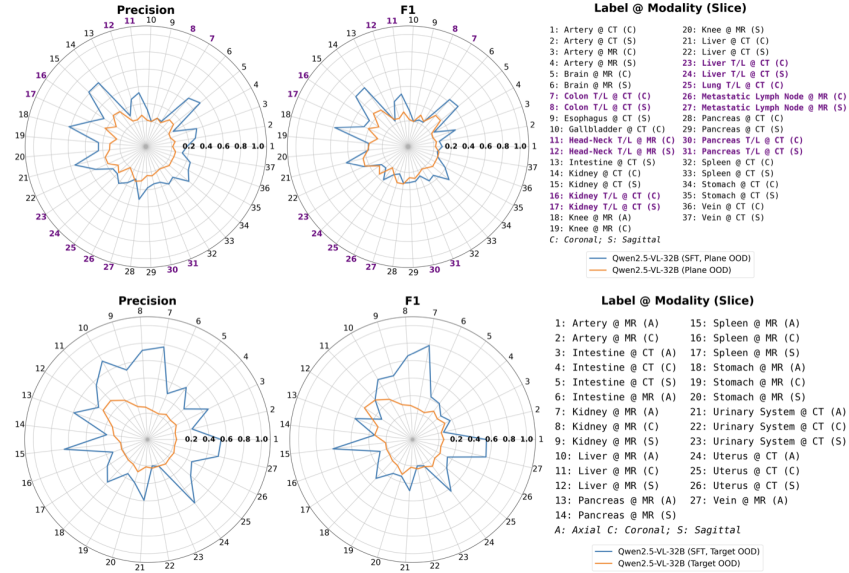


Figure 5: Detection performance on plane-OOD (top) and target-OOD (bottom) data.

Conclusion: Pretrained VLMs show limited ability in medical image detection tasks. With SFT, VLMs achieved dramatic improvement in the detection of both healthy anatomical structures and tumors/lesions. However, small object and tumors/lesions detection is still challenging.

3.2 T/L SIZE ESTIMATION PERFORMANCE

Baselines vs SFT Models: Baseline models struggle with T/L size estimation tasks, with MRE ranging from 50.1% to 116.9% (Table 3). After SFT with 5K samples, Qwen2.5-VL (7B, 32B) can predict T/L size within 10% MRE for $\sim 20\%$ of samples ($\text{MER}_{<0.1}$), with MRE at $\sim 30\%$. The best model (Qwen2.5-VL 32B, SFT) can predict T/L size with an average distance error of 12.8 mm. Figure 6 illustrates the improvement of SFT models for each T/L label.

OOD Performance: SFT models also show partial generalizability when evaluated on unseen datasets (Figure 7). While accuracy declines compared to in-distribution settings, the relative improvement over baselines remains evident, implying that learned quantitative reasoning skills transfer across domains to some extent. However, performance drops highlight that dataset diversity and annotation consistency remain critical for robust generalization.

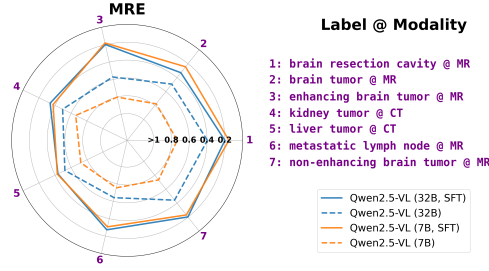


Figure 6: SFT improves T/L size estimation performance.

Table 3: VLM performance on tumor/lesion size estimation tasks. Mean relative error (MRE), success rate (SR), and $MRE_{<k}$ are reported in %, while mean absolute error (MAE) is in millimeters. Improvement after SFT is shown in parentheses.

Model	MAE ↓	MRE ↓	SR ↑	$MRE_{<0.1}$ ↑	$MRE_{<0.2}$ ↑	$MRE_{<0.3}$ ↑
Qwen2.5-VL (7B)	32.3	71.5	100.0	0.3	1.7	3.2
Qwen2.5-VL (32B)	24.2	51.9	100.0	2.7	11.8	19.9
Lingshu (32B)	23.9	66.1	100.0	10.4	29.7	43.4
InternVL3 (38B)	22.6	50.1	100.0	4.5	15.0	24.2
Gemma3 (27B)	30.7	70.7	100.0	1.4	4.9	8.7
MedGemma (4B)	38.9	116.8	100.0	1.1	4.5	7.6
MedGemma (27B)	38.9	116.9	100.0	1.1	4.3	7.5
Llama3.2-Vision (11B)	25.7	61.9	99.1	4.6	14.0	21.1
LLaVA-OneVision (72B)	26.2	83.0	100.0	4.8	17.7	29.4
LLaVA-Med-v1.5 (7B)	48.8	74.7	22.6	0.3	0.6	1.1
MedDr (40B)	30.1	73.8	100.0	1.6	4.7	9.4
HuatuoGPT-Vision (34B)	28.9	89.1	100.0	8.5	26.2	42.3
HealthGPT-L14	23.6	61.3	98.9	8.0	22.6	35.8
Qwen2.5-VL (7B, SFT _{5K})	13.2 (-19.1)	30.6 (-40.9)	100.0	20.8 (+20.5)	41.5 (+39.8)	62.1 (+58.9)
Qwen2.5-VL (32B, SFT _{5K})	12.8 (-11.4)	30.2 (-21.7)	100.0	19.6 (+16.9)	43.2 (+31.4)	63.2 (+43.3)

Conclusion: Although pretrained VLMs can follow instructions to output numerical values in required formats, they fail to estimate T/L size accurately. With SFT, VLMs can reduce MRE from $> 50\%$ to $\sim 30\%$ and show certain generalizability on OOD data.

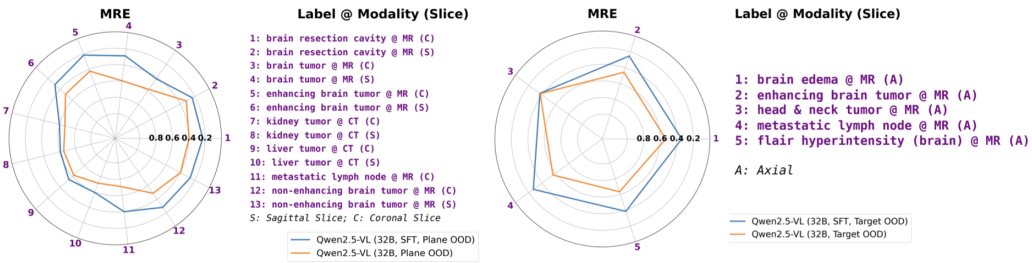


Figure 7: T/L size estimation performance on plane-ODD (left) and target-ODD (right) data.

3.3 A/D MEASUREMENT PERFORMANCE

Baselines vs SFT Models: Baselines perform poorly on A/D measurement tasks, with distance error ranging from 17.3 mm to 81.6 mm and angle error from 29.9° to 66.1° (Table 4). After SFT with 5K samples, the distance error reduces to less than 4.5 mm and angle error to less than 4° . Significant improvement in MRE is shown in Figure 8. The distributions of ground truths and VLM predictions are more closely aligned in the SFT models. (Appendix A.7).

Failure Model Analysis: Baselines perform poorly on A/D measurement tasks, often predicting only a limited set of values (Appendix A.7). As shown in Figure 8, VLMs fail on certain angle

Table 4: VLM performance on angle and distance measurement tasks. Relative errors (RE), $RE_{<0.1}$, and success rate (SR) are reported in %. Absolute errors (AE) are given in millimeters (distance) and degrees (angle).

Model	Ceph-Bio-400								FeTA24			
	Distance				Angle				Distance			
	AE ↓	RE ↓	SR ↑	$RE_{<0.1}$ ↑	AE ↓	RE ↓	SR ↑	$RE_{<0.1}$ ↑	AE ↓	RE ↓	SR ↑	$RE_{<0.1}$ ↑
Qwen2.5-VL (7B)	56.3	80.1	100	0.9	55.1	3691	98.7	3.4	27.5	61.6	100	2.0
Qwen2.5-VL (32B)	59.7	86.0	100	0.2	41.0	3287	100	2.0	19.5	49.6	100	6.0
Lingshu (32B)	17.3	26.1	100	15.5	31.2	4648	100	10.4	38.5	120.9	100	0.0
InternVL3 (38B)	22.1	38.2	100	11.1	66.1	4405	100	4.3	19.7	54.7	100	11.0
Gemma3 (27B)	22.6	32.3	100	22.4	33.6	7862	100	13.8	18.1	40.2	100	15.0
MedGemma (4B)	43.4	67.9	100	6.5	30.5	2780	100	5.8	26.2	60.4	100	5.0
MedGemma (27B)	43.9	68.8	100	6.4	29.9	2168	100	6.4	26.2	60.4	100	5.0
Llama3.2-Vision (11B)	81.6	117.4	100	2.9	55.3	9318.6	100	8.0	29.9	73.4	100	2.0
LLava-OneVision (72B)	36.3	67.8	100	18.8	62.6	12269.1	100	0.0	46.0	130.7	100	4.0
MedDr (40B)	33.7	53.6	100	6.0	54.8	9149.2	100	0.2	37.1	78.5	100	5.0
HuatuoGPT-Vision (34B)	45.8	65.2	100	2.8	65.8	5741.1	100	9.0	40.5	83.3	100	15.0
HealthGPT-L14	66.8	97.5	100	1.1	42.9	<u>434.6</u>	100	0.0	32.1	79.4	100	5.0
Qwen2.5-VL (7B, SFT _{5K})	3.5	5.4	100	86.4	3.6	126.8	100	<u>50.1</u>	4.1	13.1	100	<u>49.0</u>
Qwen2.5-VL (32B, SFT _{5K})	<u>4.5</u>	<u>7.9</u>	100	<u>83.5</u>	<u>4.0</u>	540.2	100	50.5	4.1	<u>14.0</u>	100	54.0

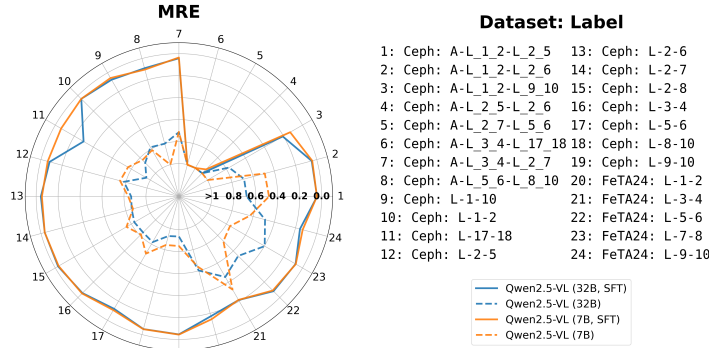


Figure 8: SFT improves VLM performance in A/D size measurement tasks. Labels start with “A” and “L” are angle and distance measurements, respectively. For example, “L-1-2” refers to distance between landmarks 1 and 2.

measurements (labels 4, 5, 6). These correspond to small angles ($< 10^\circ$), which are difficult for models to distinguish and predict accurately.

Conclusion: Pretrained VLMs fail to measure angles and distances accurately. SFT can significantly improve performance, reducing relative error to below 10%. Yet, certain small-angle measurements remain challenging.

4 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 pretrained VLMs perform poorly on these quantitative tasks, whereas supervised fine-tuning with MedVision substantially improves accuracy. Nonetheless, challenges remain, particularly in detecting and measuring small structures. We expect MedVision to provide a valuable foundation for advancing quantitative capabilities in medical VLMs.

5 REPRODUCIBILITY STATEMENT

We will release our entire pipeline of dataset, code, and model checkpoints to facilitate reproducibility. A dataset codebase is provided to reproduce and extend the MedVision dataset.

REFERENCES

- Felix Ambellan, Alexander Tack, Moritz Ehlke, and Stefan Zachow. 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, 2019.
- Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022.
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibbo 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, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- Asma Ben Abacha, Sadid A Hasan, Vivek V Datla, Dina Demner-Fushman, and Henning Müller. 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, 2019.
- Asma Ben Abacha, Mourad Sarrouiti, Dina Demner-Fushman, Sadid A Hasan, and Henning Müller. 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, 2021.
- Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. 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, 2018.
- Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pp. 1–21. Springer, 2022.
- Federico Bolelli, Luca Lumetti, Shankeeth Vinayahalingam, Mattia Di Bartolomeo, Arrigo Pellacani, Kevin Marchesini, Niels Van Nistelrooij, Pieter Van Lierop, Tong Xi, Yusheng Liu, et al. Segmenting the inferior alveolar canal in cbct volumes: the toothfairy challenge. *IEEE Transactions on Medical Imaging*, 2024.
- Federico Bolelli, Kevin Marchesini, Niels van Nistelrooij, Luca Lumetti, Vittorio Pipoli, Elisa Ficarra, Shankeeth Vinayahalingam, and Costantino Grana. Segmenting maxillofacial structures in cbct volumes. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5238–5248, 2025.
- Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P Langlotz. Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats. *arXiv preprint arXiv:2405.19538*, 2024.

- Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, Xiang Wan, and Benyou Wang. Huatuoqpt-vision, towards injecting medical visual knowledge into multimodal llms at scale, 2024. URL <https://arxiv.org/abs/2406.19280>.
- 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, Farzan Moodi, Teresa Nunes, Aaron Simon, Nico Sollmann, David Vu, Maruf Adewole, Jake Albrecht, Udunna Anazodo, Rongrong Chai, Verena Chung, Shahriar Faghani, Keyvan Farahani, Anahita Fathi Kazerooni, Eugenio Iglesias, Florian Kofler, Hongwei Li, Marius George Lingurar, Bjoern Menze, Ahmed W. Moawad, Yuri Velichko, Benedikt Wiestler, Talissa Altes, Patil Basavasagar, Martin Bendszus, Gianluca Brugnara, Jaeyoung Cho, Yaseen Dhemesh, Brandon K. K. Fields, Filip Garrett, Jaime Gass, Lubomir Hadjiiski, Jona Hattangadi-Gluth, Christopher Hess, Jessica L. Houk, Edwin Isufi, Lester J. Layfield, George Mastorakos, John Mongan, Pierre Nedelec, Uyen Nguyen, Sebastian Oliva, Matthew W. Pease, Aditya Rastogi, Jason Sinclair, Robert X. Smith, Leo P. Sugrue, Jonathan Thacker, Igor Vidic, Javier Villanueva-Meyer, Nathan S. White, Mariam Aboian, Gian Marco Conte, Anders Dale, Mert R. Sabuncu, Tyler M. Seibert, Brent Weinberg, Aly Abayazeed, Raymond Huang, Sevcen Turk, Andreas M. Rauschecker, Nikdokht Farid, Philipp Vollmuth, Ayman Nada, Spyridon Bakas, Evan Calabrese, and Jeffrey D. Rudie. The 2024 brain tumor segmentation (brats) challenge: Glioma segmentation on post-treatment mri, 2024. URL <https://arxiv.org/abs/2405.18368>.
- Arjun D Desai, Andrew M Schmidt, Elka B Rubin, Christopher Michael Sandino, Marianne Susan Black, Valentina Mazzoli, Kathryn J Stevens, Robert Boutin, Christopher Re, Garry E Gold, Brian Hargreaves, and Akshay Chaudhari. SKM-TEA: A dataset for accelerated MRI reconstruction with dense image labels for quantitative clinical evaluation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=YDMFgD_qJuA.
- 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, Smriti Joshi, Victoriya Kashtanova, Hyeon Gyu Kim, Satoshi Kondo, Christian N. Kruse, Susana K. Lai-Yuen, Hao Li, Han Liu, Buntheng Ly, Ipek Oguz, Hyungseob Shin, Boris Shirokikh, Zixian Su, Guotai Wang, Jianghao Wu, Yanwu Xu, Kai Yao, Li Zhang, Sébastien Ourselin, Jonathan Shapey, and Tom Vercauteren. Crossmoda 2021 challenge: Benchmark of cross-modality domain adaptation techniques for vestibular schwannoma and cochlea segmentation. *Medical Image Analysis*, 83:102628, January 2023. ISSN 1361-8415. doi: 10.1016/j.media.2022.102628. URL <http://dx.doi.org/10.1016/j.media.2022.102628>.
- Sunan He, Yuxiang Nie, Hongmei Wang, Shu Yang, Yihui Wang, Zhiyuan Cai, Zhixuan Chen, Yingxue Xu, Luyang Luo, Huiling Xiang, Xi Lin, Mingxiang Wu, Yifan Peng, George Shih, Ziyang Xu, Xian Wu, Qiong Wang, Ronald Cheong Kin Chan, Varut Vardhanabhuti, Winnie Chiu Wing Chu, Yefeng Zheng, Pranav Rajpurkar, Kang Zhang, and Hao Chen. Gsco: Towards generalizable ai in medicine via generalist-specialist collaboration, 2024. URL <https://arxiv.org/abs/2404.15127>.
- Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering, 2020a. URL <https://arxiv.org/abs/2003.10286>.
- Yuting He, Guanyu Yang, Jian Yang, Yang Chen, Youyong Kong, Jiasong Wu, Lijun Tang, Xiaomei Zhu, Jean-Louis Dillenseger, Pengfei Shao, et al. Dense biased networks with deep priori anatomy and hard region adaptation: Semi-supervised learning for fine renal artery segmentation. *Medical image analysis*, 63:101722, 2020b.
- Yuting He, Guanyu Yang, Jian Yang, Rongjun Ge, Youyong Kong, Xiaomei Zhu, Shaobo Zhang, Pengfei Shao, Huazhong Shu, Jean-Louis Dillenseger, et al. Meta grayscale adaptive network for 3d integrated renal structures segmentation. *Medical image analysis*, 71:102055, 2021.

- 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, Ed Walczak, Sean McSweeney, Ranveer Vasdev, Chris Hornung, Rafat Solaiman, Jamee Schoepfoerster, Bailey Abernathy, David Wu, Safa Abdulkadir, Ben Byun, Justice Spriggs, Griffin Struyk, Alexandra Austin, Ben Simpson, Michael Hagstrom, Sierra Virnig, John French, Nitin Venkatesh, Sarah Chan, Keenan Moore, Anna Jacobsen, Susan Austin, Mark Austin, Subodh Regmi, Nikolaos Papanikolopoulos, and Christopher Weight. The kits21 challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase ct, 2023.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL <https://arxiv.org/abs/2106.09685>.
- Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 590–597, 2019.
- Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in neural information processing systems*, 35:36722–36732, 2022.
- Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.
- Maxime Kayser, Cornelius Emde, Oana-Maria Camburu, Guy Parsons, Bartłomiej Papiez, and Thomas Lukasiewicz. Explaining chest x-ray pathologies in natural language. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 701–713. Springer, 2022.
- 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, Ariana Familiar, Keyvan Farahani, Andrea Franson, Anurag Gottipati, Shuvanjan Haldar, Juan Eugenio Iglesias, Anastasia Janas, Elaine Johansen, Blaise V Jones, Neda Khalili, Florian Kofler, Dominic LaBella, Hollie Anne Lai, Koen Van Leemput, Hongwei Bran Li, Nazanin Maleki, Aaron S McAllister, Zeke Meier, Bjorn Menze, Ahmed W Moawad, Khanak K Nandolia, Julija Pavaine, Marie Piraud, Tina Pous-saint, Sanjay P Prabhu, Zachary Reitman, Jeffrey D Rudie, Mariana Sanchez-Montano, Ibraheem Salman Shaikh, Nakul Sheth, Wenxin Tu, Chunhao Wang, Jeffrey B Ware, Benedikt Wiestler, Anna Zapaishchykova, Miriam Bornhorst, Michelle Deutsch, Maryam Fouladi, Margot Lazow, Leonie Mikael, Trent Hummel, Benjamin Kann, Peter de Blank, Lindsey Hoffman, Mariam Aboian, Ali Nabavizadeh, Roger Packer, Spyridon Bakas, Adam Resnick, Brian Rood, Arastoo Vossough, and Marius George Linguraru. The brain tumor segmentation in pediatrics (brats-peds) challenge: Focus on pediatrics (cbtn-connect-dipgr-asnr-miccai brats-peds), 2024. URL <https://arxiv.org/abs/2404.15009>.
- 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, Victor Alves, Jan Egger, Daniel Capellán-Martín, Abhijeet Parida, Austin Tapp, Xinyang Liu, Maria J. Ledesma-Carbayo, Jay B. Patel, Thomas N. McNeal, Maya Viera, Owen McCall, Albert E. Kim, Elizabeth R. Gerstner, Christopher P. Bridge, Katherine Schumacher, Michael Mix, Kevin Leu, Shan McBurney-Lin, Pierre Nedelec, Javier Villanueva-Meyer, David R. Raleigh, Jonathan Shapey, Tom Vercauteren, Kazumi Chia, Marina Ivory, Theodore Barfoot, Omar Al-Salihi, Justin Leu, Lia M. Halasz, Yuri S. Velichko, Chunhao Wang, John P. Kirkpatrick, Scott R. Floyd, Zachary J. Reitman, Trey C. Mullikin, Eugene J. Vaios, Christina Huang, Ulas Bagci, Sean Sachdev, Jona A. Hattangadi-Gluth,

- Tyler M. Seibert, Nikdokht Farid, Connor Puett, Matthew W. Pease, Kevin Shiue, Syed Muhammad Anwar, Shahriar Faghani, Peter Taylor, Pranav Warman, Jake Albrecht, András Jakab, Mana Moassefi, Verena Chung, Rong Chai, Alejandro Aristizabal, Alexandros Karargyris, Hasan Kassem, Sarthak Pati, Micah Sheller, Nazanin Maleki, Rachit Saluja, Florian Kofler, Christopher G. Schwarz, Philipp Lohmann, Philipp Vollmuth, Louis Gagnon, Maruf Adewole, Hongwei Bran Li, Anahita Fathi Kazerooni, Nourel Hoda Tahon, Udunna Anazodo, Ahmed W. Moawad, Bjoern Menze, Marius George Linguraru, Mariam Aboian, Benedikt Wiestler, Ujjwal Baid, Gian-Marco Conte, Andreas M. Rauschecker, Ayman Nada, Aly H. Abayazeed, Raymond Huang, Maria Correia de Verdier, Jeffrey D. Rudie, Spyridon Bakas, and Evan Calabrese. Analysis of the 2024 brats meningioma radiotherapy planning automated segmentation challenge, 2025. URL <https://arxiv.org/abs/2405.18383>.
- Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018.
- Sarah Leclerc, Erik Smistad, Joao Pedrosa, Andreas Østvik, Frederic Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Pierre-Marc Jodoin, Thomas Grenier, et al. Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. *IEEE transactions on medical imaging*, 38(9):2198–2210, 2019.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024a. URL <https://arxiv.org/abs/2408.03326>.
- Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*, 2023.
- Wenxuan Li, Chongyu Qu, Xiaoxi Chen, Pedro RAS Bassi, Yijia Shi, Yuxiang Lai, Qian Yu, Huimin Xue, Yixiong Chen, Xiaorui Lin, et al. Abdomenatlas: A large-scale, detailed-annotated, & multi-center dataset for efficient transfer learning and open algorithmic benchmarking. *Medical Image Analysis*, pp. 103285, 2024b. URL <https://github.com/MrGiovanni/AbdomenAtlas>.
- Tianwei Lin, Wenqiao Zhang, Sijing Li, Yuqian Yuan, Binhe Yu, Haoyuan Li, Wanggui He, Hao Jiang, Mengze Li, Xiaohui Song, Siliang Tang, Jun Xiao, Hui Lin, Yueting Zhuang, and Beng Chin Ooi. Healthgpt: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation, 2025. URL <https://arxiv.org/abs/2502.09838>.
- Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents, 2023. URL <https://arxiv.org/abs/2303.07240>.
- Claudia Lindner, Ching-Wei Wang, Cheng-Ta Huang, Chung-Hsing Li, Sheng-Wei Chang, and Tim F Cootes. Fully automatic system for accurate localisation and analysis of cephalometric landmarks in lateral cephalograms. *Scientific reports*, 6(1):33581, 2016.
- Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering, 2021. URL <https://arxiv.org/abs/2102.09542>.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.
- Luca Lumetti, Vittorio Pipoli, Federico Bolelli, Elisa Ficarra, and Costantino Grana. Enhancing patch-based learning for the segmentation of the mandibular canal. *IEEE Access*, 12:79014–79024, 2024.
- Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyan Wang, Xin Liu, Shucheng Cao, Qi Zhang, Shangqing Liu, Yunpeng Wang, Yuhui Li, Jian He, and Xiaoping Yang. Abdomenct-1k: Is abdominal organ segmentation a solved problem? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6695–6714, 2022. doi: 10.1109/TPAMI.2021.3100536.

- Jun Ma, Yao Zhang, Song Gu, Cheng Ge, Shihao Mae, Adamo Young, Cheng Zhu, Xin Yang, Kangkang Meng, Ziyang Huang, Fan Zhang, Yuanke Pan, Shoujin Huang, Jiacheng Wang, Mingze Sun, Rongguo Zhang, Dengqiang Jia, Jae Won Choi, Natália Alves, Bram de Wilde, Gregor Koehler, Haoran Lai, Ershuai Wang, Manuel Wiesenfarth, Qiongjie Zhu, Guoqiang Dong, Jian He, Junjun He, Hua Yang, Bingding Huang, Mengye Lyu, Yongkang Ma, Heng Guo, Weixin Xu, Klaus Maier-Hein, Yajun Wu, and Bo Wang. 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, 2024. ISSN 2589-7500. doi: [https://doi.org/10.1016/S2589-7500\(24\)00154-7](https://doi.org/10.1016/S2589-7500(24)00154-7). URL <https://www.sciencedirect.com/science/article/pii/S2589750024001547>.
- Kelly Payette, Hongwei Bran Li, Priscille De Dumast, Roxane Licandro, Hui Ji, Md Mahfuzur Rahman Siddiquee, Daguang Xu, Andriy Myronenko, Hao Liu, Yuchen Pei, et al. Fetal brain tissue annotation and segmentation challenge results. *Medical image analysis*, 88:102833, 2023.
- Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cian Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, Lu Yang, Kejia Chen, Per Björnsson, Shashir Reddy, Ryan Brush, Kenneth Philbrick, Mercy Asiedu, Ines Mezerreg, Howard Hu, Howard Yang, Richa Tiwari, Sunny Jansen, Preeti Singh, Yun Liu, Shekoofeh Azizi, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Riviere, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Elena Buchatskaya, Jean-Baptiste Alayrac, Dmitry Lepikhin, Vlad Feinberg, Sebastian Borgeaud, Alek Andreev, Cassidy Hardin, Robert Dadashi, Léonard Hussenot, Armand Joulin, Olivier Bachem, Yossi Matias, Katherine Chou, Avinatan Hassidim, Kavi Goel, Clement Farabet, Joelle Barral, Tris Warkentin, Jonathon Shlens, David Fleet, Victor Cotruta, Omar Sanseviero, Gus Martins, Phoebe Kirk, Anand Rao, Shravya Shetty, David F. Steiner, Can Kirmizibayrak, Rory Pilgrim, Daniel Golden, and Lin Yang. Medgemma technical report, 2025. URL <https://arxiv.org/abs/2507.05201>.
- Pengfei Shao, Chao Qin, Changjun Yin, Xiaoxin Meng, Xiaobing Ju, Jie Li, Qiang Lv, Wei Zhang, and Zhengquan Xu. Laparoscopic partial nephrectomy with segmental renal artery clamping: technique and clinical outcomes. *European urology*, 59(5):849–855, 2011.
- Pengfei Shao, Lijun Tang, Pu Li, Yi Xu, Chao Qin, Qiang Cao, Xiaobing Ju, Xiaoxin Meng, Qiang Lv, Jie Li, et al. Precise segmental renal artery clamping under the guidance of dual-source computed tomography angiography during laparoscopic partial nephrectomy. *European urology*, 62(6):1001–1008, 2012.
- Sanjay Subramanian, Lucy Lu Wang, Sachin Mehta, Ben Bogin, Madeleine Van Zuylen, Sravanthi Parasa, Sameer Singh, Matt Gardner, and Hannaneh Hajishirzi. Mediat: A dataset of medical images, captions, and textual references. *arXiv preprint arXiv:2010.06000*, 2020.
- 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, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xi-aohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Naveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Pettrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huiizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian

- Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wi-
eting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh,
Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine,
Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael
Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Ni-
lay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Ruben-
stein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya
Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu,
Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti
Sheth, Siim Pöder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi
Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry,
Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein
Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat
Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas
Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Bar-
ral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam
Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena
Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier
Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot.
Gemma 3 technical report, 2025a. URL <https://arxiv.org/abs/2503.19786>.
- 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. Lingshu: A generalist
foundation model for unified multimodal medical understanding and reasoning, 2025b. URL
<https://arxiv.org/abs/2506.07044>.
- Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck,
Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. Totalseg-
mentator: robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial
Intelligence*, 5(5):e230024, 2023.
- Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. Towards generalist
foundation model for radiology by leveraging web-scale 2d&3d medical data. *Nature Communi-
cations*, 16(1):7866, 2025.
- 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 Wald-
mannstetter, Florian Kofler, Fernando Navarro, Martin Menten, Ivan Ezhov, Daniel Rueckert,
Iris N. Vos, Ynte M. Ruigrok, Birgitta K. Velthuis, Hugo J. Kuijf, Pengcheng Shi, Wei Liu, Ting
Ma, Maximilian R. Rokuss, Yannick Kirchhoff, Fabian Isensee, Klaus Maier-Hein, Chengcheng
Zhu, Huilin Zhao, Philippe Bijlenga, Julien Hämmerli, Catherine Wurster, Laura Westphal, Jeroen
Bisschop, Elisa Colombo, Hakim Baazaoui, Hannah-Lea Handelsmann, Andrew Makmur, James
Hallinan, Amrith Soundararajan, Bene Wiestler, Jan S. Kirschke, Roland Wiest, Emmanuel Mon-
tagnon, Laurent Letourneau-Guillon, Kwanseok Oh, Dahye Lee, Adam Hilbert, Orhun Utku
Aydin, Dimitrios Rallios, Jana Rieger, Satoru Tanioka, Alexander Koch, Dietmar Frey, Abdul
Qayyum, Moona Mazher, Steven Niederer, Nico Disch, Julius Holzschuh, Dominic LaBella,
Francesco Galati, Daniele Falcetta, Maria A. Zuluaga, Chaolong Lin, Haoran Zhao, Zehan Zhang,
Minghui Zhang, Xin You, Hanxiao Zhang, Guang-Zhong Yang, Yun Gu, Sinyoung Ra, Jongyun
Hwang, Hyunjin Park, Junqiang Chen, Marek Wodzinski, Henning Müller, Nesrin Mansouri,
Florent Autrusseau, Cansu Yalçın, Rachika E. Hamadache, Clara Lisazo, Joaquim Salvi, Adrià
Casamitjana, Xavier Lladó, Uma Maria Lal-Trehan Estrada, Valeriia Abramova, Luca Giancardo,
Arnau Oliver, Paula Casademunt, Adrian Galdran, Matteo Delucchi, Jialu Liu, Haibin Huang,
Yue Cui, Zehang Lin, Yusheng Liu, Shunzhi Zhu, Tatsat R. Patel, Adnan H. Siddiqui, Vin-
cent M. Tutino, Maysam Orouskhani, Huayu Wang, Mahmud Mossa-Basha, Yuki Sato, Sven
Hirsch, Susanne Wegener, and Bjoern Menze. Benchmarking the cow with the topcow challenge:
Topology-aware anatomical segmentation of the circle of willis for cta and mra, 2025. URL
<https://arxiv.org/abs/2312.17670>.

- Yongcheng Yao and Weitian Chen. Quantifying knee cartilage shape and lesion: From image to metrics. In Shandong Wu, Behrouz Shabestari, and Lei Xing (eds.), *Applications of Medical Artificial Intelligence*, pp. 162–172, Cham, 2025. Springer Nature Switzerland.
- Yongcheng Yao, Junru Zhong, Liping Zhang, Sheheryar Khan, and Weitian Chen. Cartimorph: A framework for automated knee articular cartilage morphometrics. *Medical Image Analysis*, 91: 103035, 2024. ISSN 1361-8415. doi: <https://doi.org/10.1016/j.media.2023.103035>.
- Tianhong Zhou, Yin Xu, Yingtao Zhu, Chuxi Xiao, Haiyang Bian, Lei Wei, and Xuegong Zhang. Drvd-bench: Do vision-language models reason like human doctors in medical image diagnosis?, 2025. URL <https://arxiv.org/abs/2505.24173>.
- 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, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL <https://arxiv.org/abs/2504.10479>.
- Yongshuo Zong, Oisín Mac Aodha, and Timothy M. Hospedales. Self-supervised multimodal learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

A APPENDIX

A.1 DATASET EXAMPLES

Figure 9 shows preprocessed images of each datasets and example of three types of quantitative annotations.

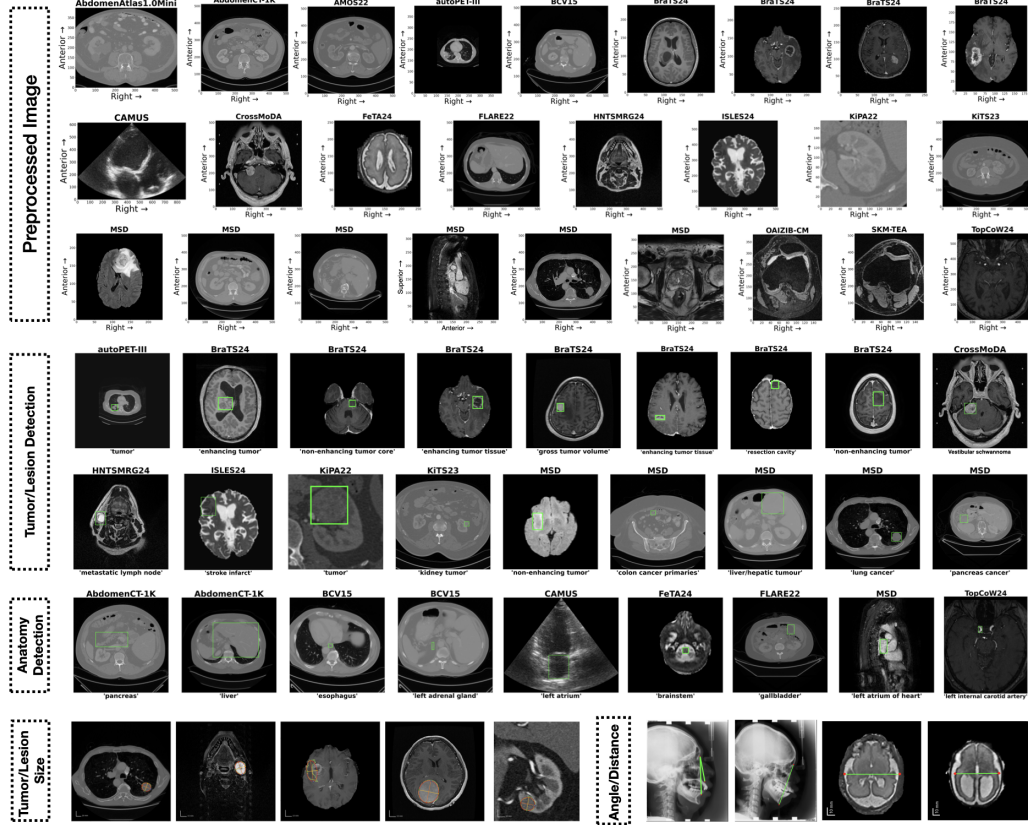


Figure 9: Example preprocessed images and annotations from MedVision dataset.

A.2 VLMS VERSION

Table 5 lists the HuggingFace model IDs of the evaluated open-weight VLMS.

A.3 PROMPT TEMPLATES

We designed task-specific prompts to guide the models in generating relevant outputs. Prompts were crafted to be clear and concise, providing necessary context without overwhelming the model.

Detection Task Prompt. For detection tasks, the prompt was designed to instruct the model to identify and localize specific anatomical structures or lesions within the image. The model is expected to return the normalized coordinates of the lower-left and upper-right corners of the bounding box. The prompt template is as follows:

Task:
Given the input medical image: {image_description},
return the coordinates of the lower-left and upper-right corner

Table 5: HuggingFace model ID of the evaluated open-weight VLMs.

Model	HuggingFace Model ID
Qwen2.5-VL (7B)	Qwen/Qwen2.5-VL-7B-Instruct
Qwen2.5-VL (32B)	Qwen/Qwen2.5-VL-32B-Instruct
Lingshu (32B)	lingshu-medical-mlm/Lingshu-32B
InternVL3 (38B)	OpenGVLab/InternVL3-38B
Gemma3 (27B)	google/gemma-3-27b-it
MedGemma (4B)	google/medgemma-4b-it
MedGemma (27B)	google/medgemma-27b-it
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	lintw/HealthGPT-L14

of the bounding box for the {label_name}.

Format requirement:

{format_prompt}.

The format prompt for this task is as follows:

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.

Tumor/Lesion Size Measurement Prompt. For tumor/lesion size measurement tasks, the prompt instructs the model to measure the longest and shortest diameters of the specified tumor/lesion. The model should return the lengths in millimeters. The prompt template is as follows:

Task:

Given the input medical image: {image_description}, estimate the major and minor axis lengths of the ellipse enclosing the {label_name}, in {metric_unit}.

Additional information:

{pixel_size_prompt}.

Format requirement:

{format_prompt}.

where we provide the pixel size information in {pixel_size_prompt} following the format:

The pixel size for this image is {adjusted_pixel_width} (width) x {adjusted_pixel_height} (height).

Note that the pixel size is adjusted according to the model’s image processing strategy. The format prompt for this task is as follows:

The answer should be two decimal numbers separated by a comma

without any units or additional text. The first is the major axis length, and the second is the minor axis length.

Angle/Distance Measurement Prompt. For angle/distance measurement tasks, the prompt instructs the model to measure specific anatomical angles or distances based on provided landmark names. The model should return the measurement in the specified unit. The prompt template is as follows:

```
Task:
Given the input medical image: {image_description},
{task_prompt}
Additional information:
{pixel_size_prompt}
Format requirement:
{format_prompt}.
```

An example of {task_prompt} for distance measurement is:

```
estimate the distance of {distance_name} in {metric_unit},
which is the distance between 2 landmark points:
(landmark 1) {p1},
(landmark 2) {p2}.
```

An example of {task_prompt} for angle measurement is:

```
estimate the angle of {angle_name} in {metric_unit},
which is the angle between 2 lines:
(line 1) the line connecting {l1p1} and {l1p2},
(line 2) the line connecting {l2p1} and {l2p2}.
```

The format prompt for this task is as follows:

```
The answer should be a single decimal number
without any units or additional text.
```

A.4 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 6, upon which we designed appropriate pixel size adjustment for conveying physical spacing information to VLMs.

A.5 DETAILS OF DETECTION PERFORMANCE

A.5.1 ANATOMY LEVEL DETECTION PERFORMANCE

Figure 10 shows the detection performance of VLMs in various anatomical groups. Labels are first categorized into these anatomy groups. Then, weighted average metrics are calculated for each group. Tumor/lesion labels are highlighted in color.

A substantial improvement is observed in the SFT models across all metrics. In Figure 10, anatomical groups are arranged in descending order of recall. Overall, detection performance is lower for tumor/lesion groups than for healthy anatomy, likely due to greater variability in tumor size, shape, and location. The limited number of tumor/lesion samples may also contribute to reduced performance.

Table 6: Image resize strategies of VLMs. ‡ aspect ratio preserved

Model	Image Size	Image Resize Strategy
Qwen2.5-VL	dynamic	reshape to a size divisible by 28x28 patch
Lingshu	dynamic	reshape to a size divisible by 28x28 patch
InternVL3	448x448	reshape to a fixed size
Gemma3	896x896	reshape to a fixed size
MedGemma	896x896	reshape to a fixed size
Llama3.2-Vision	dynamic	reshape [‡] to fit in a tiled canvas (with patches of size 560x560)
LLaVA-OneVision	dynamic	reshape [‡] to fit in a tiled canvas (with patches of size 384x384)
LLaVA-Med	336x336	reshape to a fixed size
MedDr	448x448	reshape to a fixed size
HuatuoGPT-Vision	336x336	reshape to a fixed size
HealthGPT-L14	336x336	reshape to a fixed size

[illegible]

Figure 10: Detection performance of VLMs at anatomy level. Tumor/lesion labels are highlighted in color.

A.5.2 LABEL LEVEL DETECTION PERFORMANCE

Figure 11 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 11. For each box-image-ratio group and label combination, we visualized recall, precision, and F1 score. Figure 11 is a detailed map of VLM performance at various target sizes and labels. We observed that detection performance generally improves with larger box-image-ratio. Significant performance improvement is observed in the SFT models.

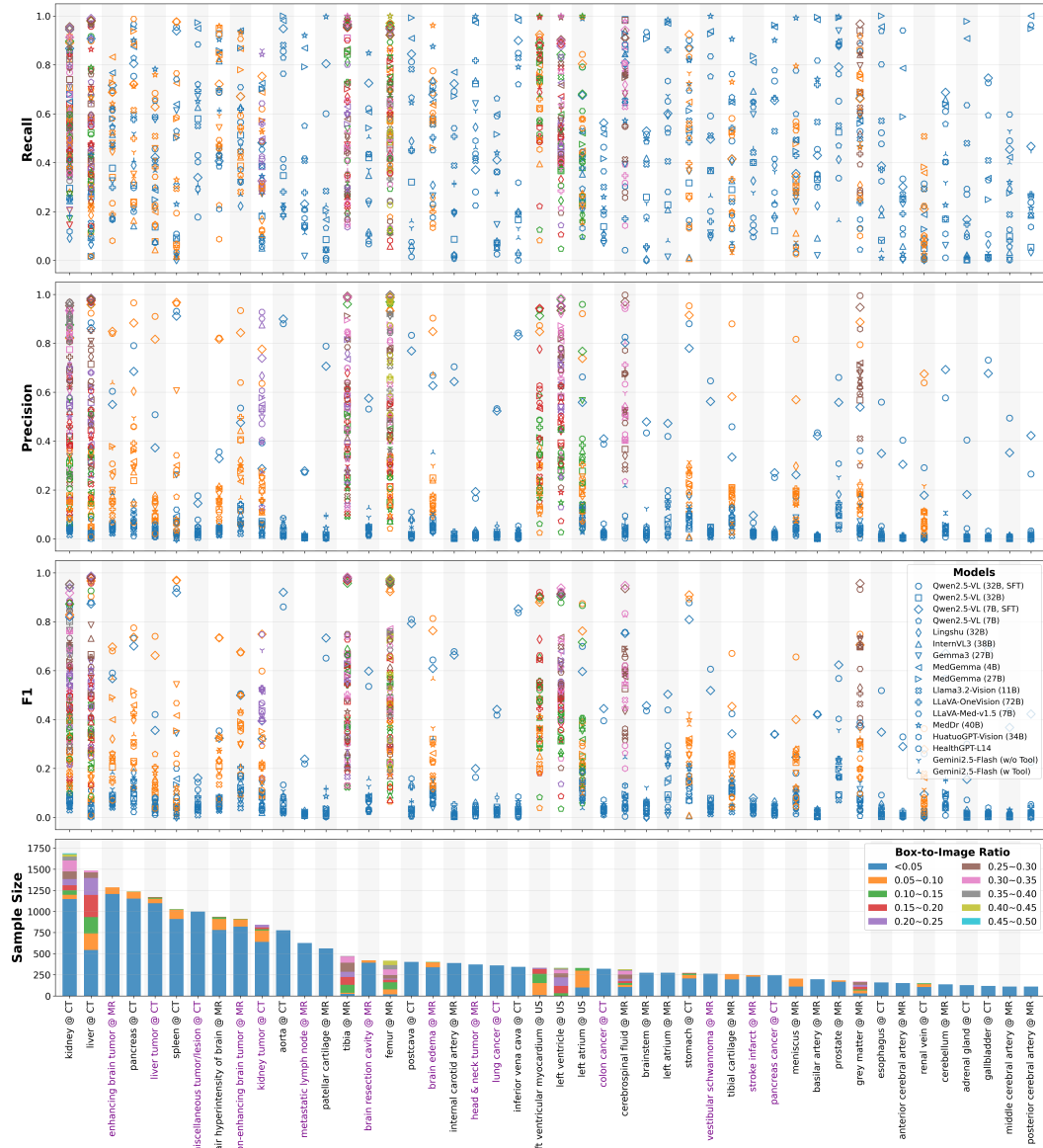


Figure 11: Label level detection performance at each box-image-ratio group.

A.5.3 EFFECT OF TARGET SIZE

Figure 11 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 12, precision and recall is positively correlated to relative target size (box-image-ratio). The SFT models consistently outperform the baselines across all target sizes.

VLMs vs Random Guess Model: We compared the detection performance of baselines and SFT VLMs with a random guess model. Most pretrained models perform slightly better than random guess, while the medical VLM HuatuoGPT-Visoin (34B) standouts from the rest baselines.

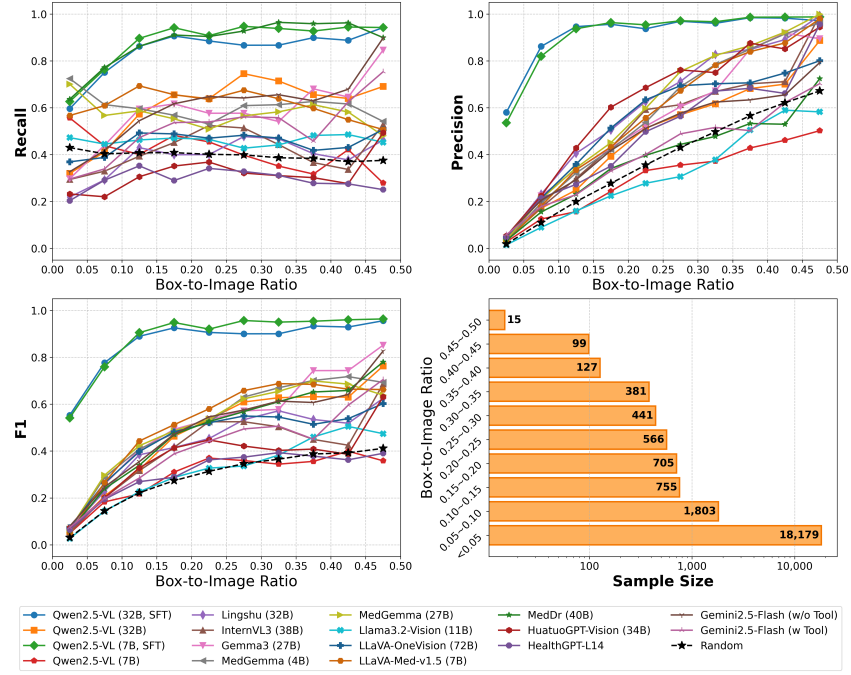


Figure 12: Effect of target size, quantified by box-image-ratio, on detection performance of VLMs.

A.6 DETAILS OF T/L SIZE ESTIMATION PERFORMANCE

Figure 13 shows the distribution of ground truths and VLM predictions in tumor/lesion size estimation tasks.

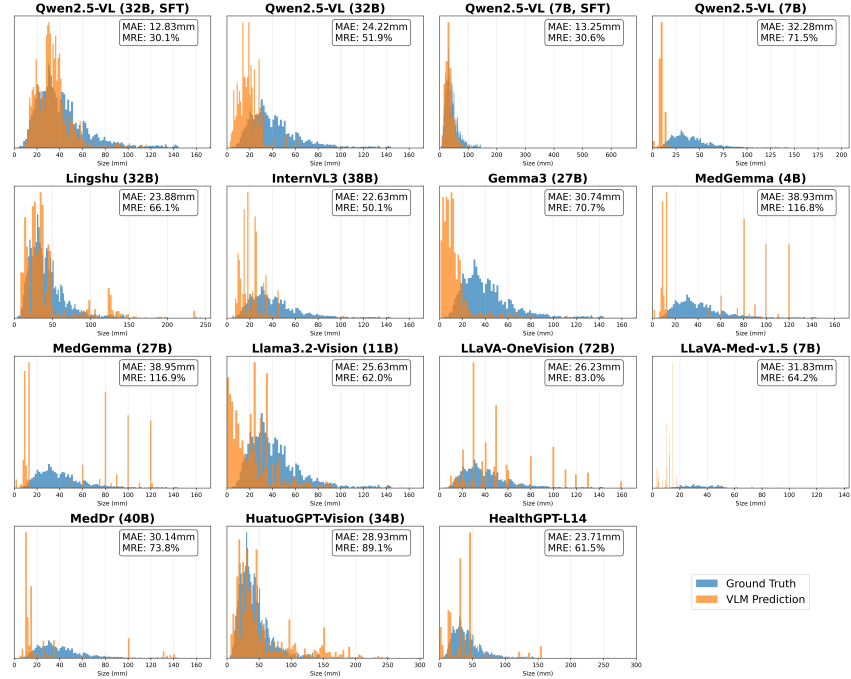


Figure 13: Distribution of ground truths and VLM predictions in tumor/lesion size estimation tasks.

A.7 DETAILS OF A/D MEASUREMENT PERFORMANCE

Figure 14 and Figure 15 show the distribution of ground truths and VLM predictions in angle and distance measurement tasks, respectively. We observed a common behavior among baselines where VLMs tend to predict a limited set of values, shown as spikes in the distribution plots. With SFT, the diversity of model predictions increases and better aligns with the ground truth distribution.

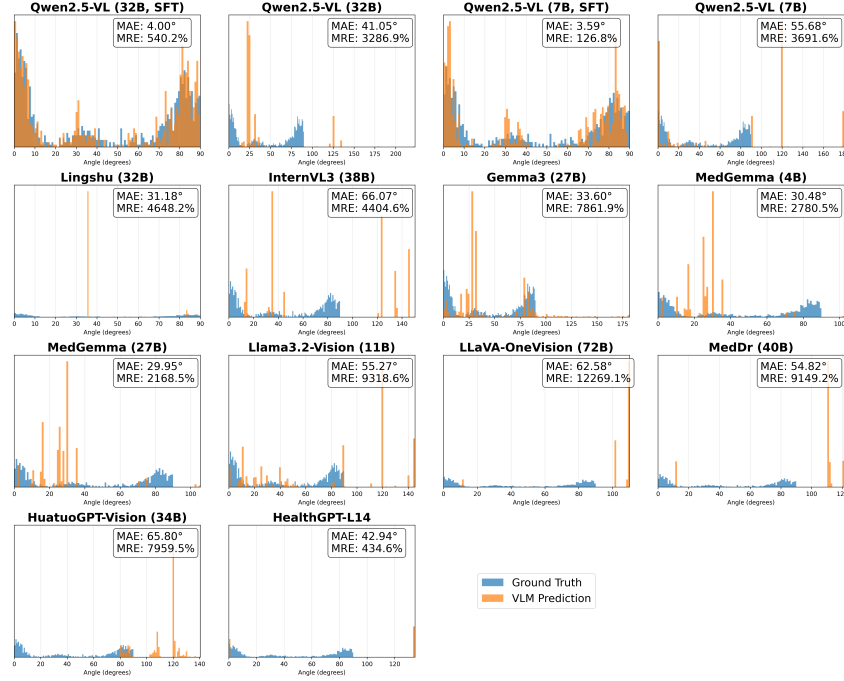


Figure 14: Distribution of ground truths and VLM predictions in angle measurement tasks.

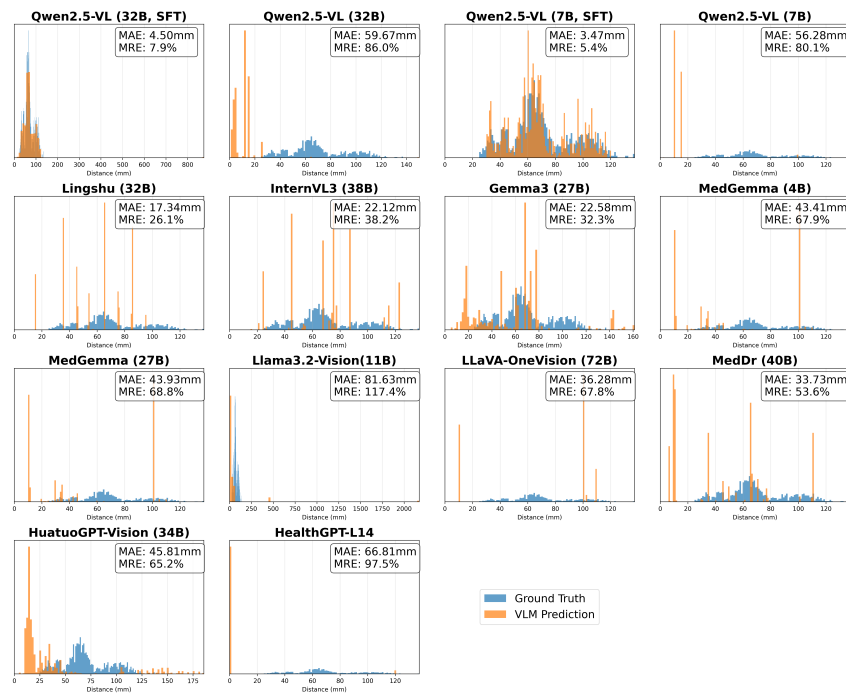


Figure 15: Distribution of ground truths and VLM predictions in distance measurement tasks.