MELBA Journal Machine Learning for Biomedical Imaging

CT-Bench: A Comprehensive Benchmark for Multimodal AI in Computed Tomography Analysis

Qingqing Zhu [1,†], Qiao Jin [1,†], Tejas S. Mathai [1], Yin Fang [1], Zhizheng Wang [1], Yifan Yang [1], Maame Sarfo-Gyamfi [1,2], Benjamin Hou [1], Ran Gu [1], Praveen T. S. Balamuralikrishna [1], Kenneth C. Wang [3], Ronald M. Summers [1], Zhiyong Lu [1,*]

1 National Institutes of Health, Bethesda, MD, USA
2 Howard University College of Medicine, Washington, D.C., USA
3 Radiology, U.S. Department of Veterans Affairs, USA
† Equal contribution

Abstract

Publicly available computed tomography (CT) resources with lesion-level image, spatial, and textual annotations remain limited, restricting reproducible development and evaluation of multimodal medical AI. We present **CT-Bench**, a CT lesion imagetext resource and visual question answering benchmark constructed by extending DeepLesion with curated lesion-level text, standardized lesion-size representations, structured attributes, and QA tasks. CT-Bench contains **20,335 lesion instances** from **7,795 CT studies** and **3,793 patients**, including CT key-slice images, bounding-box annotations, PACS-derived lesion descriptions, lesion size measurements, and associated metadata. The resource also includes a multiple-choice QA benchmark with **2,850 question-answer pairs** across seven lesion analysis tasks: image-to-description matching, multi-slice CT-to-description matching, description-to-image retrieval, description-to-bounding-box localization, lesion size estimation, image-based attribute recognition, and multi-slice CT attribute recognition. Hard negative answer choices are included to evaluate fine-grained image-text alignment and lesion-level reasoning. The resource is available at https://kaggle.com/datasets/cd1661d6d6aeab08b8eb99b58885b4489d76fc5ac07d5aac76cae577e6426e2f with documentation, metadata files, and reproducibility code. Validation includes annotation review by trained annotators and medical experts, hard-negative verification, human-reader comparison, and baseline experiments with general and medical vision-language models.

Keywords

computed tomography, open data, medical image computing, multimodal learning, lesion analysis, visual question answering

Article informations

https://doi.org/10.59275/j.melba.2026-b141

Volume 2026, Received: 2026-06-19, Published 2026-09-21

Corresponding author: zhiyong.lu@nih.gov

Special issue: MICCAI Open Data 2026 x MELBA
Guest editors: AT, MS et al.

1. Background

Computed tomography (CT) is widely used in clinical diagnosis, treatment planning, and disease monitoring across multiple anatomical regions. In routine radiology practice, many CT findings are interpreted at the lesion level: radiologists localize a lesion, assess its appearance across neighboring slices, measure its size, and describe its anatomical context and imaging characteristics (Yan et al., 2017; Zhu et al., 2025). Multimodal AI systems that connect CT images with lesion-level text could support research on lesion localization, image-text retrieval, structured lesion description, size estimation, and CT visual question answering (VQA) (Sun et al., 2023). However, the development and reproducible evaluation of such systems require datasets in which CT images are paired with spatial annotations, lesion-specific textual descriptions, size information, and metadata (Tian et al., 2024).

Existing public resources only partially meet these needs. DeepLesion Yan et al. (2017)provides a large collection of CT lesion key slices with bounding boxes, but it does not include lesion-specific textual descriptions or structured size descriptions aligned with each lesion. CT-RATE provides volumetric chest CT studies paired with radiology reports and supports CT-report learning, but it focuses on thoracic CT and does not provide 2D key-slice lesion-level annotations with bounding boxes and per-lesion descriptions (Hamamci et al., 2026). Multimodal biomedical image-text resources such as ROCOv2 (Rückert et al., 2024), PMC-OA (Lin et al., 2023), and MedICaT (Subramanian et al., 2020) contain captions from scientific articles, which are useful for broad image-text pretraining but are not directly derived from PACS lesion bookmarks or original clinical lesion measurements . In chest radiography, resources such as MIMIC-CXR demonstrate the value of paired imaging and clinical text, but an analogous lesion-level CT resource with spatial and textual annotations remains limited (Johnson et al., 2019; Zhu et al., 2023).

Related limitations also exist in medical VQA benchmarks. VQA-RAD (Lau et al., 2018), PathVQA (He et al., 2020), SLAKE (Liu et al., 2021), VQA-Med (Ben Abacha et al., 2021), and PMC-VQA (Zhang et al., 2023b) have enabled evaluation of image-question answering in radiology, pathology, and mixed biomedical imaging. However, many existing VQA resources are relatively small, focus on modalities other than CT, use open-ended question formats that complicate standardized scoring, or do not emphasize lesion-level CT reasoning with bounding boxes and hard negative choices (Jin et al., 2024). CT interpretation also differs from many 2D natural-image or single-projection medical-image tasks because clinically relevant information may depend on cross-sectional anatomy, neighboring slices, lesion size, and subtle differences among visually similar findings.

CT-Bench is an extension of DeepLesion rather than a CT imaging dataset. It retains DeepLesion key-slice images, lesion bookmarks, bounding boxes, measurements, and metadata, and adds curated PACS-derived lesion descriptions, standardized lesion-size representations, ontology-derived attributes, and hard-negative QA tasks. These additions transform DeepLesion into an AI-ready resource for reproducible multimodal lesion-level CT research. A concise comparison of the components inherited from DeepLesion and those added in CT-Bench is provided in Appendix Table S1.

2. Summary

CT-Bench was developed to support reproducible research on multimodal AI for lesion-level CT interpretation. As summarized in Figure 1, the resource has two complementary components: the CT-Bench Lesion Image and Metadata Set, which links CT lesion images with bounding boxes, PACS-derived lesion descriptions, lesion size measurements, and associated metadata; and the CT-Bench QA Benchmark, which provides multiple-choice tasks for evaluating lesion-level image–text alignment, localization, size estimation, attribute recognition, and multi-slice CT reasoning.

The objective of CT-Bench is to provide an AI-ready resource for developing and evaluating models that connect CT images with clinically meaningful lesion-level text. The dataset contains 20,335 lesion instances from 7,795 CT studies and 3,793 patients, split into training, validation, and test subsets. The QA benchmark contains 2,850 question–answer pairs across seven lesion-analysis tasks and includes hard negative answer choices to make evaluation sensitive to subtle lesion-level differences.

CT-Bench is designed for AI-readiness and reproducibility. Each lesion instance is associated with standardized identifiers, structured metadata, lesion-level text, lesion size information, and bounding-box coordinates. Lesion descriptions are mapped to structured attributes, including body part, lesion type, and lesion properties, using a RadLex-based lesion ontology. These structured fields support filtering, stratified analysis, ontology-based QA construction, and reproducible benchmarking.

The intended audience includes researchers in medical image computing, radiology AI, multimodal representation learning, medical visual question answering, and clinical natural language processing. CT-Bench is intended for research and benchmarking only and is not approved for standalone clinical decision-making.

3. Discussion

Several limitations should be considered. CT-Bench is derived from retrospective clinical CT data and may reflect the patient population, scanner distribution, imaging protocols, lesion selection process, and reporting practices of the source data. Bounding boxes provide lesion localization but do not represent full volumetric segmentation. In addition, the QA benchmark uses controlled multiple-choice tasks and does not evaluate full diagnostic workflows, longitudinal comparison, treatment response assessment, or report generation in real clinical settings.

Potential biases and representation concerns should be explicitly considered. Users should evaluate dataset balance across sex, age, anatomical region, lesion type, lesion size, and imaging protocol where metadata are available. Models trained or evaluated on CT-Bench may inherit biases from the source cohort, lesion selection process, annotation workflow, and imaging protocols; therefore, stratified reporting is recommended whenever possible.

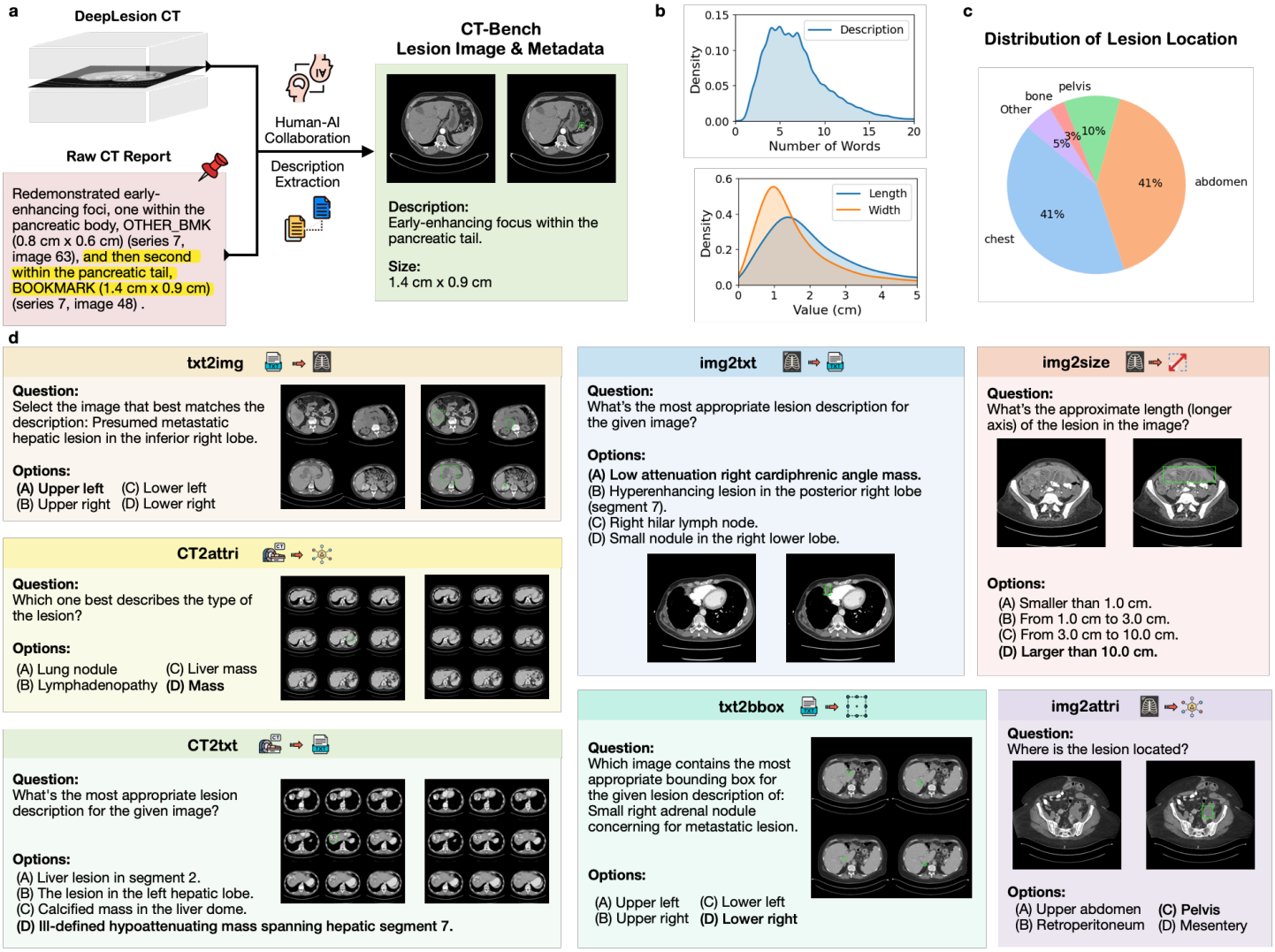

Figure 1: Overview of CT-Bench. (a) Example of data annotation, showing the transformation of original reports from PACS into detailed descriptions enriched with size information for CT-Bench: Lesion Image & Metadata Set. (b) Analysis of description: word count and value distribution. (c) Distribution of lesion locations. (d) CT-Bench: components of the QA benchmark cases.

CT-Bench is intended for research, benchmarking, and method development only. It is not approved for standalone clinical diagnosis, triage, treatment planning, or patient management. Users must not attempt re-identification, linkage with external identifiable sources, or redistribution that violates the dataset license or upstream data-use conditions.

4. Resource Availability

Repository and DOI. CT-Bench is publicly available at https://kaggle.com/datasets/cd1661d6d6aeab08b8eb99b58885b4489d76fc5ac07d5aac76cae577e6426e2f with DOI https://doi.org/10.34740/kaggle/ds/7422536. The repository contains de-identified CT lesion images, lesion-level annotations, metadata files, ontology mappings, train/validation/test split files, QA benchmark files, benchmark construction scripts, and baseline evaluation code.

Released files. The released resource includes CT lesion images, bounding-box annotations, PACS-derived lesion descriptions, lesion size measurements, metadata, structured ontology-derived lesion attributes, QA benchmark files, and reproducibility code. The metadata file `CT-Bench-Final-Merged-with-DeepLesion-Metadata.csv` integrates CT-Bench lesion annotations with available metadata from the original DeepLesion release. The file `CT-Bench_Lesion_Ontology_Mapping.csv` provides structured lesion attributes, including body part, lesion type, and lesion properties mapped to a RadLex-based lesion ontology.

Potential use cases and target tasks. CT-Bench is intended for research on multimodal AI for CT lesion analysis. Potential use cases include lesion-level CT image–

text retrieval, lesion description generation, visual question answering, lesion localization, description-to-bounding-box matching, lesion size estimation, lesion attribute classification, ontology-based lesion analysis, and evaluation of spatial grounding in medical vision-language models. CT-Bench is not intended for standalone clinical diagnosis, triage, treatment planning, patient management, or regulatory claims of clinical safety.

Licensing. The CT-Bench data and annotations are released under **CC BY-NC-SA 4.0**.

Ethical approval and consent. DeepLesion was created from retrospective NIH Clinical Center PACS data under institutional review board approval; its original publication describes anonymization by replacing patient identifiers, accession numbers, and series numbers with research indices . CT-Bench is a secondary derived resource based on a subset of DeepLesion images and associated PACS bookmark text. CT-Bench curation, linkage, de-identification, and release were conducted under NIH IRB protocol, with a waiver of informed consent for retrospective secondary use. No prospective recruitment or new imaging was performed, and all released files were de-identified in accordance with this approval and applicable upstream data-use conditions.

Data sensitivity and privacy protection. CT-Bench consists of de-identified clinical imaging data and derived lesion-level annotations. Direct identifiers were removed from image files, metadata files, filenames, annotation files, and QA files before release. Removed or replaced fields include names, medical record numbers, accession numbers, direct date fields, and other direct protected health information. Patient, study, image, and lesion identifiers were replaced with research identifiers.

Use of external AI services during data creation. GPT-4 and GPT-4V were used during annotation support and evaluation through Microsoft Azure services to support secure and privacy-compliant handling. The released dataset contains de-identified images and annotations rather than raw protected health information. Details of the annotation workflow, human review, and model-assisted processing are provided in Section 5 and in the repository documentation.

Contact. Questions about data access, licensing, ethics, privacy, responsible use, or suspected data-quality issues should be directed to Zhiyong Lu at `zhiyong.lu@nih.gov`.

5. Methods

5.1 Data source, cohort, and study design

CT-Bench was constructed as a retrospective lesion-level computed tomography (CT) image–text resource derived from the DeepLesion dataset and associated Picture Archiving and Communication System (PACS)-derived lesion bookmarks (Yan et al., 2017). The DeepLesion source version used in this work contains 33,688 bookmarked radiology images from 10,825 CT studies of 4,477 unique patients. These source data provide CT key-slice images, lesion bookmarks, bounding-box annotations, lesion measurements, and image-level metadata, but they do not include curated lesion-specific textual descriptions or standardized lesion size descriptions aligned with each bookmarked finding.

CT-Bench extends this source resource by adding curated lesion-level textual descriptions, standardized textual lesion-size representations, structured ontology-derived attributes, merged metadata files, and multiple-choice visual question answering benchmark cases. After applying the inclusion, exclusion, de-identification, and annotation-quality criteria described below, the final CT-Bench Lesion Image and Metadata Set contains 20,335 lesion instances from 7,795 CT studies and 3,793 patients. Each retained lesion instance is linked to a CT key-slice image, bounding-box coordinates, a PACS-derived lesion description, a standardized size measurement, and associated patient-, study-, image-, and lesion-level metadata where available.

No prospective recruitment or new imaging was performed. Ethical provenance, approval, and de-identification are described in Section 4.

5.2 Inclusion and exclusion criteria

Lesion instances were included if they were associated with a CT image from the DeepLesion source dataset, had a valid lesion bookmark or bounding-box annotation, had recoverable PACS-derived text referring to the bookmarked finding, and passed de-identification and quality-control checks. Lesions were excluded if the image or metadata were unavailable, unreadable, corrupted, or inconsistent; if the bounding box was missing or invalid; if the PACS-derived text could not be mapped unambiguously to the bookmarked lesion; if the lesion size could not be extracted or standardized when required for a downstream task; or if the record contained protected health information that could not be safely de-identified. These criteria were used to create a lesion-level resource suitable for multimodal AI tasks involving CT images, bounding boxes, lesion descriptions, size measurements, and structured metadata.

5.3 Dataset composition and demographics

The final CT-Bench Lesion Image and Metadata Set contains 20,335 lesion instances from 7,795 CT studies and 3,793 patients. The dataset is divided into training, validation, and test subsets at the patient level to avoid patient overlap between splits. The training set contains 15,380

Table 1: Composition of the CT-Bench Lesion Image and Metadata Set. Splits are disjoint at the patient level.

Subset	Patients	Male	Female	Studies	Lesions
Training	2,835	1,516	1,319	5,875	15,380
Validation	342	181	161	667	1,758
Test	616	367	249	1,253	3,197
Total	3,793	2,064	1,729	7,795	20,335

lesions from 5,875 studies and 2,835 patients; the validation set contains 1,758 lesions from 667 studies and 342 patients; and the test set contains 3,197 lesions from 1,253 studies and 616 patients. The cohort includes 2,064 male and 1,729 female patients. At the patient level, the median age was 55 years . The cohort included 174 patients younger than 18 years, 754 patients aged 18–39 years, 1,399 patients aged 40–59 years, 1,401 patients aged 60–79 years, and 65 patients aged 80 years or older.

5.4 Available imaging and protocol metadata

CT-Bench integrates lesion annotations with metadata inherited from the original DeepLesion release. The merged metadata file includes patient, study, image, and lesion identifiers; patient age and sex; key-slice index; bounding-box coordinates; measurement coordinates; pixel-level lesion diameters; normalized lesion location; slice range; pixel spacing; slice spacing; image size; DICOM window values; split assignment; and structured lesion attributes. These fields support reproducible preprocessing, lesion localization, physical size estimation, and stratified analysis.

The released metadata include image size, pixel spacing, slice spacing, and DICOM window values for all 20,335 lesion records. Most images are 512×512 pixels. The median in-plane pixel spacing is 0.816 mm/pixel [0.732–0.877], and the median slice spacing is 2.0 mm [1.0–5.0]. Scanner manufacturer, scanner model, tube voltage, tube current, reconstruction algorithm, and contrast phase are not available in the merged release metadata and are therefore documented as unavailable rather than imputed.

5.5 Released data format and preprocessing

The released CT-Bench files include de-identified CT lesion images, lesion-level annotation files, metadata files, ontology mapping files, train/validation/test split files, and QA benchmark files. CT lesion images are released as PNG files. Lesion annotations, metadata, ontology mappings, split files, and QA benchmark files are released as CSV or Json files.

For image preprocessing, we followed the DeepLesion PNG conversion and windowing convention used in the public DeepLesion release https://nihcc.app.box.com/v/DeepLesion/ and https://huggingface.co/datasets/farrell236/DeepLesion. The source CT images are DICOM-derived CT slices represented as PNG files. In the 16-bit PNG representation, pixel values are stored as unsigned integers with an offset; Hounsfield Unit (HU) values can be recovered by subtracting 32768 from the stored pixel intensity. For the model-ready images released with CT-Bench, we applied the corresponding DICOM window values provided in the metadata, clipped the HU values to the window range, and rescaled the result to 8-bit PNG intensity values in the range $[0, 255]$.

The preprocessing pipeline linked each lesion image to patient, study, image, and lesion identifiers; inherited bounding-box coordinates from the DeepLesion metadata; standardized lesion descriptions and lesion size strings; converted DICOM-derived CT intensities to released PNG images using the windowing procedure above; generated image variants with and without bounding-box overlays where required for QA tasks; constructed patient-level train/validation/t split files; merged CT-Bench annotations with available DeepLesion metadata; generated QA benchmark files and answer-choice metadata; and ran quality-control checks for image readability, metadata consistency, split integrity, and annotation validity. The repository includes preprocessing code and a data dictionary describing each released field, its type, allowed values, missingness, and source.

5.6 Anonymization and privacy protection

All released data were de-identified before distribution. Direct identifiers were removed from image files, metadata files, filenames, annotation files, and QA files. Removed or replaced fields include names, medical record numbers, accession numbers, direct date fields, and other direct protected health information. Patient, study, image, and lesion identifiers were replaced with research identifiers. Metadata fields were reviewed before release to reduce the risk of protected health information leakage. Records that failed de-identification checks were excluded from release or corrected before inclusion. Additional ethical and data-governance details are provided in Section 4.

5.7 Ground-truth definition

The ground-truth lesion description is defined as the final medically reviewed structured description linked to the bookmarked lesion and corresponding bounding box. The ground-truth size is defined as the final standardized lesion size measurement extracted from the PACS-derived bookmarked sentence. The ground-truth bounding box is inherited from the source lesion bookmark annotation and verified during curation and quality control. For ontology-derived tasks, the ground-truth attribute is defined as the final structured body-part, lesion-type, or lesion-property

label mapped from the curated lesion description to the RadLex-based lesion ontology.

5.8 Lesion description annotation procedure

Lesion description and size annotations were created using a three-stage human-in-the-loop workflow that combined GPT-4-assisted pre-annotation, annotator review, iterative refinement, and medical-expert verification (Figure 2). In Stage 1, GPT-4 generated preliminary structured descriptions and size fields for 200 CT cases based on radiologist-authored PACS annotations. Annotators reviewed and refined these preliminary annotations through collaborative discussion, and the corrected cases were used to fine-tune GPT-4 for annotation support. In Stage 2, the fine-tuned model generated pre-annotations for additional batches of cases. Annotators reviewed and corrected these outputs through iterative feedback rounds, and the corrected annotations were used for further refinement. In Stage 3, the final fine-tuned model was applied to the remaining training, validation, and test data to generate preliminary lesion descriptions and size measurements.

Final labels were not accepted automatically. Validation and test cases were reviewed by two annotators and verified by a medically trained expert. Training cases were reviewed by one annotator and then verified by a medical professional. The annotation team included contributors with backgrounds in medical image processing, computer science, and medicine. Medically trained personnel contributed to annotation guideline development, review, adjudication, and final verification, especially for ambiguous cases involving multiple lesions, uncertain references, or difficult terminology. Annotation quality statistics are reported in Section 6.

5.9 Semantic attribute extraction and ontology mapping

To support structured search, stratified analysis, and attribute-based QA construction, CT-Bench includes ontology-derived lesion attributes. A fine-tuned GPT-4 model was used to map curated lesion descriptions to fine-grained attributes covering body part, lesion type, and lesion property. These attributes were mapped to a RadLex-based lesion ontology (Langlotz, 2006). The released file `CT-Bench_Lesion_Ontology_Mapping.csv` contains the lesion description, lesion size, mapped lesion type, body-part terms, and lesion-property terms. Structured lesion type labels are available for 20,103 lesions (98.9%), body-part labels for 19,960 lesions (98.2%), and imaging property labels for 4,841 lesions (23.8%). These fields support ontology-based filtering, reproducible benchmark generation, and evaluation of lesion attribute recognition.

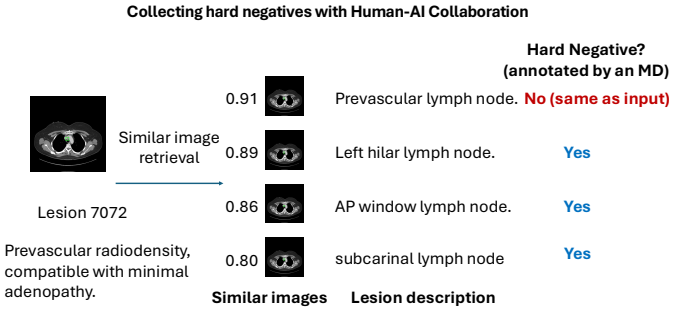

Table 2: Illustration of the selection process for "hard negative" cases using the BiomedCLIP model and MD validation.

Table 3: Distribution of CT-Bench QA Benchmark cases across task types. "w/o" denotes without bounding-box context and "w/" denotes with bounding-box context.

	Img2txt		CT2txt		Txt2img		Txt2bbox		Img2attrib		CT2attrib		Img2size		Total
Subset	w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/o	w/	
Validation	100	100	100	100	100	100	–	50	50	50	50	50	50	50	950
Test	200	200	200	200	200	200	–	100	100	100	100	100	100	100	1,900
Total	300	300	300	300	300	300	–	150	150	150	150	150	150	150	2,850

5.10 QA benchmark construction

The CT-Bench QA Benchmark was constructed from the CT-Bench Lesion Image and Metadata Set to evaluate lesion-level multimodal CT understanding. It contains 2,850 multiple-choice QA pairs across seven tasks: image-to-description matching (`Img2txt`), multi-slice CT to description matching (`CT2txt`), description-to-image retrieval (`Txt2img`), description-to-bounding-box localization (`Txt2bbox`), lesion size estimation (`Img2size`), image-based attribute recognition (`Img2attrib`), and multi-slice CT attribute recognition (`CT2attrib`).

Each QA item contains one correct answer and three distractors. Randomly sampled distractors were often too easy because they could differ substantially in anatomy, appearance, or lesion type. Therefore, hard negatives were generated to create more challenging answer choices. For image-to-description and description-to-image tasks, BiomedCLIP (Zhang et al., 2023a) was used to retrieve visually similar lesion cases, and medically trained reviewers selected clinically plausible hard negatives. This process is visualized in Figure 2. For `Txt2bbox`, candidate bounding boxes from visually similar retrieved images were inserted as alternative choices when they did not overlap the target lesion. For attribute tasks, structured ontology-derived labels were used to generate candidate body-part, lesion-type, and lesion-property choices, followed by medical review. For multi-slice tasks, nine consecutive slices were used, consisting of the key slice, four preceding slices, and four following slices.

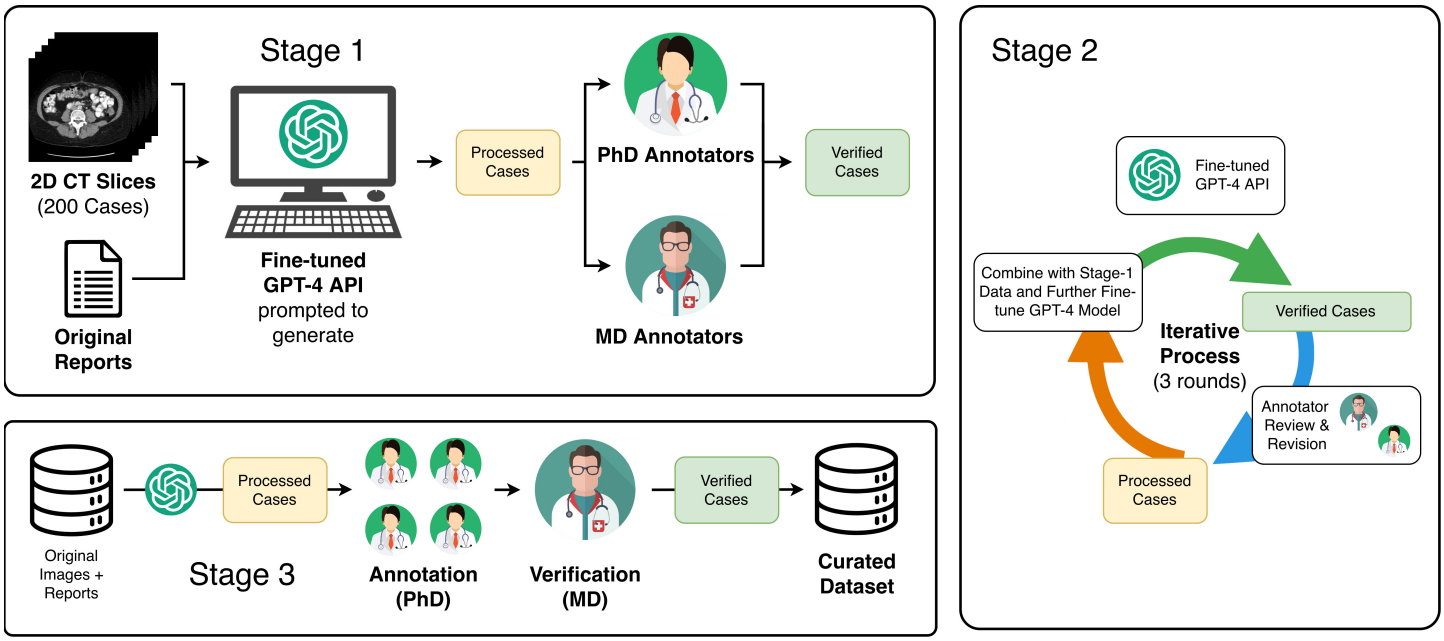

Figure 2: Three-stage annotation and verification workflow used to construct CT-Bench lesion descriptions and size measurements. GPT-4-assisted pre-annotations were reviewed by human annotators and verified by medically trained experts before release.

5.11 Known data artifacts and limitations

Several data artifacts should be considered when using CT-Bench. First, the dataset is derived from retrospective clinical CT data and may reflect the scanner distribution, imaging protocols, reporting style, and lesion selection process of the source data. Second, some CT acquisition fields are unavailable because CT-Bench inherits metadata from the upstream DeepLesion release; missing fields are documented in the data dictionary rather than imputed. Third, lesion descriptions are derived from PACS bookmarks and report text, which may contain ambiguous references, multiple lesions in the same sentence, adjacent non-lesion findings, or comparative language that cannot always be verified from a single image. Fourth, bounding boxes provide lesion localization but do not represent full volumetric segmentation. Fifth, the QA benchmark uses multiple-choice answer formats and should be interpreted as a controlled benchmark for lesion-level multimodal reasoning, not as a complete simulation of clinical diagnosis.

6. Validation

6.1 Annotation quality control

The lesion descriptions and size annotations were validated during the three-stage annotation process described in Section 5. In the initial annotation stage, annotator performance varied substantially by background. Annotators with PhD-level training in medical image analysis or computer science but without formal medical training achieved accuracies between 56.0% and 73.0%, whereas the medically trained annotator achieved 93.0%. This difference motivated the use of explicit annotation guidelines, iterative feedback, and medical-expert verification.

During the second stage, GPT-assisted pre-annotations were reviewed and corrected by human annotators over three iterative rounds. Annotator accuracy improved to approximately 90% across rounds, and the fine-tuned GPT-4 pre-annotation model achieved stable accuracy of 89.0%, 88.0%, and 88.0% across the three rounds. Final labels were not accepted automatically. In the final large-scale annotation stage, GPT-assisted outputs were reviewed by human annotators and verified by medically trained experts. Validation and test annotations were reviewed by two annotators and then verified by a medical expert; training annotations were reviewed by one annotator and then verified by a medical professional. Detailed annotator-level quality-control results are provided in the supplementary material.

We assessed the completeness and internal consistency of the merged CT-Bench metadata file. Bounding-box coordinates, measurement coordinates, pixel-level lesion diameters, normalized lesion locations, image size, DICOM window values, pixel spacing, patient sex, and patient age were available for all 20,335 lesion records. The training, validation, and test splits were disjoint at the patient level, with no patient appearing in more than one split. Normalized lesion locations were within the expected range for all lesions. Coordinate checks identified no inverted bounding boxes; seven bounding boxes extended marginally outside

Table 4: Validation and example AI-use-case summary for CT-Bench. Accuracy is averaged across evaluated QA tasks. Human-reader results were computed on sampled cases, whereas model results were computed on the QA benchmark test set.

Setting	w/o BBox	w/ BBox	Total average
Random baseline	25.0	25.0	25.0
BiomedCLIP, untuned	34.9	41.0	38.1
BiomedCLIP, fine-tuned with CT-Bench	37.1	62.0	50.3
Senior radiologist average	68.3	90.3	80.4
Junior doctor	45.5	68.5	58.2

the image boundary and are documented for transparent reuse.

Structured lesion type labels were available for 20,103 lesions (98.9%), body-part labels for 19,960 lesions (98.2%), and imaging-property labels for 4,841 lesions (23.8%). The textual lesion size field was available for all lesions. These checks identify which metadata fields are complete, which fields are optional, and which records may require additional caution during downstream analysis. Full metadata completeness statistics are provided in the supplementary material.

6.2 Clinical validation by human readers

To assess whether the CT-Bench QA Benchmark Component reflects clinically meaningful lesion-analysis tasks, we conducted a human-reader evaluation with two senior radiologists and one junior doctor with radiology experience. Each reader reviewed randomly sampled cases across the benchmark tasks under conditions with and without bounding-box information. We selected 100 cases in total.

The two senior radiologists achieved total average accuracies of 79.2% and 81.5%, respectively. Their performance was higher when bounding boxes were provided, reaching 90.0% and 90.5% average accuracy in the bounding-box setting. The junior doctor achieved a lower total average accuracy of 58.2%, suggesting that the benchmark is sensitive to clinical expertise. These results indicate that CT-Bench evaluates clinically relevant visual and textual reasoning rather than trivial image–text matching.

6.3 Example AI use case

To demonstrate the usability of CT-Bench for multimodal AI development and evaluation, we performed example experiments using the two released components of the resource. First, we evaluated lesion-level captioning on the CT-Bench Lesion Image and Metadata Set, where the goal is to generate a clinically meaningful lesion description from a CT lesion image. Model outputs were compared with the curated radiologist-derived lesion descriptions using (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE(Lin, 2004), BERTScore (Zhang et al., 2019) and Cosine Similarity (Zhang et al., 2019) . Fine-tuning RadFM on the CT-Bench Lesion Image and Metadata Set improved performance over untuned models across the reported captioning metrics, demonstrating that the released lesion image–text pairs can support model adaptation for structured lesion description generation. Full captioning results are provided in the supplementary material.

Second, we evaluated multiple-choice visual question answering on the CT-Bench QA Benchmark Component. The benchmark uses a standardized four-choice format across lesion description, image–text retrieval, bounding-box localization, lesion size estimation, and attribute-recognition tasks. A random baseline achieved 25.0% accuracy. Among untuned models, BiomedCLIP achieved the strongest average performance, reaching 34.9% accuracy without bounding boxes and 41.0% accuracy with bounding-box inputs. After fine-tuning on CT-Bench, BiomedCLIP achieved 62.0% average accuracy in the bounding-box setting, demonstrating that the resource can support model adaptation for lesion-level CT reasoning.

Model performance remained below the senior-radiologist average, indicating that CT-Bench remains a challenging benchmark. These results support the use of CT-Bench for research, model development, and standardized evaluation, but they should not be interpreted as evidence that models trained or evaluated with CT-Bench are ready for standalone clinical decision making. Full task-level model results, captioning results, and model-configuration details are provided in the supplementary material.

Conflicts of Interest

Author RMS receives royalties from iCAD, Philips, and ScanMed.

Acknowledgements

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH). The contributions of the NIH authors are considered Works of the United States Government. The findings and conclusions presented in this paper are those of the authors

and do not necessarily reflect the views of the NIH or the U.S. Department of Health and Human Services.

References

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

Asma Ben Abacha, Mourad Sarrouti, 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.

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

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

Qiao Jin, Fangyuan Chen, Yiliang Zhou, Ziyang Xu, Justin M Cheung, Robert Chen, Ronald M Summers, Justin F Rousseau, Peiyun Ni, Marc J Landsman, et al. Hidden flaws behind expert-level accuracy of multimodal gpt-4 vision in medicine. *ArXiv*, pages arXiv–2401, 2024.

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

Curtis P Langlotz. Radlex: a new method for indexing online educational materials, 2006.

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.

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

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. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 525–536. Springer, 2023.

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

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

Johannes Rückert, Louise Bloch, Raphael Brüngel, Ahmad Idrissi-Yaghir, Henning Schäfer, Cynthia S Schmidt, Sven Koitka, Obioma Pelka, Asma Ben Abacha, Alba G. Seco de Herrera, et al. Rocov2: Radiology objects in context version 2, an updated multimodal image dataset. *Scientific Data*, 11(1):688, 2024.

Sanjay Subramanian, Lucy Lu Wang, Sachin Mehta, Ben Bogin, Madeleine van Zuylen, Sravanthi Parasa, Sameer Singh, Matt Gardner, and Hannaneh Hajishirzi. Medicat: A dataset of medical images, captions, and textual references. *arXiv preprint arXiv:2010.06000*, 2020.

Zhaoyi Sun, Mingquan Lin, Qingqing Zhu, Qianqian Xie, Fei Wang, Zhiyong Lu, and Yifan Peng. A scoping review on multimodal deep learning in biomedical images and texts. *Journal of Biomedical Informatics*, page 104482, 2023.

Shubo Tian, Qiao Jin, Lana Yeganova, Po-Ting Lai, Qingqing Zhu, Xiuying Chen, Yifan Yang, Qingyu Chen, Won Kim, Donald C Comeau, et al. Opportunities and challenges for chatgpt and large language models in biomedicine and health. *Briefings in Bioinformatics*, 25(1):bbad493, 2024.

Ke Yan, Xiaosong Wang, Le Lu, and Ronald M Summers. Deeplesion: Automated deep mining, categorization and detection of significant radiology image findings using large-scale clinical lesion annotations. *arXiv preprint arXiv:1710.01766*, 2017.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023a.

Tianyi Zhang, Varsha Kishore, Felix Wu, et al. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

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

Qingqing Zhu, Tejas Sudharshan Mathai, Pritam Mukherjee, Yifan Peng, Ronald M Summers, and Zhiyong Lu. Utilizing longitudinal chest x-rays and reports to pre-fill radiology reports. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 189–198. Springer, 2023.

Qingqing Zhu, Benjamin Hou, Tejas Sudarshan Mathai, Pritam Mukherjee, Qiao Jin, Xiuying Chen, Zhizheng Wang, Ruida Cheng, Ronald M Summers, and Zhiyong Lu. How well do multimodal llms interpret ct scans? an auto-evaluation framework for analyses. *Journal of Biomedical Informatics*, 168:104864, 2025.

Appendix A. Supplementary Dataset Construction Details

A.1 Annotation challenge examples

Table S2 provides examples of report-derived lesion descriptions and the corresponding standardized CT-Bench annotations. These examples illustrate common challenges encountered during curation, including multiple lesions in the same sentence, comparative phrasing, adjacent findings, and ambiguous anatomical references.

A.2 Annotation prompt and standardization rules

The following instruction was used to guide GPT-assisted pre-annotation and annotator review:

> You are a helpful assistant for extracting relevant descriptions within a bounding box in CT image reports. The descriptions of the bounding box are labeled with `BOOKMARK`, `SIZE_BMK`, or `INDEX_BMK`, while `OTHER_BMK` describes other bounding boxes and should not be included. The descriptions should contain clear location, property, and, if applicable, indication. Use `Unclear Description` if the location or property information is missing. Please also extract the size of the bounding box. You should output a JSON dictionary formatted as `description: str, size: str`.

The following annotation rules were used to improve consistency:

- **Use of "enlarged".** The term was retained for lymph nodes when enlargement was clinically meaningful, but avoided for nodules when the term did not add valid information.
- **Use of comparative terms.** Terms such as "largest" were avoided when the comparison could not be verified from a single image.
- **Singular and plural forms.** Ambiguous plural expressions such as "foci" were standardized to clearer lesion terms such as "focus" or "focal lesion" when appropriate.
- **Description completeness.** Descriptions were required to include clear anatomical location and lesion property information when available. Cases with insufficient location or property information were marked as unclear or excluded when reliable annotation was not possible.

A.3 Three-stage annotation workflow

The CT-Bench lesion descriptions and size measurements were created with a three-stage human-in-the-loop workflow. In Stage 1, GPT-4 generated preliminary descriptions for 200 CT cases, which were then reviewed by annotators. These corrected cases were used to fine-tune GPT-4 for annotation support. In Stage 2, the fine-tuned model generated pre-annotations for additional batches of cases, followed by iterative human correction and feedback. In Stage 3, the final fine-tuned model was applied to the remaining training, validation, and test data; all preliminary outputs were reviewed by human annotators and verified by medically trained experts before release.

A.4 GPT-4 fine-tuning details for annotation support

To support efficient annotation, GPT-4 was fine-tuned using supervised learning on the curated lesion annotation dataset. The final training file contained 500 curated cases and approximately 329,000 tokens. Training was conducted for three epochs with batch size 1 and the default learning rate on the Azure OpenAI platform. The fine-tuned model was used to generate preliminary annotations in Stages 2 and 3, but final labels were accepted only after human review and medical-expert verification.

A.5 DeepLesion metadata integration and PNG preprocessing

CT-Bench integrates curated lesion annotations with metadata inherited from DeepLesion. The merged file `CT-Bench-Final-Merged-with-DeepLesion-Metadata.csv` includes patient, study, image, and lesion identifiers; patient age and sex; key-slice index; bounding-box coordinates; measurement coordinates; pixel-level lesion diameters; normalized lesion location; slice range; pixel spacing; slice spacing; image size; DICOM window values; split assignment; and structured lesion attributes.

For image preprocessing, CT-Bench follows the DeepLesion PNG conversion and windowing convention used in the public DeepLesion release . In the 16-bit PNG representation, pixel values are stored as unsigned integers with an offset; Hounsfield Unit (HU) values can be recovered by subtracting 32768 from the stored intensity:

$$I_{\mathrm{HU}} = I_{16} - 32768. \tag{1}$$

Given a window lower bound $W_{\min}$ and upper bound $W_{\max}$ from the `DICOM_windows` metadata field, the released 8-bit PNG intensity was computed as

$$I_8 = 255 \times \frac{\mathrm{clip}(I_{\mathrm{HU}}, W_{\min}, W_{\max}) - W_{\min}}{W_{\max} - W_{\min}}. \tag{2}$$

The resulting values were rounded and stored as unsigned 8-bit PNG images. Images with bounding-box overlays were generated only for benchmark variants that required visible localization cues; the underlying lesion image and metadata were preserved.

A.6 Metadata completeness and consistency checks

Table S1: Comparison of the components inherited from DeepLesion and those added in CT-Bench. CT-Bench is a derived extension of a quality-controlled subset of DeepLesion rather than a newly collected CT imaging cohort.

Aspect	Inherited from DeepLesion	Added in CT-Bench
Resource role	Source CT images and lesion annotations.	A derived multimodal lesion-level resource and benchmark; no new imaging cohort was collected.
Scale	33,688 bookmarked images from 10,825 CT studies and 4,477 patients.	A quality-controlled subset of 20,335 lesion instances from 7,795 CT studies and 3,793 patients.
Images and localization	CT key-slice images, lesion bookmarks, bounding boxes, measurements, and image-level metadata.	Curated linkage of the retained images and annotations with lesion-level text and benchmark metadata.
Lesion text	No curated lesion-specific descriptions or standardized textual size descriptions aligned with each lesion.	Curated PACS-derived lesion descriptions and standardized textual lesion-size representations.
Structured attributes	Not part of the source components used for CT-Bench.	RadLex-based body-part, lesion-type, and lesion-property labels.
QA benchmark	No unified lesion-level multimodal QA benchmark.	2,850 four-choice questions across seven lesion-analysis tasks, including clinically plausible hard negatives.
Organization and release	Original DeepLesion images and metadata.	Patient-disjoint splits, merged metadata, ontology mappings, QA files, preprocessing scripts, and baseline evaluation code.

Appendix B. Supplementary Benchmark Construction Details

B.1 Hard-negative construction

Randomly sampled distractors were often visually or semantically dissimilar from the target case, making the multiple-choice tasks too easy. To make the benchmark clinically meaningful, CT-Bench uses hard negatives selected from visually or semantically similar lesions. For image-to-description and description-to-image tasks, BiomedCLIP was used to retrieve visually similar lesion cases, and medically trained reviewers selected clinically plausible incorrect choices. For `Txt2bbox`, bounding boxes from visually similar retrieved images were used as alternative choices only when they did not overlap the target lesion. For attribute tasks, distractors were generated from ontology-derived body-part, lesion-type, and lesion-property labels and then reviewed by medically trained reviewers.

B.2 Task input formats

B.3 Task definitions

`Img2txt`. The model is given a single CT lesion image and four candidate lesion descriptions. The correct answer is the curated CT-Bench lesion description for the target lesion. The model selects the most appropriate description.

`CT2txt`. The model is given nine consecutive CT slices centered on the lesion key slice and four candidate descriptions. The task tests whether neighboring slices improve lesion interpretation.

`Txt2img`. The model is given a lesion description and four candidate lesion images arranged as choices. The task tests image retrieval conditioned on lesion-level clinical text.

`Txt2bbox`. The model is given a lesion description and four candidate images with bounding boxes. The task tests whether the model can identify the image containing the bounding box that best matches the description.

`Img2size`. The model is given a lesion image and four size categories: smaller than 1.0 cm, 1.0–3.0 cm, 3.0–10.0 cm, and larger than 10.0 cm. The target is the lesion long-axis size category.

`Img2attrib`. The model is given a lesion image and four candidate ontology-derived attributes. The attribute may refer to lesion location, lesion type, or lesion property.

`CT2attrib`. The model is given nine consecutive CT slices and four candidate ontology-derived attributes. This task extends `Img2attrib` to multi-slice input.

Appendix C. Supplementary Experimental Details and Results

C.1 Model configurations

C.2 Fine-tuning settings

RadFM and BiomedCLIP were fine-tuned as representative models from the medical vision-language and CLIP-style model families. RadFM was trained with batch size 4, four-step gradient accumulation, maximum token length 2048, and four epochs. BiomedCLIP was trained with the Adam optimizer, learning rate 1×10^{-5}, batch size 64, and 200 epochs. Cross-entropy loss was used for multiple-choice QA tasks. Fine-tuning was performed using PyTorch on

Table S2: Examples of lesion descriptions from PACS-derived text with corresponding standardized annotations. Each case includes the original sentence, the final lesion description, the extracted size, and the main annotation challenge.

Original sentence	Ground-truth description	Ground-truth size	Annotation challenge
Unchanged right lower lobe lung nodule OTHER_BMK (0.6 cm x 0.4 cm) and right lower lobe fissural likely lymph node BOOKMARK (0.6 cm x 0.3 cm)	Right lower lobe fissural likely lymph node.	0.6 cm x 0.3 cm	Multiple lesions in the same region; requires disambiguation.
Bilateral small renal hypodensities, for example on the right measuring OTHER_BMK (0.5 cm x 0.4 cm), another smaller more inferior low-density area measuring BOOKMARK (0.6 cm x 0.4 cm)	Small low-density area in the right kidney.	0.6 cm x 0.4 cm	Vague comparative language.
Representative large left axillary node BOOKMARK (5.3 cm x 3.2 cm). At the site of preexisting atherosclerotic disease in the infrarenal abdominal aorta, new focal saccular aneurysm measuring OTHER_BMK (4.3 cm x 2.9 cm)	Large left axillary node.	5.3 cm x 3.2 cm	Multiple findings; need to isolate lymph node from vascular finding.
In the pelvis, increase in the left upper pelvic node at the level of the iliac bifurcation BOOKMARK (3.5 cm x 1.4 cm), with adjacent fluid collection OTHER_BMK (4.5 cm x 3.8 cm)	Left upper pelvic node at the level of the iliac bifurcation.	3.5 cm x 1.4 cm	Confounding adjacent findings.
Unchanged low-density lesions, for example gastrohepatic BOOKMARK (2.8 cm x 1.8 cm), inferiorly OTHER_BMK (2.5 cm x 2.1 cm), and pelvis OTHER_BMK (2.3 cm x 1.9 cm)	Low-density lesion in the gastrohepatic region.	2.8 cm x 1.8 cm	Multiple similar lesions requiring regional disambiguation.

Table S3: Annotation quality during the development of CT-Bench. Stage 1 reports initial annotation accuracy. Stage 2 reports accuracy across three iterative GPT-assisted annotation and review rounds. "–" indicates that the annotator or model did not participate in that round.

Annotator / model	Stage 1	Stage 2 R1	Stage 2 R2	Stage 2 R3
Annotator 1 (PhD)	56.0	90.0	95.0	86.0
Annotator 2 (PhD)	73.0	93.0	89.0	89.0
Annotator 3 (PhD)	61.5	85.0	91.0	86.0
Annotator 4 (PhD)	60.0	–	88.0	85.0
Annotator 5 (MD)	93.0	96.0	94.0	96.0
Fine-tuned GPT-4	–	89.0	88.0	88.0

NVIDIA A100 GPUs.

C.3 Evaluation metrics

For lesion captioning, generated descriptions were compared with curated lesion descriptions using BLEU-1, METEOR, ROUGE-1, BERTScore F1, and Sentence-BERT cosine similarity. For QA benchmarking, accuracy was used as the primary metric because each item contains one correct answer among four choices. Human-reader results were computed on sampled cases, whereas model QA results were computed on the QA benchmark test set unless otherwise noted.

C.4 Full lesion-captioning results

C.5 Full QA benchmark test results

C.6 Validation-set QA results

C.7 Average inference time

C.8 Qualitative captioning case study

Table S4: Metadata completeness and consistency checks for the CT-Bench merged metadata file. Percentages are computed over all 20,335 lesion records unless otherwise noted.

Field or quality-control check	Result
Bounding-box coordinates	20,335 / 20,335 available (100.0%)
Measurement coordinates	20,335 / 20,335 available (100.0%)
Pixel-level lesion diameters	20,335 / 20,335 available (100.0%)
Normalized lesion location	20,335 / 20,335 available and within range (100.0%)
Image size, pixel spacing, and DICOM windows	20,335 / 20,335 available (100.0%)
Patient age and sex	20,335 / 20,335 lesion records available (100.0%)
Patient-level split overlap	No patient overlap across training, validation, and test splits
Bounding-box sanity check	20,328 / 20,335 within image bounds; 7 marginal boundary cases; 0 inverted boxes
Textual lesion size field	20,335 / 20,335 available (100%)
Structured lesion type labels	20,103 / 20,335 available (98.9%)
Structured body-part labels	19,960 / 20,335 available (98.2%)
Structured imaging-property labels	4,841 / 20,335 available (23.8%)
Possibly noisy flag	10 / 20,335 lesions marked as possibly noisy (0.05%)

Table S5: Distribution of lesion long-axis size estimated from pixel-level lesion diameters and in-plane pixel spacing. Percentages are computed over all 20,335 lesions.

Long-axis size category	Lesions	Percentage
< 1 cm	3,632	17.9%
1–< 3 cm	11,809	58.1%
3–< 10 cm	4,681	23.0%
≥ 10 cm	213	1.0%
Total	20,335	100.0%

Table S6: Examples from CT-Bench_Lesion_Ontology_Mapping.csv. Lesion descriptions are mapped to structured lesion type, body-part, and property attributes using a RadLex-based lesion ontology.

Description	Size	Type	Body part	Property
Low-density renal lesion in the left kidney.	0.9 cm x 0.8 cm	kidney cyst; cyst	upper abdomen; left kidney; retroperitoneum	hypoattenuatic
Small rounded enhancing lesion in the pancreatic tail.	0.5 cm x 0.4 cm	pancreatic tail; upper abdomen; abdomen; enhancing	pancreatic tail; upper abdomen; abdomen; pancreas	enhancing
Enhancing mass within the pancreatic tail.	1.9 cm x 1.6 cm	mass	pancreatic body; pancreatic tail; upper abdomen	enhancing
Enhancing mass within the pancreatic body.	1.6 cm x 1.1 cm	mass	pancreatic body; pancreas; upper abdomen; abdomen	enhancing

Table S7: Coverage and frequent structured labels in CT-Bench. Labels are multi-label; therefore, percentages do not sum to 100%.

Field	Coverage	Most frequent labels
Lesion type	20,103 lesions (98.9%)	mass: 5,458 (26.8%); enlargement: 4,887 (24.0%); nodule: 4,006 (19.7%); lymphadenopathy: 3,687 (18.1%); lung nodule: 2,518 (12.4%)
Body part	19,960 lesions (98.2%)	chest: 9,207 (45.3%); abdomen: 9,100 (44.8%); lymph node: 6,425 (31.6%); upper abdomen: 5,953 (29.3%); lung: 4,759 (23.4%)
Imaging property	4,841 lesions (23.8%)	hypoattenuation: 1,575 (7.7%); enhancing: 1,008 (5.0%); large: 404 (2.0%); calcified: 396 (1.9%); solid: 315 (1.5%)

Table S8: Summary of input formats for CLIP-style models and language-model-based vision-language models across CT-Bench QA tasks.

Task	CLIP-style model input	Vision-language model input
`Img2txt`	Image + four candidate descriptions	Prompt + image + question + four descriptions as choices A–D
`CT2txt`	Nine slices encoded individually and combined with weighting + four descriptions	Prompt + nine-slice concatenated image + question + four descriptions as choices A–D
`Txt2img`	Description + four candidate images	Prompt + 2x2 image grid + question + choices A–D corresponding to image locations
`Txt2bbox`	Description + four candidate images with bounding boxes	Prompt + 2x2 image grid with bounding boxes + question + choices A–D corresponding to image locations
`Img2size`	Image + four size descriptions	Prompt + image + question + four size ranges as choices A–D
`Img2attrib`	Image + four attribute descriptions	Prompt + image + question + four attributes as choices A–D
`CT2attrib`	Nine slices encoded individually and combined with weighting + four attribute descriptions	Prompt + nine-slice concatenated image + question + four attributes as choices A–D

Table S9: Summary of model configurations used in CT-Bench experiments. API-based models were evaluated through the corresponding cloud services; open-source models were evaluated locally.

Model	Version / checkpoint	Source	Fine-tuning	Key settings
GPT-4V	2024-02-15-preview	Azure OpenAI API	No	temperature 0; max token length 4000; top-p 0.95
Gemini	Pro 1.5	Google API	No	temperature 0; top-p 1.0; top-k 32; max output tokens 1000
BiomedCLIP	PubMedBERT-based	GitHub / model release	Yes	context length 256
PMC-CLIP	PMC_CLIP_beta	GitHub / model release	No	max length 77
Dragonfly	Biomed checkpoint	GitHub / model release	No	max tokens 1024; temperature 0
RadFM	Released checkpoint	GitHub / model release	Yes	max length 2048; temperature 0
LLaVA-Med	LLaVA-Med v1.5	GitHub / model release	No	temperature 0
BiomedGPT	Released checkpoint	GitHub / model release	No	num beams 5; no-repeat n-gram size 3; max length 1000

Table S10: Full lesion-captioning performance on the CT-Bench Lesion Image and Metadata Set. Results are reported with and without bounding-box guidance. Bold indicates the best zero-shot result, and underlining indicates the best fine-tuned result.

Model	**BLEU-1**		**METEOR**		**ROUGE-1**		**BERTScore F1**		**CosineSim**	
	w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/o	w/
Zero-shot / untuned										
RadFM	0.0713	0.0779	0.0984	0.1034	0.0573	0.0604	**0.8518**	0.8525	0.3262	0.3325
Dragonfly	**0.0842**	**0.1685**	0.1045	**0.2242**	0.0307	**0.1250**	0.8430	**0.8681**	0.2638	0.4082
LLaVA-Med	0.0756	0.0802	0.1555	0.1666	0.0748	0.0804	0.8462	0.8484	0.3914	0.4014
GPT-4V	0.0656	0.0916	0.1405	0.1915	0.0586	0.0819	0.8424	0.8549	0.3895	0.4456
Gemini	0.0754	0.1021	**0.1666**	0.2027	**0.0709**	0.0954	0.8499	0.8582	**0.4149**	**0.4730**
Fine-tuned										
RadFM (w/o BBox)	0.2258	–	0.2142	–	0.1302	–	0.8811	–	0.4567	–
RadFM (w/ BBox)	–	0.2448	–	0.2382	–	0.1507	–	0.8878	–	0.4911

Table S11: Full task-level accuracy comparison on the CT-Bench QA Benchmark test set. "w/o" and "w/" denote evaluation without and with bounding-box guidance, respectively. Bold indicates the best zero-shot result, and underlining indicates the best fine-tuned result. Human evaluation is included as an expert reference and was performed on a sampled subset of cases.

Model / evaluator	**Img2txt**		**CT2txt**		**Txt2img**		**Txt2bbox**		**Img2attrib**		**CT2attrib**		**Img2size**		**Average**		
	w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/o	w/	Total
Zero-shot baselines																	
Random	25.00	25.00	25.00	25.00	25.00	25.00	–	25.00	25.00	25.00	25.00	25.00	25.00	25.00	25.00	25.00	25.00
PMC-CLIP	30.00	33.00	30.50	30.50	24.50	24.50	–	19.00	27.00	28.00	26.00	29.00	20.00	16.00	27.00	26.80	26.89
BiomedCLIP	**38.00**	**42.50**	**37.00**	**42.00**	26.50	**34.50**	–	35.00	44.00	**57.00**	37.00	**42.00**	30.00	38.00	**34.89**	**41.00**	**38.11**
RadFM	25.00	25.00	25.00	24.50	3.50	5.00	–	3.00	18.00	18.00	17.00	19.00	0.00	0.00	15.78	14.90	15.31
LLaVA-Med	23.50	22.50	25.50	25.00	0.00	0.00	–	0.00	21.00	20.00	21.00	21.00	0.00	0.00	15.56	13.60	14.53
Dragonfly	27.00	26.50	24.50	26.50	29.50	26.50	–	26.00	**49.00**	51.00	**44.00**	**42.00**	**37.00**	33.00	32.44	31.10	31.72
GPT-4V	27.50	26.50	24.00	22.00	28.00	25.00	–	30.00	37.00	44.00	31.00	38.00	30.00	23.00	28.56	28.20	28.36
Gemini	27.00	38.50	27.00	25.00	**32.50**	30.00	–	**36.00**	40.00	51.00	36.00	37.00	32.00	**50.00**	31.22	36.10	33.79
Fine-tuned models																	
RadFM w/o box	0.00	0.00	0.00	0.00	0.00	0.00	–	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RadFM w/ box	0.00	0.00	0.00	0.00	0.00	0.00	–	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BiomedCLIP w/o box	42.00	42.00	41.00	43.00	32.00	36.00	–	29.00	54.00	45.00	46.00	48.00	43.00	36.00	41.44	40.00	40.66
BiomedCLIP w/ box	38.50	68.00	40.50	62.00	29.00	57.50	–	67.00	44.00	67.00	44.00	65.00	30.00	46.00	37.11	62.00	50.26
Human evaluation																	
Senior radiologist 1	100.0	90.0	80.0	60.0	80.0	70.0	–	100.0	60.0	100.0	40.0	100.0	40.0	100.0	66.0	90.0	79.20
Senior radiologist 2	70.0	100.0	80.0	70.0	80.0	70.0	–	100.0	60.0	100.0	60.0	100.0	80.0	100.0	70.5	90.5	81.50
Junior doctor	50.0	80.0	70.0	50.0	30.0	40.0	–	80.0	60.0	80.0	40.0	100.0	0.0	40.0	45.5	68.5	58.15

Table S12: Validation-set accuracy on the CT-Bench QA Benchmark. "w/o" and "w/" denote evaluation without and with bounding-box guidance, respectively. The `Txt2bbox` task is defined only in the bounding-box setting.

Model	**Img2txt**		**CT2txt**		**Txt2img**		**Img2size**		**Txt2bbox**	**Img2attrib**		**CT2attrib**	
	w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/	w/o	w/	w/o	w/
Random	25.00	25.00	25.00	25.00	25.00	25.00	25.00	25.00	25.00	25.00	25.00	25.00	25.00
PMC-CLIP	23.00	23.00	22.00	24.00	24.50	24.50	20.00	16.00	19.00	24.00	24.00	22.00	20.00
BiomedCLIP	36.00	37.00	35.00	34.00	26.50	34.50	44.00	44.00	35.00	42.00	50.00	37.00	42.00
RadFM	20.00	20.00	17.00	19.00	9.00	7.00	0.00	0.00	0.00	18.00	18.00	17.00	19.00
LLaVA-Med	21.00	21.00	25.50	25.00	0.00	0.00	0.00	0.00	0.00	18.00	18.00	28.00	22.00
Dragonfly	27.00	28.00	22.00	22.00	27.00	20.00	34.00	34.00	26.00	49.00	51.00	44.00	42.00
GPT-4V	33.00	37.00	31.00	36.00	28.00	31.00	22.00	40.00	32.00	30.00	44.00	26.00	26.00
Gemini	34.00	39.00	34.00	36.00	29.00	40.00	48.00	50.00	36.00	40.00	40.00	40.00	40.00

Table S13: Average inference time of baseline models. GPT-4V and Gemini were accessed through cloud APIs, and latency includes network and server queue time; therefore, cloud and local inference times are not directly comparable.

Model	Avg. inference time (s)	Deployment type
BiomedCLIP	≈0.06	Local (A100 GPU)
PMC-CLIP	≈0.06	Local (A100 GPU)
RadFM	≈0.77	Local (A100 GPU)
LLaVA-Med	≈2.14	Local (A100 GPU)
Dragonfly	≈4.29	Local (A100 GPU)
GPT-4V	≈2.98	API (Azure Cloud)
Gemini	≈2.98	API (Google Cloud)

Table S14: Representative lesion-captioning case study. The ground-truth description is compared with outputs from untuned and fine-tuned models.

Model	Output	Alignment	Key observations
Ground truth	A nodule in the superior segment of the left lower lobe.	–	–
RadFM	There are multiple nodular opacities scattered throughout the bilateral lungs. The largest is in the right upper lobe with a draining vessel.	Misaligned	Overinterprets findings and mislocalizes the lesion.
Tuned RadFM (w/ BBox)	Left lower lobe nodule.	Most aligned	Correctly identifies a nodule in the left lower lobe but lacks the specific superior-segment location.
LLaVA-Med	The image is a CT scan of the chest. It shows a large mass in the right upper lobe of the lung.	Misaligned	Mislocalizes the finding and describes it as a mass rather than a nodule.
GPT-4V	The image shows a transverse section of a chest CT scan with a well-circumscribed, round, hyperdense lesion in the right lung parenchyma.	Misaligned	Incorrectly identifies the lesion in the right lung and lacks diagnostic specificity.
Gemini	There is a small, well-defined nodule in the right lower lobe. The nodule is approximately 5 mm in diameter. No other significant findings.	Misaligned	Mislocalizes the nodule and adds unsupported size information.
Dragonfly	CT scan of the chest showing a 1.1 cm nodule in the left lower lobe.	Most aligned	Correctly identifies a nodule in the left lower lobe, but does not specify the superior segment and adds size information not in the ground truth.