Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Developing Generalist Foundation Models from a Multimodal Dataset for 3D Computed Tomography

Ibrahim Ethem Hamamci

ibrahim.hamamci@uzh.ch

University of Zurich / ETH AI Center

Sezgin Er

Istanbul Medipol University

Furkan Almas

Istanbul Medipol University

Ayse Gulnihan Simsek

Istanbul Medipol University

Sevval Nil Esirgun

Istanbul Medipol University

Irem Dogan

Istanbul Medipol University

Muhammed Furkan Dasdelen

Istanbul Medipol University

Omer Faruk Durugol

Istanbul Medipol University

Bastian Wittmann

University of Zurich https://orcid.org/0000-0002-6308-2926

Tamaz Amiranashvili

University of Zurich / ETH AI Center https://orcid.org/0000-0001-8914-3427

Enis Simsar

ETH Zurich

Mehmet Simsar

Istanbul Medipol University

Emine Bensu Erdemir

Istanbul Medipol University

Abdullah Alanbay

Istanbul Medipol University

Anjany Sekuboyina

University of Zurich / ETH AI Center

Berkan Lafci

University of Zurich / ETH AI Center
Christian Bluethgen
University Hospital Zurich https://orcid.org/0000-0001-7321-5676
Mehmet Kemal Ozdemir
Istanbul Medipol University
Bjoern Menze
University of Zurich https://orcid.org/0000-0003-4136-5690

Biological Sciences - Article

Keywords:

Posted Date: October 28th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-5271327/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Version of Record: A version of this preprint was published at Nature Biomedical Engineering on February 12th, 2026. See the published version at https://doi.org/10.1038/s41551-025-01599-y.

Developing Generalist Foundation Models from a Multimodal Dataset for 3D Computed Tomography

Ibrahim Ethem Hamamci[1,2,†,*], Sezgin Er[3,†], Furkan Almas[3], Ayse Gulnihan Simsek[3], Sevval Nil Esirgun[3], Irem Dogan[3], Muhammed Furkan Dasdelen[3], Omer Faruk Durugol[3], Bastian Wittmann[1,2], Tamaz Amiranashvili[1,2], Enis Simsar[4], Mehmet Simsar[3], Emine Bensu Erdemir[3], Abdullah Alanbay[3], Anjany Sekuboyina[1,2], Berkan Lafci[1,2], Christian Bluethgen[1], Mehmet Kemal Ozdemir[3], and Bjoern Menze[1,2]

[1] Department of Quantitative Biomedicine, University of Zurich, Zurich, Switzerland
[2] ETH AI Center, ETH Zurich, Zurich, Switzerland
[3] International School of Medicine, Istanbul Medipol University, Istanbul, Turkey
[4] Department of Computer Science, ETH Zurich, Zurich, Switzerland

†These authors contributed equally to this work. *Corresponding author: {ibrahim.hamamci@uzh.ch}

Abstract. While computer vision has achieved tremendous success with multimodal encoding and direct textual interaction with images via chat-based large language models, similar advancements in medical imaging AI—particularly in 3D imaging—have been limited due to the scarcity of comprehensive datasets. To address this critical gap, we introduce CT-RATE, the first dataset that pairs 3D medical images with corresponding textual reports. CT-RATE comprises 25,692 non-contrast 3D chest CT scans from 21,304 unique patients. Through various reconstructions, these scans are expanded to 50,188 volumes, totaling over 14.3 million 2D slices. Each scan is accompanied by its corresponding radiology report. Leveraging CT-RATE, we develop CT-CLIP, a CT-focused contrastive language-image pretraining framework designed for broad applications without the need for task-specific training. We demonstrate how CT-CLIP can be used in two tasks: multi-abnormality detection and case retrieval. Remarkably, in multi-abnormality detection, CT-CLIP outperforms state-of-the-art fully supervised models across all key metrics, effectively eliminating the need for manual annotation. In case retrieval, it efficiently retrieves relevant cases using either image or textual queries, thereby enhancing knowledge dissemination. By combining CT-CLIP's vision encoder with a pretrained large language model, we create CT-CHAT, a vision-language foundational chat model for 3D chest CT volumes. Finetuned on over 2.7 million question-answer pairs derived from the CT-RATE dataset, CT-CHAT surpasses other multimodal AI assistants, underscoring the necessity for specialized methods in 3D medical imaging. Collectively, the open-source release of CT-RATE, CT-CLIP, and CT-CHAT not only addresses critical challenges in 3D medical imaging but also lays the groundwork for future innovations in medical AI and improved patient care.

Introduction

Three-dimensional (3D) medical imaging, such as computed tomography (CT) and magnetic resonance imaging (MRI), provides a more detailed and comprehensive view of a patient's condition compared to two-dimensional (2D) imaging [1]. Among these, CT stands out due to its speed and widespread availability, making it the preferred modality for diagnosing various abnormalities [2]. As global demand for CT scans continues to grow, the need for rapid and accurate diagnostic results becomes more pressing. However, the manual generation of reports from these complex images is time-consuming and prone to errors, underscoring the urgent need for automation [3]. Artificial intelligence (AI) has the potential to address this challenge by efficiently identifying pathologies, reducing the workload on radiologists, and potentially saving lives in situations where immediate radiological expertise is not available.

AI has already begun to revolutionize the interpretation of medical images [4, 5, 6, 7, 8]. Recent advances in automated diagnosis, classification, and segmentation have been driven by numerous publicly available datasets [9, 10, 11, 12, 13]. Beyond supervised methods, general AI systems have been developed using open-source image-text pairs in self-supervised learning, particularly in 2D imaging [14]. These versatile AI foundation models integrate comprehensive natural language annotations into the learning process without requiring task-specific training, thereby eliminating the need for explicit manual image annotations. Unconstrained by predefined classes or tasks during training, they enhance adaptability and have been successfully applied to tasks such as disease classification [15], case retrieval [16], and interactive diagnostics via chat interfaces [14]. Despite these significant advancements in 2D imaging, the application of advanced AI techniques to 3D imaging remains underexplored. A major challenge is the lack

of open-source 3D medical imaging datasets paired with textual reports, which require extensive resources, specialized knowledge, and technical expertise to create [17, 18, 19]. Currently, the largest 3D medical imaging datasets available contain only a few thousand cases and do not include paired textual data [20, 21, 22], limiting the development of advanced AI models. Moreover, 3D models face higher computational demands and have fewer available pretrained open-source models compared to their 2D counterparts [23, 24]. Previous studies have only begun to address these challenges; for example, one involving chest CT is limited by its supervised approach, restricting detection to predefined abnormalities and offering only a small portion of its dataset publicly without paired textual reports [20].

To address these challenges, we present three related contributions: First, we introduce CT-RATE, a dataset that pairs 50,188 reconstructed 3D chest CT volumes (over 14.3 million 2D slices) from 25,692 scans of 21,304 patients with corresponding radiology reports. Each volume is annotated with 18 abnormalities, supporting extensive validation and fine-tuning. CT-RATE is fully open-source, making it the first public 3D medical imaging dataset with paired textual reports and the largest CT dataset with multi-abnormality labels. Second, leveraging CT-RATE, we develop CT-CLIP, a novel CT-focused contrastive language-image pretraining framework. CT-CLIP integrates extensive natural language annotations into the learning process, enabling the model to comprehend image-based semantic knowledge and perform various tasks without specific training. Our evaluation shows CT-CLIP's ability to detect multiple abnormalities in a zero-shot setting [25], outperforming state-of-the-art fully supervised methods on two test sets from different countries. These findings indicate that explicit labels are not essential for achieving high performance in 3D medical imaging interpretation, given the availability of corresponding radiology reports. Third, to extend CT-CLIP's capabilities, we combine its vision encoder with a pretrained large language model to create CT-CHAT, a vision-language chat model for 3D chest CT volumes. CT-CHAT is fine-tuned with over 2.7 million question-answer pairs derived from the CT-RATE dataset. CT-CHAT significantly outperforms open-source vision-language AI assistants, such as LLaVA 1.6 [26], LLaVA-Med [27], and CXR-LLaVA [28], underscoring the need for specialized tools in 3D medical imaging.

In this article, we introduce pivotal contributions to medical AI. We address the scarcity of open-source 3D imaging datasets paired with textual reports by introducing CT-RATE. Then, we create CT-CLIP, a versatile foundation model that effectively integrates linguistic and visual data from CT-RATE via contrastive learning [29]. We demonstrate CT-CLIP's remarkable adaptability and performance across a variety of tasks without task-specific training. Moreover, we develop a vision-language AI assistant, CT-CHAT, by finetuning a pretrained large language model through visual instruction using CT encodings created by CT-CLIP's vision encoder. The open-source release of CT-RATE, CT-CLIP, and CT-CHAT addresses critical challenges in 3D medical imaging, such as dataset scarcity, model versatility, and learning paradigms, leading to groundbreaking research and substantial advancements in the field.

Results

CT-RATE: Pairing CT volumes with radiology text reports

We curate and open-source CT-RATE, a dataset comprising 25,692 non-contrast 3D chest CT volumes and corresponding radiology text reports from 21,304 unique patients (Figure **1a**). These 3D volumes, reconstructed using various methods to suit different window settings [30], total over 14.3 million 2D slices. CT-RATE is divided into training and internal validation sets for experiments (see **creating the CT-RATE dataset** section of **Methods**). Additionally, to evaluate model performance under distribution shifts, we incorporate RAD-ChestCT, an external validation set from an independent study in a different country [20] (see **adding an external validation dataset to CT-RATE** section of **Methods**). Importantly, we do not use the RAD-ChestCT dataset during the training of our models, all development is conducted exclusively using the CT-RATE dataset.

Conventional supervised machine learning approaches for medical image classification necessitate manual annotation of large datasets, a process that requires significant time and expertise [31]. Although our self-supervised learning approach does not require abnormality labels, the baseline supervised model and finetuning methods do. To meet this need, we enrich CT-RATE with global abnormality labels extracted from the radiology reports using an automated text classifier (Figure **1b**). This process involves finetuning a pretrained language model, RadBERT [32], similar to recent works [33]. Within CT-RATE, we manually annotate 1,000 reports for 18 distinct abnormalities for this finetuning procedure. The finetuned model's precision, recall, and F1 scores, detailed in Supplementary Table **2**, show robust classification across abnormalities, ensuring the CT-RATE dataset is enriched with high-quality labels.

Figure **2** summarizes the CT-RATE dataset alongside RAD-ChestCT. Our dataset spans a wide age range (18-102 years) (Figure **2a**) and includes diverse pathologies related to both heart and lung

conditions (Figure **2b**). The reports vary in length and are authored by numerous radiologists, reflecting a variety of writing styles (Figure **2c**). The dataset also varies in CT slice numbers (Figure **2d**), slice thickness, axial pixel (XY) spacing (Figure **2e**), and includes data from different CT manufacturers and a balanced sex distribution (Figure **2f**). This heterogeneity is crucial for developing robust models capable of handling out-of-distribution cases. Supplementary Table **1** provides detailed statistics on abnormalities, manufacturers, and sex distributions in the training and internal validation sets. In conclusion, CT-RATE does not represent a curated study dataset, but rather reflects real-world scenarios characterized by heterogeneity. This diversity ensures that the models trained on CT-RATE are well-suited for deployment in clinical settings, offering reliable performance across a broad range of conditions and imaging scenarios.

CT-CLIP: Developing a visual-language foundation model with CT-RATE

Unlike other methods utilizing supervised machine learning for multi-abnormality classification of CT volumes, our study employs radiology reports, which provide a rich source of semantic knowledge. This significantly enhances the understanding of CT volumes and supports a wide range of potential zero-shot applications. Accordingly, we develop CT-CLIP, the first foundational model for chest CT-focused contrastive language-image pretraining, adapting a technique from a previous study [34]. We utilize both the impression and findings sections of radiology reports combined for CT-CLIP due to their superior performance compared to utilizing either one alone, as demonstrated by the abnormality classification scores in an ablation study (Supplementary Figure **1a**).

CT-CLIP comprises a vision transformer and a text transformer that generate two distinct embedding vectors: one for CT volumes and another for reports, respectively (see **developing the CT-CLIP model** section of **Methods**). The vision transformer is adapted from CT-ViT [23], incorporating a two-stage spatial and causal transformer, enabling 3D attention across volumes. CT-CLIP is trained to maximize the similarity between embedding vectors for paired volume-text data, while minimizing similarity for unpaired data. Our training approach leverages contrastive learning [35] (Figures **1c**, **1d**), which enables the training of a foundation model suitable for various applications without task-specific training, including multi-abnormality detection, volume-to-volume retrieval, and report-to-volume retrieval. Subsequent sections demonstrate how CT-CLIP can be applied to these downstream tasks without additional training. To further enhance CT-CLIP's capabilities, we explore finetuning methods for detection to achieve greater accuracy and adapt CT-CLIP to create a vision-language AI assistant.

We analyze the latent spaces of reports and CT volumes generated by CT-CLIP, by computing t-SNE projections of these spaces. Figure **4d** shows the distribution according to the number of abnormalities present in a case. Figure **4e** depicts the age distribution on the t-SNE projections. The well-clustered t-SNE projections confirm CT-CLIP's ability to accurately map latent spaces with closely aligned visual and linguistic information. For instance, in both image and text latents, the alignment between age and abnormality distributions indicates that older patients are more prone to abnormalities, which the model effectively identifies. Extended Data Figures **1**, **2** provide additional t-SNE projections for each abnormality using volume and report embeddings, respectively.

Surpassing supervised levels with CT-CLIP's zero-shot multi-abnormality detection

CT-CLIP adopts a zero-shot classification approach for multi-abnormality detection without explicit training on specific abnormality labels. Zero-shot classification refers to the ability of a model to categorize previously unseen classes without being explicitly trained on those specific categories [34]. Instead of relying on predefined labels, CT-CLIP leverages the semantic information embedded in paired radiology reports to infer multiple abnormalities. For instance, a model trained to recognize features like "reticulation," "consolidation," or "nodules" in common lung diseases could potentially identify a rare lung condition it was not explicitly trained on by recognizing combinations of these features.

Zero-shot classification with foundation models is typically applied as a multi-class classification task, assigning a single label to each image [34]. However, a medical image may present multiple abnormalities, necessitating a unique approach. Inspired by recent work on 2D chest radiographs [15], we generate positive and negative prompts for each abnormality. The embeddings for the input 3D CT volume and both prompts are projected into a shared latent space, and the contrastive loss between these projections is calculated based on the distance between the image and language spaces (see **CT-CLIP for zero-shot multi-abnormality detection** section of **Methods** for more details). This process identifies the relevant features in both modalities. The model then outputs logits for each prompt, which are interpreted as the likelihood of abnormality presence, as shown in Figure **1e**. We conduct a systematic evaluation of CT-CLIP's zero-shot detection capabilities on both internal and external validation sets (Figure **4a**).

The careful engineering of prompts is crucial in zero-shot classification tasks with foundation models, as it significantly impacts performance [36]. Therefore, we evaluate seven distinct prompts commonly used in our reports, with detailed information provided in the online methods section. Figure **4b** presents the mean AUROC, F1 score, accuracy, and precision for these various prompts. Based on its consistently superior performance, we use prompt 7 ("{Abnormality} is {∅/not} present.") in further experiments (see **CT-CLIP for zero-shot multi-abnormality detection** section of **Methods** for more details).

To evaluate CT-CLIP's zero-shot multi-abnormality detection, we benchmark it against CT-Net [20], a state-of-the-art fully supervised model for 3D chest CT volumes. CT-Net requires manual annotation for each abnormality, which demands significant time and expertise [31]. We choose CT-Net as a baseline to demonstrate that CT-CLIP can surpass traditional fully supervised performance without relying on manual annotations or being restricted to predefined classes. Since the pretrained CT-Net model is not available and only a small portion of its dataset is open-source, we train CT-Net on CT-RATE using our abnormality labels for a fair comparison (see **benchmarking the CT-CLIP models for multi-abnormality detection** section of **Methods**). Despite using a different dataset for training, the retrained CT-Net's classification scores are comparable to those reported in the original study [20]. Figure **4a** shows the retrained CT-Net's performance metrics, while Extended Data Figures **3**, **4** and Supplementary Figures **2**, **6** detail classification scores and confusion matrices.

We compare our zero-shot approach to this supervised baseline and observe significantly higher performance metrics for our approach. Specifically, our zero-shot CT-CLIP model yields a significantly higher mean AUROC (0.128 higher in internal validation and 0.085 higher in external validation), mean F1 score (0.076 higher in internal validation and 0.055 higher in external validation), mean accuracy (0.87 higher in internal validation and 0.055 higher in external validation), and mean precision (0.084 higher in internal validation and 0.051 higher in external validation) compared to the supervised baseline. Results on the external validation set demonstrate that the zero-shot CT-CLIP model performs well not only on similar data distributions but also under distribution shifts, indicating its strong generalization to external datasets. A two-sided unpaired permutation test, utilizing 10,000 permutations, highlights statistically significant improvements ($p < 0.05$) in AUROC, F1 score, and accuracy for CT-CLIP as a zero-shot classifier over the supervised (baseline) model, detailed in Supplementary Table **5**.

Extended Data Figures **3**, **4** provide scores for all abnormalities on the internal and external validation sets, respectively. Supplementary Figures **3**, **7** show confusion matrices on the internal and external validation sets, respectively. Notably, on the internal validation set, all abnormalities (18 out of 18) exhibit higher scores in our zero-shot approach compared to the supervised baseline in mean AUROC. In conclusion, CT-CLIP excels by just leveraging existing radiology text reports in electronic health records, thus avoiding the need for expensive manual annotations. Moreover, CT-CLIP's zero-shot capability offers flexible predictions beyond predefined classes and adapts efficiently to changing needs post-training.

Since the availability of sufficiently large datasets is crucial for machine learning, especially for transformer models, which require more data than convolutional neural networks [37, 38], we explore the impact of dataset size on classification performance. We train CT-CLIP on different portions of the CT-RATE dataset, specifically 9.8% (equivalent to the open-source portion of the RAD-ChestCT dataset [20], currently the largest publicly available chest CT dataset), 20%, 40%, 60%, 80%, and 100%. Figure **4c** reports the classification performance across different dataset sizes. Notably, a consistent trend emerges: larger dataset sizes correlate with improved scores. This finding highlights the role of dataset size in enhancing CT-CLIP's zero-shot classification performance and underscores the value of our extensive CT-RATE dataset for computational research. It also offers a clear avenue for future improvements, suggesting that further expanding the dataset or incorporating more diverse data could lead to even better performance.

Enhancing CT-CLIP's detection performance with finetuning

To enhance CT-CLIP's multi-abnormality detection performance, we explore a task-specific finetuning method, ClassFine, inspired by conventional linear probing, by adding a new classification layer to the pretrained vision transformer of CT-CLIP (Figure **3a**). Supplementary Figure **1b** indicates that training only the final layer may be sufficient for finetuning (see **finetuning CT-CLIP for multi-abnormality detection** section of **Methods**). While the ClassFine approach improves detection performance (Figure **4a**), it restricts the model to predefined labels as any supervised classification method. Extended Data Figures **3**, **4** present abnormality scores, with confusion matrices in Supplementary Figures **5**, **9** on the internal and external validation sets, respectively. To address the limitations of ClassFine, we develop the VocabFine approach, a novel open-vocabulary finetuning method inspired by WISE-FT [39] (see **finetuning CT-CLIP for multi-abnormality detection** section of **Methods**). VocabFine, like CT-CLIP's zero-shot inference, is finetuned to improve label prediction accuracy (Figure **3c**). Specifically,

we use positive and negative prompts for abnormalities, calculate the logits for these volumes with pretrained CT-CLIP, and align these logits with the ground truth labels. While the finetuned model with VocabFine retains the open-vocabulary inference capability of CT-CLIP (Figure **3d**), it also consistently outperforms ClassFine (Figure **4a**), further emphasizing the value of our open vocabulary finetuning. Abnormality-specific improvements are detailed in Extended Data Figures **3**, **4**, while Supplementary Figure **1b** underscores the importance of training all weights in finetuning. Supplementary Figures **4**, **8** provide confusion matrices on the internal and external validation sets, respectively.

Retrieving relevant cases with CT-CLIP

In clinical practice, radiologists frequently refer to past cases with similar findings to guide diagnosis and treatment decisions. This process, known as case retrieval, enhances diagnostic accuracy and builds confidence in decision-making. However, manually searching for comparable cases is time-consuming and inefficient. An automated, robust solution can streamline case retrieval, saving both time and effort.

Beyond zero-shot detection, CT-CLIP serves as a powerful tool for image-to-image retrieval [40], identifying the relevant volumes from a query volume by comparing cosine similarities between the embeddings (Figure **5a**). We test on both internal and external validation sets using the vision transformer of CT-CLIP, finetuned models, and the supervised baseline. We utilize the Mean Average Precision at K (MAP@K) metric, where K represents the number of returned images (see **benchmarking the CT-CLIP models for CT volume retrieval** section of **Methods**). As K increases, MAP@K scores typically decrease due to less accurate matches. Results, as in Figure **5c**, show comparable performance among methods, with all significantly outperforming random retrieval. Although external validation MAP@K scores are higher, the fold increase over random retrieval is lower (three times on the external vs. six times on the internal set), due to differing abnormality distributions. This suggests that examining fold change offers a more accurate comparison. Finetuning does not significantly alter the vision transformer's weights, as CT-CLIP and the finetuned models perform similarly across MAP@K metrics.

CT-CLIP also functions as a text-to-image retrieval tool [41], evaluating cosine similarities between target volume embeddings and query report embeddings (Figure **5b**). Given the limitations of the supervised model and the ClassFine approach in encoding text, we employ CT-CLIP and the VocabFine approach for our experiments. Due to the lack of reports in the RAD-ChestCT dataset, only the internal validation set is used for report-to-volume retrieval. We evaluate performance using Recall@K, where K represents the number of returned images, reflecting the percentage of correctly matched images within the top K returns (see **benchmarking the CT-CLIP models for CT volume retrieval** section of **Methods**). While Recall@K scores increase with more returned volumes, considering fold change relative to random performance is more meaningful. Figure **5d** shows that CT-CLIP performs significantly higher than the random baseline. However, the Recall@K scores for VocabFine are substantially lower than CT-CLIP, consistent with the expectation that finetuning a foundation model for a specific task might diminish its efficacy for other tasks. The reduced performance with VocabFine in report-to-volume retrieval further indicates significant modifications to the text transformer weights due to finetuning.

CT-CHAT: Developing a multimodal AI assistant with CT-RATE and CT-CLIP

Observing CT-CLIP's remarkable capabilities for open-vocabulary inference, we build on this foundation to enable an interactive exchange between textual queries and diagnostic image information. To achieve this, we develop CT-CHAT, a multimodal generalist AI assistant specifically designed for 3D chest CT. CT-CHAT integrates both visual and language processing to perform various tasks, making it a versatile tool for clinical use. To train CT-CHAT, we create a visual question-answering (VQA) dataset, comprising over 2.7 million question-answer pairs derived from CT-RATE using Llama 3.1 7B [42] (see **creating the dataset for CT-CHAT** section of **Methods**). This dataset consists of several components, including long-answer questions, short-answer questions, multiple-choice questions, and report generation tasks (Extended Data Figure **7**), reflecting the diverse types of inquiries encountered in clinical practice.

CT-CHAT's architecture leverages the strengths of CT-CLIP by using its pretrained vision transformer to encode 3D chest CT volumes. These encodings are then passed to a large language model (LLM) through a multimodal projector, as shown in Figure **6a**. We perform an ablation study evaluating CT-CHAT's performance with 4 different state-of-the-art LLMs, as detailed in Figure **6d**. We select 70B Meta Llama 3.1 [42], outperforming others in all key metrics, for the optimal configuration of CT-CHAT.

We create CT-CHAT by finetuning the multimodal projector and Low-Rank Adaptation (LORA) weights [43] for the LLM (see **developing the CT-CHAT model** section of **Methods**). Prior to this, we perform a medical concept alignment step, where the LLM is frozen, and the multimodal projector

is trained solely to generate radiology reports. This step slightly enhances performance by aligning the multimodal projector with biomedical concepts (Supplementary Table **10b**). The ablation study for the volume provision proves CT-CHAT extracts relevant information from the given 3D volumes (Supplementary Table **10a**). Although CT-CHAT shows the ability to retrieve information from the given image, it is also expected to hallucinate some pathologies when the input volume is not provided. Thus, we implement a guardrail after the ablation study, with the system prompt (Supplementary Table **9**). Extended Data Figure **9** provides an example of the guardrail in case the volume is not provided.

Finally, CT-CHAT generates contextually relevant and clinically accurate responses across a variety of tasks, from simple question-answering to the generation of detailed radiology reports as shown in Extended Data Figure **8**. The model is also capable of dynamic task switching, facilitated by special tokens that guide the LLM in generating the appropriate type of response based on the specific task. This flexibility makes CT-CHAT a powerful tool for enhancing radiologists' workflows, reducing the time and effort required for interpretation, and improving diagnostic accuracy in 3D medical imaging.

Surpassing other multimodal AI assistants with CT-CHAT

Given the absence of generalist models for 3D imaging, we evaluate CT-CHAT against several state-of-the-art open-source vision-language AI models, including LLaVA 1.6 (Mistral 7B and Vicuna 13B) [26], LLaVA-Med [27], and CXR-LLaVA [28]. To ensure a fair comparison, these models, designed for 2D imaging tasks, are tested using Digitally Reconstructed Radiographs (DRRs) of 3D chest CT volumes—a method that ensures accurate representation in a 2D context, as suggested by [44], rather than using random or central 2D slices of a 3D volume (see **benchmarking the CT-CHAT model** section of **Methods**). In contrast, CT-CHAT directly processes 3D chest CT, giving it a significant advantage in understanding and interpreting the complex anatomical and pathological information within these scans.

CT-CHAT generates clinically accurate and detailed answers to questions about the given CT image, as shown in Figure **6b**. To evaluate the performance, we employ multiple metrics, including BLEU-1, METEOR, ROUGE-L, and CIDEr, which assess the quality of the generated text. Additionally, inspired by previous work [27], we utilize an LLM-based metric, named the Llama score, to evaluate the clinical accuracy of responses, with the models' outputs compared against ground truths (see **benchmarking the CT-CHAT model** section of **Methods**). The results indicate that CT-CHAT, being capable of evaluating the full information of 3D CT volumes, consistently outperforms the other open-source vision-language AI assistants across all evaluated tasks. It shows superior ability in generating long and short answers, handling multiple-choice questions, and producing detailed radiology reports that align closely with clinical expectations, as shown in Figure **6c**. Extended Data Figure **5** provides a detailed breakdown, demonstrating that CT-CHAT significantly outperforms others in all tasks.

There are two key reasons why CT-CHAT significantly outperforms 2D-based models: (1) Image: CT-CHAT leverages the 3D spatial information processed by CT-CLIP's pretrained vision transformer, which captures much richer and more detailed anatomical structures than 2D models—a critical advantage for CT interpretation. (2) Text: CT-CHAT is trained on 3D CT volumes paired with radiology-specific linguistic data from CT-RATE, enabling it to generate highly precise and clinically relevant responses, far surpassing those produced by models trained on general-purpose texts. This powerful combination of CT-CLIP's capacity to handle complex 3D imaging data and the domain-specific paired image-text data from CT-RATE allows CT-CHAT to push the boundaries of multimodal tasks in 3D medical imaging. In summary, CT-CHAT not only surpasses existing 2D tools but also provides a scalable solution for enhancing chest CT interpretation, underscoring the need for generalist models in 3D medical imaging.

Discussion

Radiology reports provide comprehensive descriptions of medical images without adding extra labeling burden, as they are already part of the daily routine for radiologists [45]. Recent studies show that foundational models, trained on paired radiology reports and 2D medical images, perform comparably to fully supervised methods in various tasks [15]. However, these studies have not extended to 3D medical images, primarily because a publicly available 3D medical imaging dataset paired with text reports is absent. Therefore, CT-RATE, which pairs 3D chest CT volumes with radiology reports (Figure **1a**), represents a significant advancement as the first of its kind. In this study, we introduce CT-RATE, a novel and comprehensive dataset, which we first utilize to develop CT-CLIP (Figure **1c**). To our knowledge, CT-CLIP is the first foundation model in 3D medical imaging utilizing radiology text reports.

CT-CLIP provides a versatile solution for various tasks without requiring task-specific training. Unlike supervised methods, it eliminates the need for explicit labels and enables predictions beyond predefined

classes (Figure **1e**), making it particularly valuable when learning objectives evolve post-training. CT-CLIP's zero-shot multi-abnormality detection outperforms the state-of-the-art supervised method on the internal validation set across all metrics (Figure **4a**). Importantly, this experiment suggests that explicit labels are not necessary for high performance in 3D medical image interpretation when corresponding radiology reports are available for training. Furthermore, we externally validate CT-CLIP on an independent dataset, Rad-ChestCT, from a different country (Figure **2e**). The model's ability to generalize to different data distributions, a primary challenge in deploying medical AI [46], is confirmed as it outperforms the fully supervised method in all metrics in the external validation set as well (Figure **4a**). To establish a fully supervised baseline and fine-tune CT-CLIP, we extract multi-abnormality labels from radiology reports, making them publicly available. Following recent examples, we develop an automatic label extractor by finetuning a language model (Figure **1b**) [47]. However, developing such a system requires significant time, expertise, and domain knowledge, as evidenced by our effort in manually annotating a small portion of the reports. CT-CLIP, as a zero-shot classifier, overcomes this bottleneck by detecting abnormalities at a supervised level without the need for an automatic labeler. Finetuning experiments further confirm CT-CLIP's representational learning capabilities, demonstrating superior performance (Figure **4a**). Additionally, the effective clustering observed in t-SNE projections (Figure **4d-e**) unveils another aspect of CT-CLIP's clinical potential. The successful clustering of abnormalities (Extended Data Figures **1** and **2**) suggests that CT-CLIP could extend its utility beyond multi-abnormality detection to assist in prognosis, disease grading, and the calculation of lung-related parameters in future research.

Analyzing CT-CLIP's open-vocabulary inference capabilities, we observe significant potential for developing multimodal algorithms that address broader and more diverse clinical applications. To expand these capabilities, we finetune four state-of-the-art LLMs through visual instruction tuning with the CT tokens created by CT-CLIP, resulting in the creation of CT-CHAT. For finetuning, we create and open-source a visual question-answering dataset derived from CT-RATE (Extended Data Figure **7**). As a result, CT-CHAT significantly outperforms existing 2D models, underscoring the urgent need for generalist models specialized in 3D medical imaging (Figure **6c**). We find the choice of LLM (Supplementary Table **10a**) and the use of concept alignment pretraining (Supplementary Table **10b**) have minimal impact on CT-CHAT's accuracy. In contrast, including the volume as input is crucial, particularly in report generation (Supplementary Table **10b**). This highlights the value of the CT-RATE dataset and CT-CLIP in developing a generalist AI assistant for 3D medical imaging. Also, we demonstrate that engineering the system prompt of CT-CHAT (Supplementary Table **9**) helps mitigate pathology hallucinations when the volume is not provided (Extended Data Figure **9**). Furthermore, our model, with new special tokens, is capable of responding in various modes, including long and short answers, multiple-choice, and report generation (Extended Data Figure **8**). To broaden accessibility, we introduce a user-friendly graphical interface that allows non-experts to interact with CT-CHAT effortlessly. This interface not only makes CT-CHAT a valuable tool for educational and clinical settings but also invites a wider audience to engage with and contribute to the evolution of interactive AI-driven inference in clinical practice.

Despite these advancements, our CT-RATE dataset and the CT-CLIP and CT-CHAT models have certain limitations. First, biases inherent to our single-institution dataset may influence the performance of our self-supervised training. Although we curate a large and diverse set of volume-report pairs, the data originates from a single geographic and clinical setting, which may limit the generalizability of our models to other populations or institutions. We conduct an ablation study to assess the impact of training set size on CT-CLIP's performance, revealing that performance increases with larger training sets (Figure **4c**). This suggests that incorporating data from multiple institutions and more diverse patient populations could further enhance model robustness and generalizability. Second, CT-CLIP depends on the specific terminology used in reports and the choice of prompts during zero-shot detection. Variability in how diseases are described across radiology reports can affect detection accuracy. To assess the importance of prompt engineering, we evaluate zero-shot multi-abnormality detection performance using seven different prompts during inference (Figure **4b**). While we select the best-performing prompt from commonly used structures in radiology reports, continued optimization of prompts may further improve performance. Third, the evaluation of CT-CHAT faces limitations due to the lack of pretrained 3D medical multimodal AI assistants for benchmarking. As a result, we compare CT-CHAT against 2D models, which may not fully capture the 3D spatial information inherent in CT volumes. To address this, we utilize digitally reconstructed radiographs (DRRs) generated from our dataset [44]. This approach is suggested by prior work showing that models trained with DRRs can yield comparable results to those trained with 3D chest CT volumes [44]. However, the absence of true 3D benchmarks highlights the need for further development and validation of multimodal AI assistants in 3D medical imaging. Finally, our current models are limited to chest CT volumes. To broaden their clinical applicability, future work should focus on extending CT-RATE, CT-CLIP, and CT-CHAT to include other 3D imaging modalities such as MRI and PET scans,

as well as other anatomical regions. Expanding to different modalities and body parts would enhance the utility of our models across a wider range of medical imaging tasks and clinical scenarios.

In conclusion, the development of CT-CLIP and CT-CHAT marks a significant advancement in 3D medical imaging, particularly for interpreting chest CT volumes. Enabled by the creation of CT-RATE, the first open-source 3D imaging dataset paired with radiology reports, these models set a benchmark for multimodal AI in 3D medical imaging. The open-source availability of CT-RATE, CT-CLIP, and CT-CHAT is expected to be a valuable asset to the medical AI community, providing a solid foundation for future research. These foundational models also can be deployed in hospitals with minimal finetuning, making them especially useful in emergency settings where radiologists may not always be available, thereby enhancing diagnostic capabilities and knowledge dissemination in 3D medical imaging.

Main References

[1] Luca Salvolini, Elisabetta Bichi Secchi, Leopoldo Costarelli, and Maurizio De Nicola. "Clinical applications of 2D and 3D CT imaging of the airways–a review." In: *European journal of radiology* 34 1 (2000), pp. 9–25.

[2] Cynthia H. McCollough and Prabhakar Shantha Rajiah. "Milestones in CT: Past, Present, and Future." In: *Radiology* 309 1 (2023), e230803.

[3] Yann Erick Claessens, Marie-Pierre Debray, Florence Tubach, Anne Brun, Blandine Rammaert, Pierre Hausfater, Jean-Marc Naccache, Patrick Ray, Christophe Choquet, Marie France Carette, Charles Mayaud, Catherine Leport, and Xavier Duval. "Early Chest Computed Tomography Scan to Assist Diagnosis and Guide Treatment Decision for Suspected Community-acquired Pneumonia." In: *American journal of respiratory and critical care medicine* 192 8 (2015), pp. 974–82.

[4] Andre Esteva, Katherine Chou, Serena Yeung, Nikhil Vijay Naik, Ali Madani, Ali Mottaghi, Yun Liu, Eric J. Topol, Jeff Dean, and Richard Socher. "Deep learning-enabled medical computer vision". In: *NPJ Digital Medicine* 4.1 (2021), p. 5.

[5] Chunli Qin, Demin Yao, Yonghong Shi, and Zhijian Song. "Computer-aided detection in chest radiography based on artificial intelligence: a survey". In: *BioMedical Engineering OnLine* 17 (2018).

[6] Jana Lipková, Tiffany Y. Chen, Ming Y. Lu, Richard J. Chen, Maha Shady, Mane Williams, Jingwen Wang, Zahra Noor, Richard N. Mitchell, Mehmet Turan, Gulfize Coskun, Funda Yilmaz, Derya Demir, Deniz Nart, Kayhan Başak, Nesrin Turhan, Selvinaz Ozkara, Yara Banz, Katja E. Odening, and Faisal Mahmood. "Deep learning-enabled assessment of cardiac allograft rejection from endomyocardial biopsies". In: *Nature Medicine* 28 (2022), pp. 575–582.

[7] Ibrahim Ethem Hamamci, Sezgin Er, Enis Simsar, Anjany Kumar Sekuboyina, Mustafa Gundogar, Bernd Stadlinger, Albert C. Mehl, and Bjoern H Menze. "Diffusion-Based Hierarchical Multi-Label Object Detection to Analyze Panoramic Dental X-rays". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2023, pp. 389–399.

[8] Sarthak Pati, Siddhesh P. Thakur, Megh Manoj Bhalerao, Ujjwal Baid, Caleb M. Grenko, Brandon Edwards, Micah J. Sheller, Jose Agraz, Bhakti Baheti, Vishnu Bashyam, Parth Sharma, Babak Haghighi, Aimilia Gastounioti, Mark Bergman, Bjoern H Menze, Despina Kontos, Christos Davatzikos, and Spyridon Bakas. "GaNDLF: the generally nuanced deep learning framework for scalable end-to-end clinical workflows". In: *Communications Engineering* 2.1 (2021), p. 23.

[9] Alistair Johnson, Tom Pollard, Seth Berkowitz, Nathaniel Greenbaum, Matthew Lungren, Chih-ying Deng, Roger Mark, and Steven Horng. "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports". In: *Scientific Data* 6.1 (2019), p. 317.

[10] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 3462–3471.

[11] Ha Q. Nguyen, Khanh Lam, Linh T. Le, Hieu Pham, Dat Q. Tran, Dung B. Nguyen, Dung D. Le, Chi M. Pham, Han Tong, Diep H. Dinh, Cuong Duy Do, Luu T. Doan, Cuong N. Nguyen, Binh T. Nguyen, Que V. Nguyen, A. D. Hoang, Hien Ngoc Phan, Anh Thu Nguyen, Phuong Ho, Dat Thanh Ngo, Nghia Trung Nguyen, Nhan T. Nguyen, Minh-Son Dao, and Van Vu. "VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations". In: *Scientific Data* 9.1 (2022), p. 429.

[12] Jeremy A. Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie S. Shpanskaya, Jayne Seekins, David Andrew Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, C. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and A. Ng. "Chexpert: A large chest radiograph dataset

with uncertainty labels and expert comparison". In: *AAAI Conference on Artificial Intelligence*. Vol. 33. 01. 2019, pp. 590–597.

[13] Ibrahim Ethem Hamamci, Sezgin Er, Enis Simsar, Atif Emre Yuksel, Sadullah Gultekin, Serife Damla Ozdemir, Kaiyuan Yang, Hongwei Bran Li, Sarthak Pati, Bernd Stadlinger, Albert Mehl, Mustafa Gundogar, and Bjoern Menze. "DENTEX: An abnormal tooth detection with dental enumeration and diagnosis benchmark for panoramic X-rays". In: *ArXiv* abs/2305.19112 (2023).

[14] Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Melissa Zhao, Aaron K Chow, Kenji Ikemura, Ahrong Kim, Dimitra Pouli, Ankush Patel, Amr Soliman, Chengkuan Chen, Tong Ding, Judy J. Wang, Georg Gerber, Ivy Liang, Long Phi Le, Anil V. Parwani, Luca L Weishaupt, and Faisal Mahmood. "A Multimodal Generative AI Copilot for Human Pathology." In: *Nature* (2024), pp. 1–3.

[15] Ekin Tiu, Ellie Talius, Pujan Patel, Curt P. Langlotz, Andrew Y. Ng, and Pranav Rajpurkar. "Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning". In: *Nature Biomedical Engineering* 6 (2022), pp. 1399–1406.

[16] Brian Hu, Bhavan Kumar Vasu, and Anthony J. Hoogs. "X-MIR: EXplainable Medical Image Retrieval". In: *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (2022), pp. 1544–1554.

[17] Xuxin Chen, Ximing Wang, Kecheng Zhang, Roy Zhang, Kar-Ming Fung, Theresa C. Thai, Kathleen M Moore, Robert S. Mannel, Hong Liu, Bin Zheng, and Yuchen Qiu. "Recent advances and clinical applications of deep learning in medical image analysis". In: *Medical Image Analysis* 79 (2021), p. 102444.

[18] S. Kevin Zhou, Hayit Greenspan, Christos Davatzikos, James S. Duncan, Bram van Ginneken, Anant Madabhushi, Jerry L Prince, Daniel Rueckert, and Ronald M. Summers. "A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises". In: *Proceedings of the IEEE* 109 (2020), pp. 820–838.

[19] Martin J. Willemink, Wojciech A Koszek, Cailin Hardell, Jie Wu, Dominik Fleischmann, Hugh Harvey, Les R. Folio, Ronald M. Summers, D. Rubin, and Matthew P. Lungren. "Preparing medical imaging data for machine learning". In: *Radiology* (2020), p. 192224.

[20] Rachel Lea Draelos, David Dov, Maciej A. Mazurowski, Joseph Y. Lo, Ricardo Henao, Geoffrey D. Rubin, and Lawrence Carin. "Machine-learning-based multiple abnormality prediction with large-scale chest computed tomography volumes". In: *Medical Image Analysis* 67 (2020), p. 101857.

[21] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, AnnetteKopp-Schneider, Bennett A. Landman, Geert J. S. Litjens, Bjoern H Menze, Olaf Ronneberger, Ronald M.Summers, Bram van Ginneken, Michel Bilello, Patrick Bilic, Patrick Ferdinand Christ, Richard K. G. Do, Marc J. Gollub, Stephan Heckers, Henkjan J. Huisman, William R. Jarnagin, Maureen McHugo, Sandy Napel, Jennifer S. Goli Pernicka, Kawal S. Rhode, Catalina Tobon-Gomez, Eugene Vorontsov, James Alastair Meakin, Sébastien Ourselin, Manuel Wiesenfarth, Pablo Arbeláez, Byeonguk Bae, Sihong Chen, Laura Alexandra Daza, Jian-Jun Feng, Baochun He, Fabian Isensee, Yuanfeng Ji, Fucang Jia, Namkug Kim, Ildoo Kim, Dorit Merhof, Akshay Pai, Beomhee Park, Mathias Perslev, Ramin Rezaiifar, Oliver Rippel, Ignacio Sarasua, Wei Shen, Jaemin Son, Christian Wachinger, Liansheng Wang, Yan Wang, Yingda Xia, Daguang Xu, Zhanwei Xu, Yefeng Zheng, Amber L. Simpson, Lena Maier-Hein, and Manuel Jorge Cardoso. "The Medical Segmentation Decathlon". In: *Nature Communications* 13 (2021).

[22] Patrick Bilic et al. "The Liver Tumor Segmentation Benchmark (LiTS)". In: *Medical image analysis* 84 (2019), pp. 102680–102680.

[23] Ibrahim Ethem Hamamci, Sezgin Er, Enis Simsar, Anjany Kumar Sekuboyina, Chinmay Prabhakar, Alperen Tezcan, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Furkan Almas, Irem Dougan, Muhammed Furkan Dasdelen, Hadrien Reynaud, Sarthak Pati, Christian Bluethgen, Mehmet Kemal Ozdemir, and Bjoern H Menze. "GenerateCT: Text-conditional generation of 3D chest CT volumes". In: *European Conference on Computer Vision*. 2024.

[24] Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, K. Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. "Get3d: A generative model of high quality 3d textured shapes learned from images". In: *ArXiv* abs/2209.11163 (2022).

[25] Yongqin Xian, Bernt Schiele, and Zeynep Akata. "Zero-shot learning-the good, the bad and the ugly". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 3077–3086.

[26] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. "Improved Baselines with Visual Instruction Tuning". In: *ArXiv* abs/2310.03744 (2023).

[27] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. "LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day". In: *ArXiv* abs/2306.00890 (2023).

[28] Seowoo Lee, Jiwon Youn, Hyungjin Kim, Mansu Kim, and Soon Ho Yoon. "CXR-LLAVA: a multimodal large language model for interpreting chest X-ray images". In: *ArXiv* abs/2310.18341 (2023).

[29] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. "A simple framework for contrastive learning of visual representations". In: *ArXiv*. Vol. abs/2002.05709. 2020.

[30] Martin J. Willemink and Peter B. Noël. "The evolution of image reconstruction for CT—from filtered back projection to artificial intelligence". In: *European Radiology* 29 (2018), pp. 2185–2195.

[31] Martin J. Willemink, Wojciech A Koszek, Cailin Hardell, Jie Wu, Dominik Fleischmann, Hugh Harvey, Les R. Folio, Ronald M. Summers, D. Rubin, and Matthew P. Lungren. "Preparing Medical Imaging Data for Machine Learning." In: *Radiology* (2020), p. 192224.

[32] An Yan, Julian McAuley, Xing Lu, Jiang Du, Eric Y. Chang, Amilcare Gentili, and Chun-Nan Hsu. "RadBERT: Adapting transformer-based language models to radiology". In: *Radiology. Artificial intelligence* 4 4 (2022), e210258.

[33] Shervin Minaee, E. Cambria, and Jianfeng Gao. "Deep learning-based text classification: A comprehensive review". In: *ACM Computing Surveys* 54 (2020), pp. 1–40.

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. "Learning transferable visual models from natural language supervision". In: *International Conference on Machine Learning*. 2021, pp. 8748–8763.

[35] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. "Cross-modal contrastive learning for text-to-image generation". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 833–842.

[36] Gunwoo Yong, Kahyun Jeon, Daeyoung Gil, and Ghang Lee. "Prompt engineering for zero-shot and few-shot defect detection and classification using a visual-language pretrained model". In: *Computer-Aided Civil and Infrastructure Engineering* 38 (2022), pp. 1536–1554.

[37] Alexandre Bailly, Corentin Blanc, Élie Francis, Thierry Guillotin, Fadi Jamal, Béchara Wakim, and Pascal Roy. "Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models". In: *Computer methods and programs in biomedicine* 213 (2021), p. 106504.

[38] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. "Scaling vision transformers". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 1204–1213.

[39] Mitchell Wortsman, Gabriel Ilharco, Mike Li, Jong Kim, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. "Robust fine-tuning of zero-shot models". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 7949–7961.

[40] Chengkuan Chen, Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Andrew J. Schaumberg, and Faisal Mahmood. "Fast and scalable search of whole-slide images via self-supervised deep learning". In: *Nature Biomedical Engineering* 6 (2022), pp. 1420–1434.

[41] Shaoting Zhang, Ming Yang, Timothée Cour, Kai Yu, and Dimitris N. Metaxas. "Query specific fusion for image retrieval". In: *European Conference on Computer Vision*. 2012, pp. 660–673.

[42] Abhimanyu Dubey et al. "The Llama 3 Herd of Models". In: *ArXiv* abs/2407.21783 (2024).

[43] J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. "LoRA: Low-Rank Adaptation of Large Language Models". In: *ArXiv* abs/2106.09685 (2021).

[44] Benjamin Hou, Qingqing Zhu, Tejas Sudarshan Mathai, Qiao Jin, Zhiyong Lu, and Ronald Summers. "Shadow and Light: Digitally Reconstructed Radiographs for Disease Classification". In: *ArXiv* abs/2406.03688 (2024).

[45] Michael P Hartung, Ian C Bickle, Frank Gaillard, and Jeffrey P Kanne. "How to Create a Great Radiology Report." In: *Radiographics : a review publication of the Radiological Society of North America, Inc* 40 6 (2020), pp. 1658–1670.

[46] Christopher J. Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg C. Corrado, and Dominic King. "Key challenges for delivering clinical impact with artificial intelligence". In: *BMC Medicine* 17 (2019), pp. 1–9.

[47] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel Coelho de Castro, Anton Schwaighofer, Stephanie L. Hyland, Maria Teodora Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. "Making the most of text semantics to improve biomedical vision-language processing". In: *European Conference on Computer Vision*. 2022, pp. 1–21.

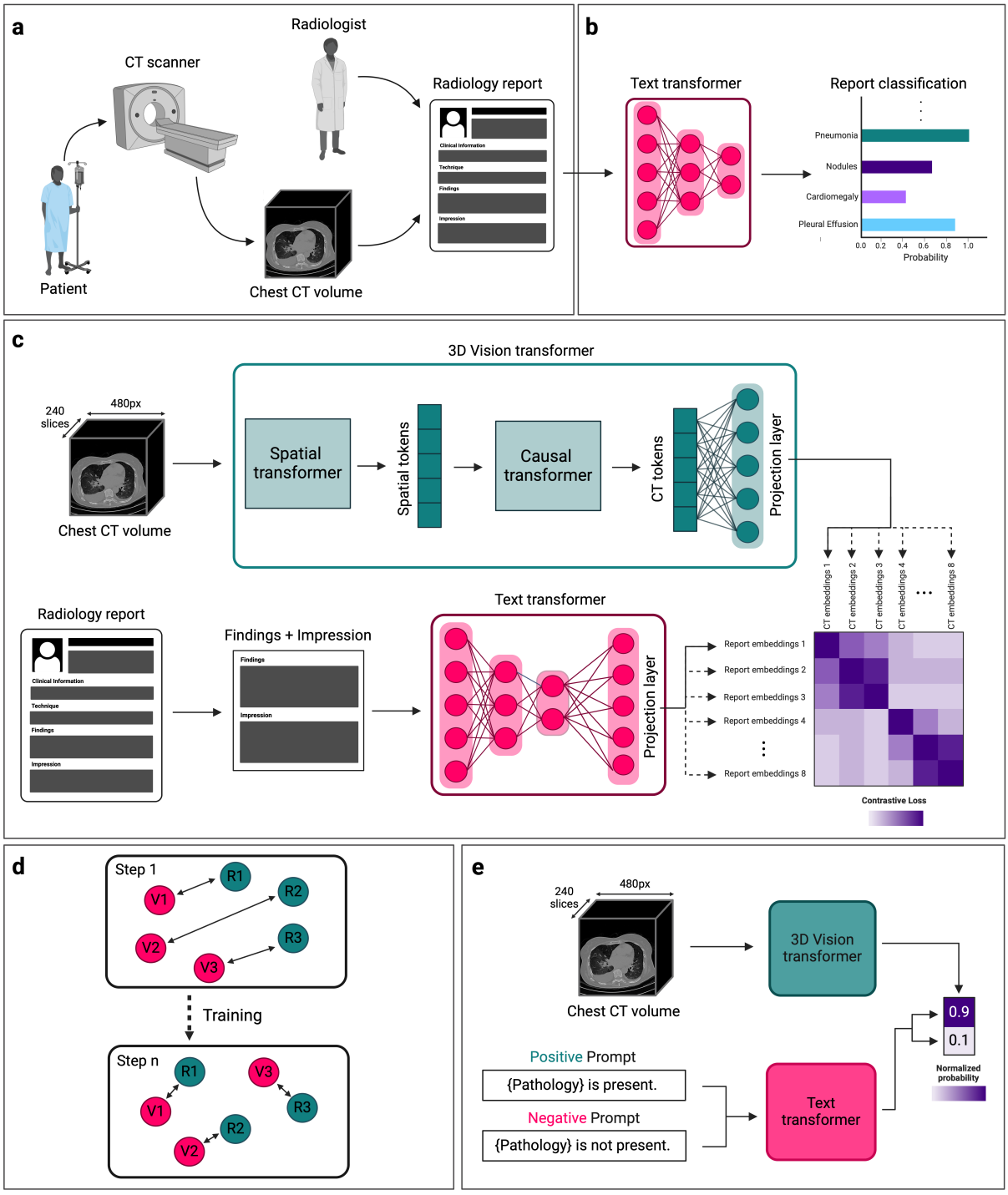


Figure 1: Overview of our dataset and CT-CLIP. a. Data acquisition protocol for the CT-RATE dataset, comprising non-contrast 3D chest CT volumes and corresponding radiology text reports written by expert radiologists. **b.** Extraction of multi-abnormality labels from radiology reports using a finetuned text encoder. **c.** Training of our CT-CLIP framework, which pairs chest CT volumes with radiology reports from the CT-RATE dataset through a contrastive learning approach. **d.** Illustration of the contrastive learning procedure, employed in the correlation matrix during training to ensure precise match recognition and mismatch differentiation. V_i shows volume embeddings and R_i shows report embeddings. **e.** Inference for zero-shot multi-abnormality detection utilizes normalized probabilities derived from softmax-applied contrastive loss for each positive and negative prompt.

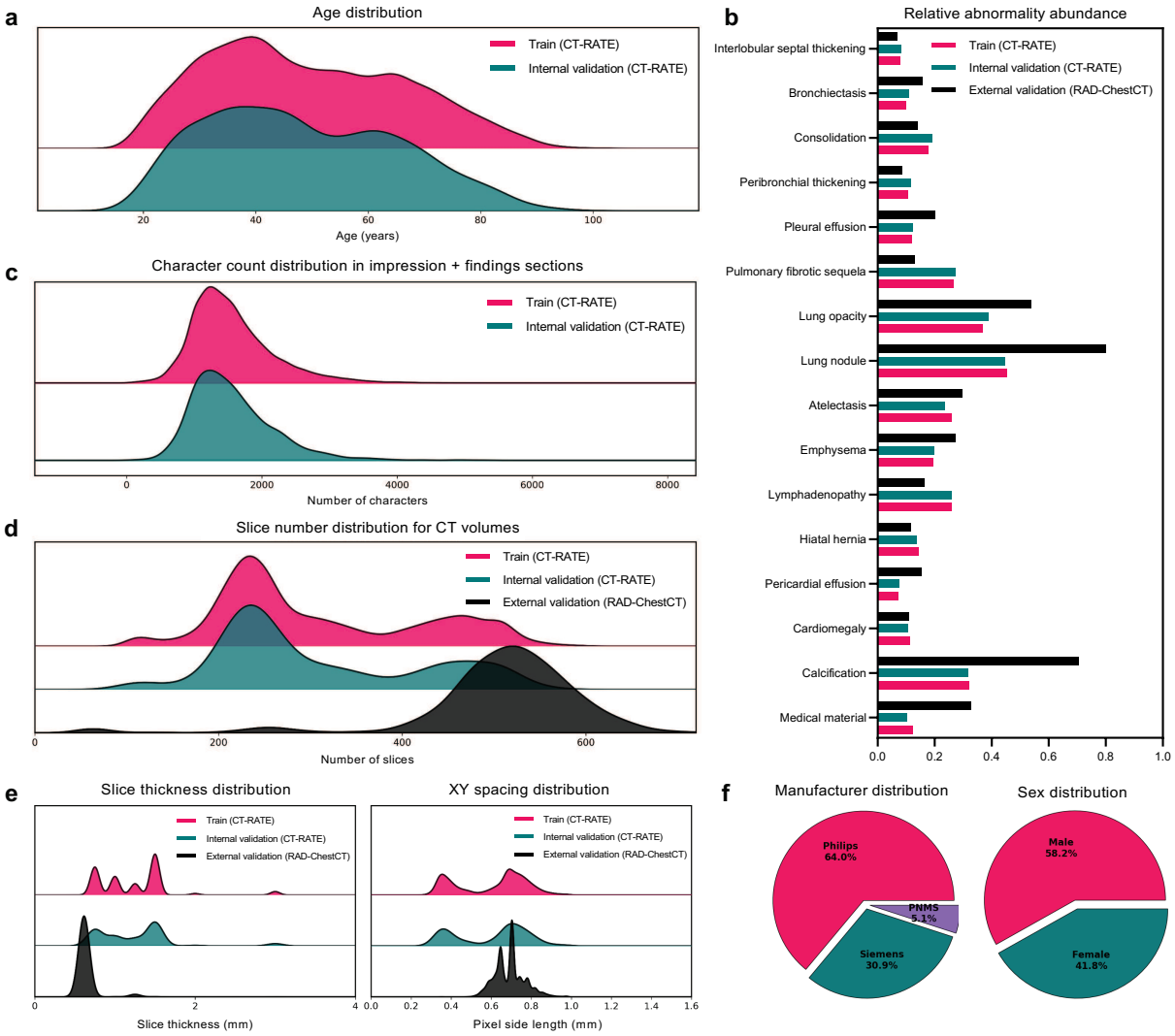


Figure 2: Comprehensive analysis of the novel CT-RATE dataset. a. CT-RATE comprises data from 21,304 unique patients, with ages ranging from 18 to 102 years. **b.** For validation and finetuning purposes, multi-abnormality labels for 18 distinct abnormalities are extracted from the corresponding radiology reports for each CT volume. **c.** The training process of the CT-CLIP model utilizes the *impression* and *findings* sections of radiology reports, which vary in length, along with the chest CT volumes. **d.** The number of slices in the chest CT volumes within CT-RATE ranges from 100 to 600. **e.** Spacings in the x-axis, y-axis, and z-axis vary across volumes; all chest CT volumes are preprocessed to ensure consistent spacing before training each model. **f.** Despite originating from a single institution, CT-RATE includes volumes acquired using scanners from three different manufacturers, ensuring variability.

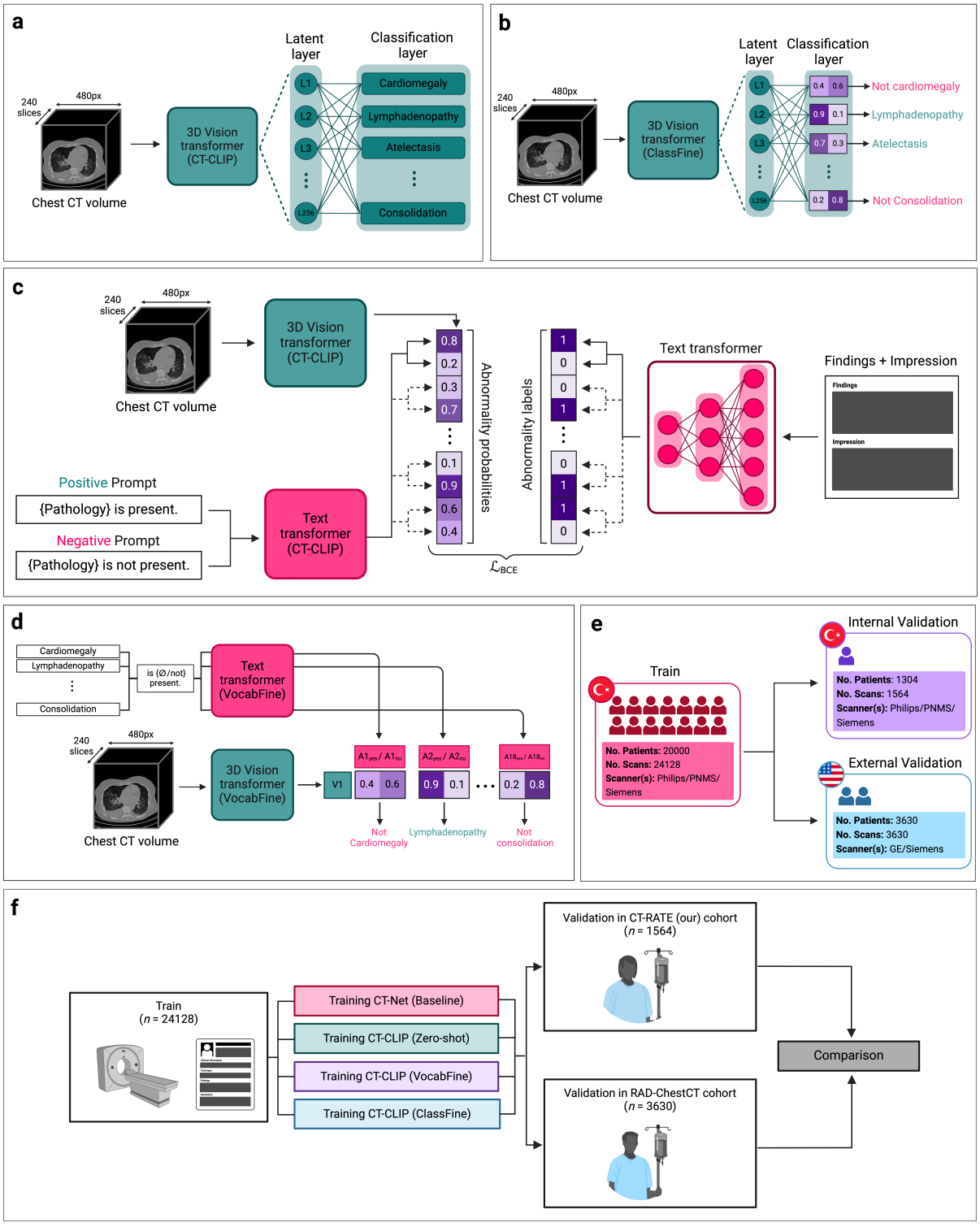


Figure 3: Finetuning CT-CLIP and the validation strategy. a. Illustration of linear probing finetuning method (ClassFine) for CT-CLIP, where a linear layer is incorporated into the vision encoder. **b.** ClassFine enables multi-abnormality classification but is limited to the classes predefined during finetuning. **c.** Illustration of CT-CLIP's open vocabulary finetuning method (VocabFine) for each abnormality. **d.** VocabFine allows for open vocabulary abnormality classification even after finetuning, although it is constrained to the prompts provided during finetuning. **e.** Validation: models are trained on the CT-RATE dataset from Türkiye and then tested on both an internal CT-RATE validation set and an external dataset from the U.S.; a detailed analysis of the CT-RATE dataset is presented in Figure 2. **f.** Comparison: a comprehensive evaluation is performed in multi-abnormality detection across two different cohorts, evaluating CT-CLIP, the two finetuned models, and a fully supervised method.

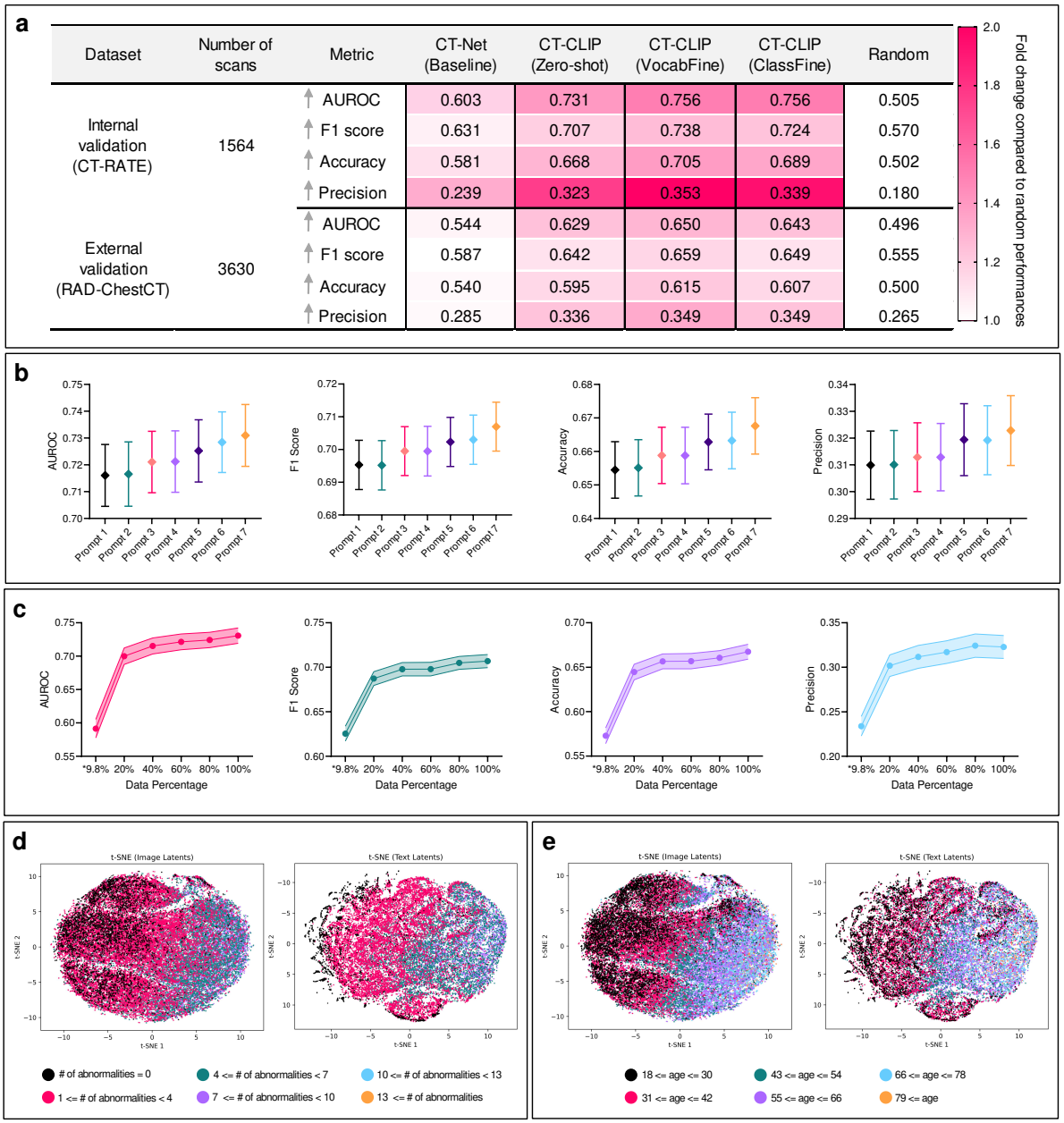


Figure 4: A detailed evaluation of CT-CLIP a. Multi-abnormality classification performance is evaluated on both internal and external validation sets using four key metrics. Our CT-CLIP outperforms the fully supervised baseline across all metrics. Performance improves with both finetuning techniques, demonstrating the robustness of the CT-CLIP's encoder in feature extraction from CT volumes. The scores are color-highlighted to emphasize the performance increase compared to random benchmarks. **b.**Different prompts are utilized for the zero-shot multi-abnormality detection task with CT-CLIP. The best-performing prompt, highlighting the importance of prompt engineering, is selected for further experiments. **c.** To explore how the performance scales with dataset size, we train CT-CLIP using different portions of CT-RATE. (9.8% of CT-RATE matches the size of the only currently available open-source chest CT dataset with multi-abnormality labels.) Contour lines denote ±1 standard deviations calculated via bootstrapping. **d.** t-SNE projections map the distribution of the number of abnormalities present in cases, utilizing the latent values from individual reports and CT volumes in CT-RATE. **e.** t-SNE projections illustrate the age distribution of patients in the dataset. The clustering shows that CT volumes and reports from older patients tend to correspond with a higher number of reported abnormalities, indicating a correlation between age and disease prevalence.

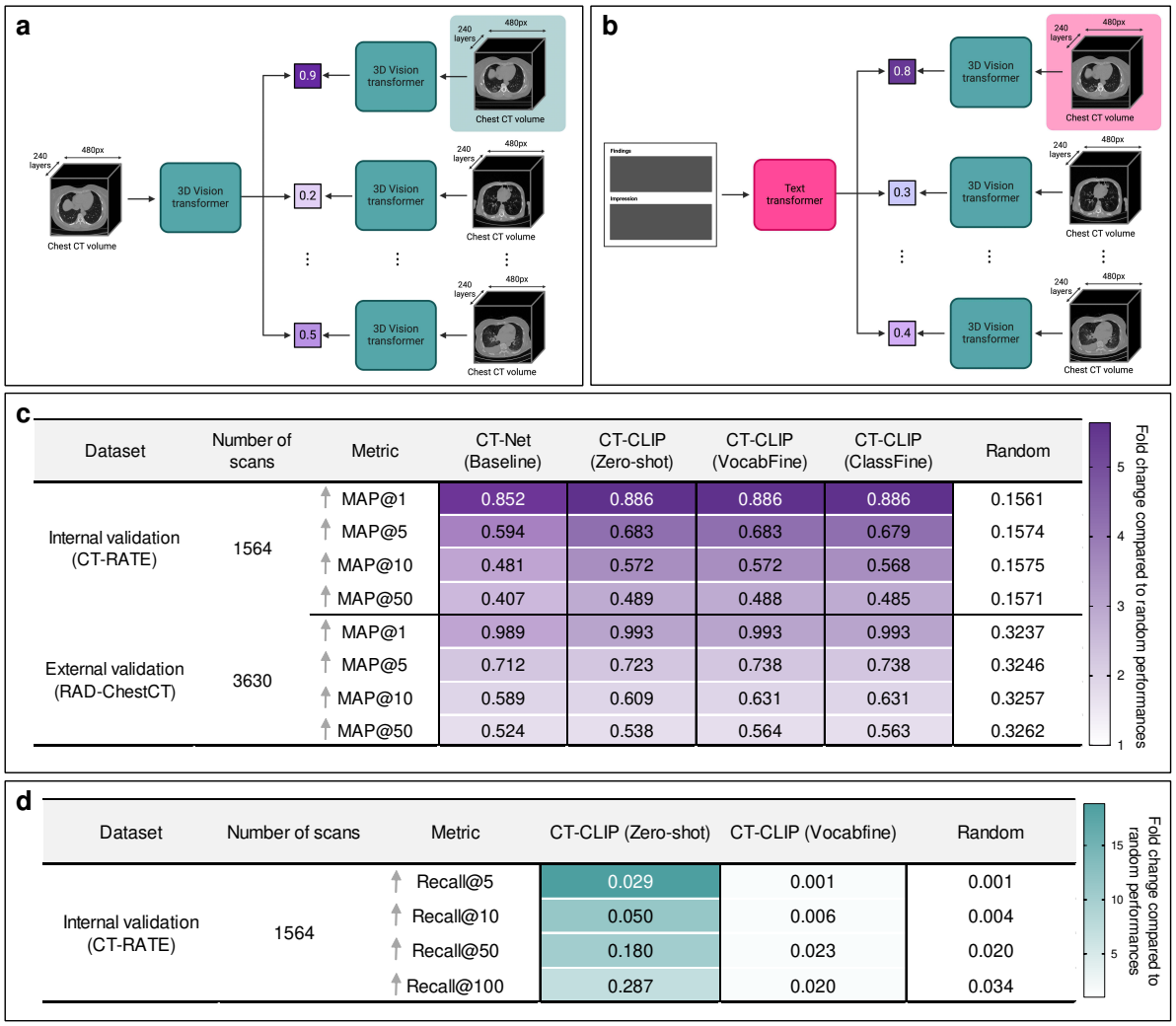


Figure 5: Utilizing CT-CLIP for retrieval of chest CT volumes a. CT-CLIP can be used for volume-to-volume retrieval utilizing the vision encoder within CT-CLIP. **b.** CT-CLIP can be used for report-to-volume retrieval utilizing both the vision and text transformers within CT-CLIP. **c.** Comprehensive evaluation of volume-to-volume retrieval performance on internal and external validation sets, with scores color-highlighted to emphasize improvements over random scores, shows similar retrieval accuracy even for the finetuned models. **d.** The evaluation of report-to-volume retrieval is limited to the internal validation set due to the absence of paired radiology reports in the external validation set. Notably, finetuning CT-CLIP for detection diminishes its retrieval performance. This reduction in performance is primarily observed in report-to-volume retrieval, indicating that the VocabFine finetuning predominantly impacts on the text transformer component of CT-CLIP.

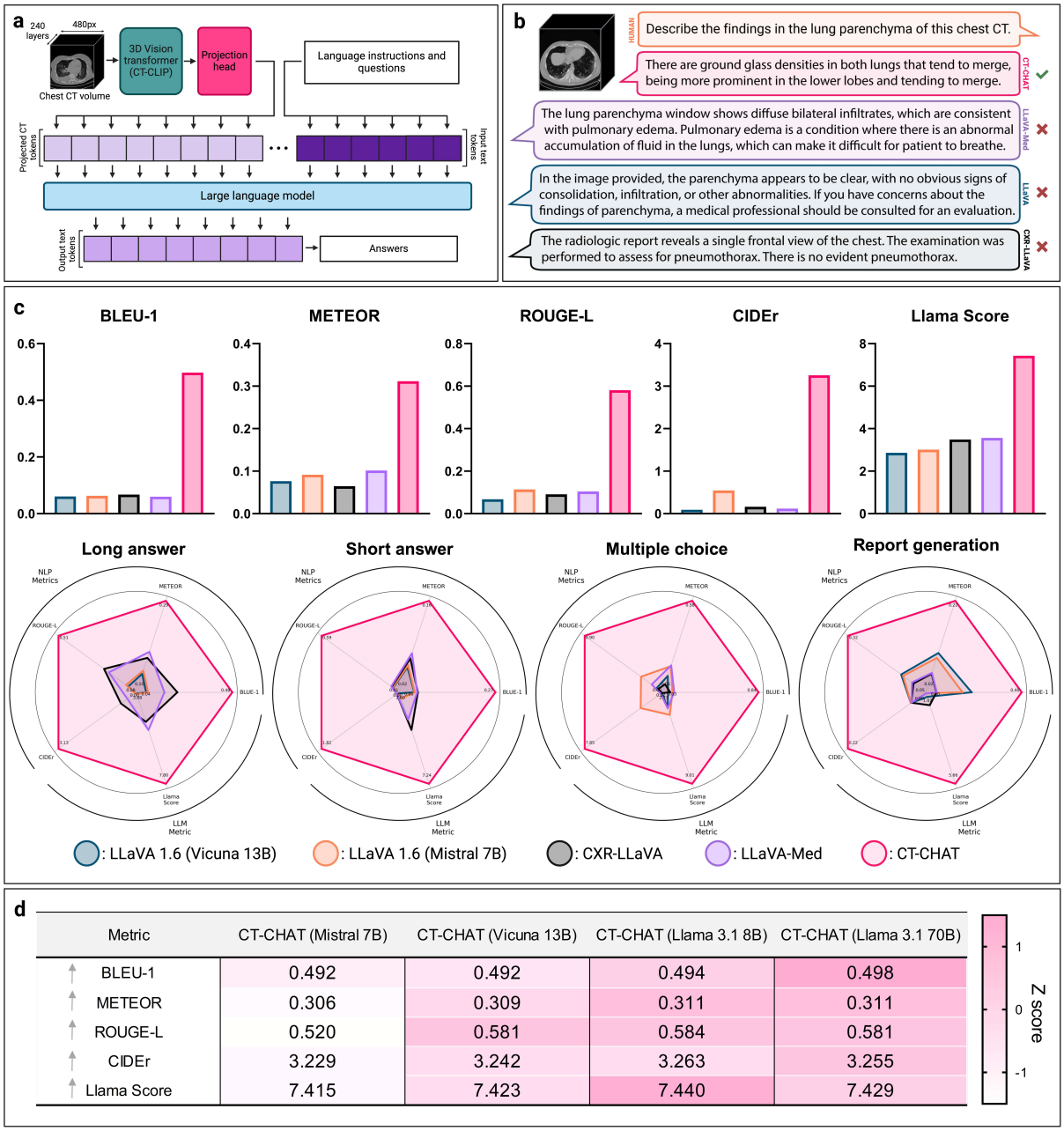


Figure 6: Overview and detailed evaluation of CT-CHAT a. The architecture of CT-CHAT, a vision-language AI assistant for 3D chest CT volumes, trained by finetuning LoRA weights on a state-of-the-art large language model using CT tokens encoded by the 3D vision encoder of CT-CLIP. **b.** An illustration showing an example CT-CHAT output, highlighting its superior performance as a vision-language assistant specifically designed for 3D chest CT volumes compared to its 2D counterparts with DRRs provided. **c.** The question-answering performance of various vision-language assistants is evaluated across five key metrics, with radar charts providing a detailed breakdown of evaluations for each of the four different visual question-answering tasks. **d.** An ablation study on the accuracy of CT-CHAT using different LLMs, shows that CT-CHAT with the Llama 3.1 model performs slightly better.

Online Methods

Ethics statement

Our study is granted with ethical approval from the Clinical Research Ethics Committee at Istanbul Medipol University (E-10840098-772.02-6841, 27/10/2023) for open-sourcing our CT-RATE dataset and associated trained models. To uphold patient privacy, we anonymize all chest CT volumes, associated metadata, and radiology text reports before conducting any computational analysis.

Creating the CT-RATE dataset

We curate a novel dataset, CT-RATE, consisting of chest CT volumes and corresponding radiology text reports from Istanbul Medipol University Mega Hospital. The CT-RATE dataset includes non-contrast chest CT volumes acquired between May 2015 and January 2023, totaling 50,188 reconstructed CT volumes from 25,692 distinct CT experiments conducted on 21,304 unique patients. We divide the cohort into two groups: 20,000 patients are allocated to training and 1,304 for validation. Figure **1a** illustrates the dataset accumulation process, and Figure **3e** shows the training and validation split.

In clinical practice, CT volumes are reconstructed with various techniques for specific diagnostic requirements. Sharper kernels, for instance, improve resolution for lung abnormalities, whereas smoother kernels are favored for assessing mediastinal issues [30]. For that reason, we incorporate every reconstructed volume from the hospital database into our dataset. This strategy facilitates a thorough evaluation of various abnormalities. The patients range in age from 18 to 102 years, with a mean of 48.8 years and a mode of 40 years. The sex distribution is 41.6% female and 58.4% male. 61.5% of the scans are taken using Philips CT scanners, 30.1% with Siemens scanners, and 8.4% with PNMS (Philips-Neusoft Medical Systems) scanners. The CT volumes are provided in various resolutions: 65.4% in 512 $\times$ 512 px, 4.2% in 768 $\times$ 768 px, and 30.4% in 1024 $\times$ 1024 px. The pixel spacing in the axial (XY) plane ranges from 0.227 to 1.416 mm, with a mean of 0.605 mm, while slice thickness ranges from 0.035 to 6 mm, with a mean of 1.231 mm. The number of slices per volume ranges from 100 to 600, averaging 304.7 with a mode of 255 slices. The radiology reports are structured into four sections: clinical information, technique, findings, and impression. The clinical information section outlines the patient's symptoms and medical history. The technique section specifies the acquisition protocol, encompassing details such as the field of view. The findings section covers anatomical and pathological observations, whereas the impression offers a diagnosis based on these findings. Figure **2** summarizes the CT-RATE dataset.

To anonymize the CT-RATE dataset, we first anonymize reports by removing all personal information pertaining to patients and medical professionals. This process involves targeted searches using regular expressions to identify and eliminate date-related information and patient-doctor identifiers. Subsequently, these anonymized reports are translated from Turkish to English using the Google Translate API, and bilingual final-year medical students review and correct all translations. They also thoroughly examine each report to remove any residual personal information and ensure coherence. Only the English versions of these reports are included in our CT-RATE dataset. For the CT volumes, we implement several quality control measures. We exclude volumes that involve contrast material or body parts other than the chest. Only the remaining data is used for developing our models and open-sourced as the CT-RATE dataset. Each chest CT volume's orientation is standardized by examining its metadata to ensure consistency. Additionally, we remove all identifying information about patients and medical professionals from the metadata. The anonymized metadata is open-sourced as part of the CT-RATE dataset. Extended Data Table **1** provides a detailed explanation of these metadata attributes.

To develop an automated label extractor model, we annotate 1,000 reports by merging findings and impression sections, as some abnormalities are not explicitly mentioned in a single section. This creates a multi-label dataset for each report, identifying 18 abnormalities. Conditions that are not referenced in either section are considered absent. Similar conditions are grouped: left and right mucoid impactions under *mucoid impaction*, and various lung opacities (lung opacity, density increase, and ground-glass opacities) under *lung opacity*. Lung and fissural nodules are categorized as *lung nodules*. The annotated reports are divided into training (80%, 800 reports) and validation (20%, 200 reports) sets. Leveraging this dataset, we finetune the RadBERT-RoBERTa-4m model [32], which was originally trained on over 4 million radiology reports. Then, we apply it to classify the remaining 24,692 reports. These classified reports serve three purposes: training the supervised baseline model, finetuning CT-CLIP, and internal evaluation of the models. Figure **1b** illustrates the automated classification process for radiology reports. Supplementary Table **2** shows the precision, recall, and F1 scores of the report classifier.

Our preprocessing approach, inspired by previous work [20], involved several key steps to standardize all chest CT volumes in the dataset. First, each volume is resized to achieve uniform spacing of 0.75

mm in the x-axis and y-axis and 1.5 mm in the z-axis. The CT volumes are then either center-cropped or padded to reach a consistent resolution of $480 \times 480 \times 240$, ensuring comprehensive inclusion of the patient's relevant area for the interpretation. Subsequently, we convert each CT volume to Hounsfield Unit (HU) values using intercept and slope values from the metadata. HU values, a measure of radiodensity, are clipped to a range of -1000 to 1000. This range represents the practical lower and upper limits of the HU scale [48]. During training, we normalize these values to a range of -1 to 1. The CT-RATE dataset contains the original raw data and all preprocessing scripts are provided in our GitHub repository.

Adding an external validation dataset to CT-RATE

For external validation, we utilize the RAD-ChestCT dataset, which includes a cohort from another country [20]. The dataset comprises 3,630 non-contrast chest CT volumes, uniformly reconstructed using a single technique. Half are conducted with Siemens CT, and the other half with GE Medical scanners. The number of slices ranges from 46 to 1277, with an average of 511.6 and a mode of 472. All CT volumes have a resolution of 512×512 pixels. The pixel spacing in the axial (XY) plane ranges from 0.189 to 0.977 mm, with an average of 0.692 mm, while the slice thickness varies from 0.125 to 5 mm, with an average of 0.706 mm. To ensure consistent evaluation, we apply identical preprocessing methods to this external dataset as used with our CT-RATE cohort. Since the RAD-ChestCT dataset lacks a *Mosaic attenuation* label, we exclude it from our external evaluation. Furthermore, while the RAD-ChestCT dataset includes a *Calcification* label, our dataset distinguishes between *Arterial wall calcification* and *Coronary artery wall calcification*. In assessing the model accuracies on the external evaluation set, we use the higher classification probabilities between these two labels for the *Calcification* prediction.

Developing the CT-CLIP model

CT-CLIP is a novel 3D adaptation of the Contrastive Language-Image Pretraining (CLIP) which is originally designed for natural image-text pairs [34] and has shown excellent performance in 2D medical data [15, 49]. It learns to identify similarities between text and images via contrastive learning (Figure **1d**), enabling efficient zero-shot classification. We extend CLIP to work with 3D CT paired with reports. It integrates a 3D encoder, previously used for generating CT [23], as the image encoder and CXR-Bert [47] as the text encoder. Similar to the 2D CLIP, our method extracts latent embeddings from input volumes and radiological reports. We then compute the contrastive loss, which measures the similarity between these embeddings. Training the model to align the 512-dimensional projection layers from text and CT volume embeddings with the identity matrix (as shown in Figure **1c**) allows identifying meaningful correlations between textual information and 3D medical images. This process enables zero-shot classification and can be further enhanced through task-specific finetuning for supervised classification.

As a 3D vision transformer, we utilize a component of CT-ViT [23], a network designed to generate low-dimensional tokens from CT volumes and reconstruct CT volumes from these tokens. CT-ViT divides CT volumes into patches measuring $20 \times 20 \times 10$ voxels in the coronal, sagittal, and axial planes, ensuring uniform spacing. These patches are processed by spatial and causal transformers to produce low-dimensional CT tokens, which are then decoded to reconstruct the CT volumes [23]. For CT-CLIP, we use only the encoder part of this network to generate CT tokens. During the training phase of CT-CLIP, the encoded CT tokens are averaged across the axial plane, then flattened and transformed into a 512-dimensional layer. This creates latent representations for the CT volumes, as shown in Figure **1c**. This approach reduces GPU memory requirements significantly through the linear transformation layer for latent representations, while still capturing information from every slice due to the axial attention mechanism in the CT-ViT. The improved memory efficiency enables the use of a larger 3D CT encoder.

For latent representations of radiology reports, we employ the CXR-Bert text encoder [47]. This encoder can process up to 512 tokens, which allows the inclusion of both impression and findings sections from reports. For reports with content of fewer than 512 tokens, we use padding with the <PAD> token, where padded tokens are masked during the attention mechanism. Each token is represented in 768 dimensions in the CXR-Bert text encoder. These tokens are then summed across the 512-token span and linearly transformed into a compact 512-dimensional layer, encapsulating the essential textual information from the reports. Figure **1c** visually represents this encoding process. Additionally, to assess the impact of including the impression and findings sections, we conduct an ablation study on impression only and findings only training in Supplementary Figure **1a**. To highlight the impact of scaling the dataset size on model performance, we conduct an ablation study. This experiment involves training the CT-CLIP model with varying sizes of the CT-RATE dataset: 9.8% (matching the size of the RAD-ChestCT dataset, the only public chest CT dataset with multi-abnormality labels [20]), 20%, 40%, 60%, 80%, and 100%. By

systematically altering the sizes of the dataset, our goal is to explore how the accuracy of the model scales with the amount of available training data. Figure **4c** illustrates the outcomes of this ablation study.

CT-CLIP for zero-shot multi-abnormality detection

In our CT-CLIP model, we implement a similar approach to CheXzero [15] for zero-shot multi-abnormality classification. This method adapts CLIP's zero-shot classification technique to handle multiple classes instead of multiple labels for the same class, which is particularly important in medical imaging where images often contain several abnormalities simultaneously. We start by extracting the model's preliminary output scores, known as logits, using both positive (e.g., "Consolidation is present.") and negative (e.g., "Consolidation is not present.") prompts. The prompting method for inference is chosen through prompt engineering experiments. We then apply softmax to these logits to convert them into a likelihood of each abnormality's presence or absence in a CT volume. Figure **1e** illustrates CT-CLIP's inference strategy for the zero-shot multi-abnormality detection task.

To optimize CT-CLIP's performance for zero-shot multi-abnormality detection, we experiment with various prompts, derived from common sentence structures in radiology reports. The positive-negative formulations for prompts 1 to 7 are as follows: "{∅/No} {abnormality}.", "Findings are {∅/not} compatible with {abnormality}.", "There is {∅/no} {abnormality}.", "{Abnormality} is {∅/not} seen.", "{∅/Not} {abnormality}.", "{Abnormality} is {∅/not} observed.", and "{Abnormality} is {∅/not} present.". We evaluate these prompts based on their classification scores (Figure **4b**) and apply the most effective one to finetune the CT-CLIP model using the VocabFine approach and during inference.

Finetuning CT-CLIP for multi-abnormality detection

To enhance the performance of the CT-CLIP model in classifying multiple abnormalities, we implement two finetuning approaches. The first approach, called VocabFine, involves finetuning the model while retaining its open vocabulary capabilities. The second approach, adapted from linear probing, is called ClassFine. This method entails training the model with an additional classification layer. We also conduct ablations to investigate the impact of freezing pretrained layers within CT-CLIP during finetuning (Supplementary Figure **1b**). Specifically, in addition to training every layer, we finetune only the projection layers of the image and text encoders in the case of VocabFine. For ClassFine, alongside training every layer, we train the model only for the newly added classification layer.

The first approach, VocabFine, is a modification of WISE-FT, a technique for finetuning foundation models [39]. In WISE-FT, the zero-shot model generates logits using classification labels as input text prompts. The model is then finetuned using binary cross-entropy (BCE) [50] to compare these logits with the actual ground truth labels. This process allows the model to learn specific classification labels while maintaining its ability to function with an open vocabulary [39]. Our VocabFine approach adapts this method for the multi-label classification. We generate both positive and negative prompts for each abnormality and then concatenate their respective logits into a unified array of 36 elements (two logits for each of the 18 abnormalities). The ground truth labels are represented as binary arrays (1s and 0s), where 1 indicates a positive logit if the ground truth is $True$ and vice versa for $False$. For example, if a volume's label is $True$ for a specific abnormality, it corresponds to $[1, 0]$ in the ground truth array; $False$ would be represented as $[0, 1]$. The training aligns these ground truth values with the computed logits. To optimize computational efficiency, we divide the 36-element array into smaller segments of 12 elements for each gradient run. Figure **3c** depicts the finetuning process in VocabFine. The inference strategies of both CT-CLIP and the finetuned model with Vocabfine are the same (Figure **3d**).

Our second finetuning strategy, ClassFine, is achieved by integrating a linear layer into the visual encoder network, a technique inspired by [51]. The architecture of the ClassFine approach is demonstrated in Figure **3a**. This modified model undergoes training using ground truth labels. The finetuned model with ClassFine is then evaluated using both internal and external evaluation sets. While this approach offers the potential for improved classification accuracy, it also makes the foundation model more specific for the trained classification task. Since ClassFine is based on fully supervised finetuning, its capabilities are primarily limited to the classes defined during this finetuning phase. Figure **3b** visualizes the inference process of ClassFine, highlighting its specific application to the predetermined classes.

Benchmarking the CT-CLIP models for multi-abnormality detection

In assessing our CT-CLIP and finetuned models, we employ the CT-Net model as a fully supervised baseline [20]. Renowned for its performance in a similar CT dataset [20], CT-Net provides a robust

comparison point. We train the CT-Net model on our CT-RATE dataset, adapting the chest CT volumes to fit the model's input requirements by resizing them to $420 \times 420 \times 402$. Consistent with our models, we focus on the abnormalities present in the CT-RATE dataset (Figure **3d**), instead of the 98 and 9 abnormalities explored in the original CT-Net study. For each abnormality, we determine classification thresholds by identifying the left-upper point on the Receiver Operating Characteristic (ROC) curve. This analysis is conducted for zero-shot, finetuned, and fully supervised classification models. Additionally, we compute the Area Under the Receiver Operating Characteristic (AUROC) values for each abnormality.

To provide a comprehensive evaluation, we also calculate accuracy, precision, and weighted F1 scores for each identified pathology. These additional metrics offer more insights into the performance of the models. To further validate the latent spaces generated by CT-CLIP for radiology reports and volumes, we conduct a t-SNE (t-distributed stochastic neighbor embedding) analysis [52]. This analysis enables us to map the positions of abnormalities visually, the distribution of abnormality groups, and other relevant attributes within the t-SNE embeddings (Figures **4d**, **4e** and Extended Data Figures **1**, **2**).

For the statistical analysis of our model evaluations, we use the non-parametric bootstrap method to examine the distribution of predictions on internal and external validation sets. This involves 1,000 iterations of random sampling of the volumes with replacement, each using sample sets of size n, matching the number of evaluated volumes. We calculate the average metric values for each sample and quantify dispersion by computing the standard deviation of these bootstrap values. For statistical significance, we employ a two-sided paired permutation test with 10,000 permutations to assess differences in performance between the two models across AUROC, weighted F1 score, accuracy, and precision metrics, following previous work [53]. Detailed comparisons are provided in Supplementary Table **5**.

Benchmarking the CT-CLIP models for CT volume retrieval

Image retrieval involves identifying the most relevant image from a pool of candidates based on either textual [54] or visual [55] input. With CT-CLIP, trained to associate CT volumes with text reports, volume retrieval becomes possible. This is achieved by computing the cosine similarity among all pairs of volumes and reports, as well as among volume pairs, within a unified embedding space. This embedding space is generated by projecting CT tokens derived from the volumes through the projection layer, Figure **1c**. To conduct volume-to-volume retrieval, we calculate the cosine similarity between the latent embeddings of a specific chest CT volume in the validation set and all other abnormal chest CT volumes within the same set. We then rank the volumes based on their cosine similarity scores to identify the most relevant matches for a given volume. This volume-to-volume retrieval process is shown in Figure **5a**.

To evaluate this retrieval task's performance, we adapt the Mean Average Precision at K (MAP@K) metric, typically used in image-to-image retrieval tasks [40]. In this metric, an image is considered relevant if it matches the classes of the target input. Precision is then calculated as the proportion of relevant images within the top K retrieved images. Relevance is determined using the intersection over the union of the abnormality labels marked as $True$ in the ground truth with those in the retrieved chest CT volume. The Average Precision at K (AP@K) is defined as follows:

$$\text{AP@K} = \frac{1}{K} \sum_{i=1}^{K} P_i \times R_i$$

Here, P_i is the precision at the i-th position, and R_i is the relevance at the i-th position. The final class retrieval accuracy is given by:

$$\text{MAP@K} = \frac{1}{n} \text{AP@K}$$

Where n is the total number of samples. It is important to note that as K increases, MAP@K may decrease due to the inclusion of less relevant images. We utilize the vision encoders of the supervised baseline, CT-CLIP, and the finetuned models for volume-to-volume retrieval and assess their performance across both internal and external validation sets using MAP@1, MAP@5, MAP@10, and MAP@50 metrics. The Recall@K metric is not applied in image-to-image retrieval due to the inherent limitation that the cosine similarity for identical images would inevitably be the highest, resulting in a consistent Recall@K value of 1. However, this issue does not arise in text-to-image retrieval since the query is based on text embeddings rather than image embeddings.

For report-to-volume retrieval, we utilize the pretrained text transformer and 3D vision transformer to generate vector embeddings for both radiology reports and CT volumes, which are transformed from the token space to the embedding space with the CT-CLIP's trained projection layers (Figure **1c**). Our goal is to find the CT volume that best matched each query report. To achieve this, we compute the cosine similarity between the latent representations of each report and all the CT volumes. Subsequently,

we rank the CT volumes based on their cosine similarity scores to identify and return the most relevant matches. The process of report-to-volume retrieval is depicted in Figure **5b**.

For the evaluation of this task, we employ the Recall@K metric, where K represents the number of returned images. This method serves as a standard evaluation tool in text-to-image retrieval tasks [41]. We assess four Recall@K metrics: Recall@5, Recall@10, Recall@50, and Recall@100. These metrics measure the frequency of the target image appearing among the top 5, 10, 50, and 100 retrieved images, respectively. It is worth noting that the external validation set lacks text reports. Therefore, we restrict the report-to-volume retrieval evaluation to the internal validation set. We assess both CT-CLIP and the finetuned model with VocabFine for this task. However, since the supervised model and ClassFine do not generate latent embeddings for reports, they are not utilized in the report-to-volume retrieval.

Creating the dataset for CT-CHAT

To train CT-CHAT, we curate a comprehensive visual question-answering (VQA) dataset derived from free-form reports and chest CT volumes, using the same train-validation split as CT-RATE. The dataset is composed of four main components, with the following details for the training set.

The first component is the long-answer questions, which are further divided into four subcategories: conversation-based questions, description questions, free-response questions, and text-only questions. The conversation, description, and free-response questions are generated from radiology reports, while the text-only questions are created using 48,343 articles related to chest CT, retrieved from the PubMed Central-OA (www.ncbi.nlm.nih.gov/pmc/tools/openftlist). In total, the long-answer component of the training set includes 281,689 conversations, with a total of 518,053 conversation questions, 138,104 description questions, 138,061 free-response questions, and 250,760 text-only questions. All questions are generated using the Llama 3.1 8B large language model [42], following the methodology described in [14]. The scripts for generating the question-answer pairs are available in the GitHub repository, and the prompts are provided in Supplementary Table **7**. The second component consists of multiple-choice questions, also generated by Llama 3.1 8B [42], following the same approach described in [14]. This component of the training set includes a total of 46,911 multiple-choice conversations, comprising 138,608 questions in total. The third component includes short-answer questions derived from radiology reports in the CT-RATE dataset using a rule-based approach, as described in [56]. These questions cover categories such as location, presence, abnormality, and disorder. A total of 1,417,828 short-answer questions are sampled to create 318,710 short-answer conversations in the training set. The fourth component involves radiology report generation questions, totaling 47,149 conversations in the training set, where each conversation contains a single report generation question along with the corresponding generated report. The report generation questions are randomly selected from a predefined set of questions (Supplementary Table **6**).

The validation set is curated similarly to the training set, using the free-form reports from the CT-RATE validation set. The validation dataset includes 33,326 conversation questions, 8,886 description questions, 8,888 free-response questions, 12,013 short-answer questions, 8,915 multiple-choice questions, and 3,039 report-generation questions. Extended Data Figure **7** shows example entries from the described dataset. Since the VQA dataset is generated by a large language model, we carefully select only accurately generated and properly formatted data—that is, data that can be successfully parsed as JSON—for training our models. Additionally, we limit the initial number of generated questions to maintain a reasonable training duration. Our VQA dataset is made publicly available alongside CT-RATE.

Developing the CT-CHAT model

Adapted from the LLaVa framework [57], CT-CHAT integrates the frozen CT-CLIP vision encoder, a multimodal projector module, and a large language model (LLM) to enhance multimodal processing capabilities. The visual encoder compresses 3D chest CT volumes into a low-dimensional feature space. The multimodal projector module then transforms these visual features into the LLM's embedding space, aligning it with text tokens. The LLM processes these visual-language tokens as input to generate contextual responses. Figure **6a** summarizes the architecture of the CT-CHAT model.

In our implementation, we evaluate and compare four different LLMs. For the primary configuration of CT-CHAT, we use the 70B Meta Llama 3.1 [42], one of the state-of-the-art open-source LLMs. This model consists of 80 transformer layers, 64 attention heads, an embedding dimension of 8,192, an FFN dimension of 28,672, and supports a context size of up to 128k tokens. For a smaller model, we utilize the 8B Meta Llama 3.1 [42], which includes 32 transformer layers, 32 attention heads, an embedding dimension of 4,096, an FFN dimension of 14,336, and a context size of up to 128k tokens. Additionally, for comparison, we implement two LLMs used in the official LLaVA 1.6 [26]: Vicuna 13B and Mistral

7B. Figure **6d** shows the scores for each of these LLMs. To dynamically adjust the model's output mode based on the specific task, we add four special tokens to the tokenizers of each model: `<long_answer>`, `<short_answer>`, `<report_generation>`, and `<multiple_choice>`. These tokens are selected for different tasks during training and inference. The multimodal projector, based on the approach described in [14], consists of an attention-pooling layer followed by a 2-layer multi-layer perceptron (MLP). The attention-pooling mechanism utilizes 256 learned latent queries in conjunction with multi-headed cross-attention to compress the final layer's features from the encoder backbone into a fixed-length sequence of image tokens. This approach optimizes both training and inference, ensuring that the sequence length remains within the LLM's context window, and making the model focus more on the important features with the attention mechanism. The MLP that follows is adapted from LLava 1.6 [26] and includes a single hidden layer activated by GeLU, transforming the image tokens to match the LLM's embedding dimension.

Training CT-CHAT involves two steps. First, in the medical concept alignment step, the LLM is frozen, and the multimodal projector is trained with the input "Please provide the radiology report for the following chest CT volume `<image>`" and the output report. This step improves subsequent finetuning performance and helps the multimodal projector to be aligned in biomedical concepts [27]. We conduct an ablation study of this step to show its effect in Supplementary Table **10b**. In the second step, medical instruction tuning, the model is trained with the CT-VQA dataset that includes long answer, short answer, multiple choice, and report generation question-answer pairs. During this step, the multimodal projector, along with Low-Rank Adaptation (LORA) weights [43] for the LLM with rank 128 and alpha 256, is finetuned. `<image>` tokens are added to indicate the embedded image input to the model, except for text-only questions and answers. The four special tokens are added to the respective question-answer pairs in the dataset, enabling the model to learn different output modes. These tokens are later used during inference to select different modes of the model. The hyperparameters for each model for pretraining and finetuning are provided in Supplementary Table **4a**, **4b**.

Benchmarking the CT-CHAT model

Given the lack of pretrained VQA models for 3D imaging, we conduct a comparative analysis of CT-CHAT against four 2D state-of-the-art open-source VQA model: LLaVA 1.6 (Mistral 7B) [26], LLaVA 1.6 (Vicuna 13B) [26], LLaVA-Med [27], and CXR-LLaVA [28], using benchmarking framework established in [14]. To address the challenge of representing 3D information using their 2D encoders and to ensure a fair comparison, we utilize Digitally Reconstructed Radiographs (DRRs) derived from CT-RATE, as suggested by [44]. This approach is chosen over the selection of random or central CT slices, which can not adequately capture all relevant pathologies. The use of DRRs is justified, as [44] demonstrates that models trained with DRRs show comparable results to those trained with 3D chest CTs.

To evaluate the baseline models and CT-CHAT models with different LLMs, we employ four NLP metrics: BLEU-1 [58], METEOR [59], ROUGE-L [60], and CIDEr [61], along with the Llama score, which measures the clinical accuracy of the generated responses compared to the ground truth, using the Llama 3.1 70B model [42]. A similar LLM-based clinical accuracy measurement approach is used in Llava-Med with the GPT-4 API [27]. Supplementary Table **8** presents the system prompt for the Llama 3.1 model, which is used to score the generated answers. We calculate scores task specifically: long answer, short answer, multiple choice, and report generation. Figure **6c** illustrates mean and task-specific scores for each model. To visualize the clinical accuracy of the models, we categorize the Llama scores for every question in each task into three ranges: 0-3 for low clinical accuracy, 4-6 for average clinical accuracy, and 7-10 for high clinical accuracy, and create Sankey plots within these ranges in Extended Data Figure **6**.

To evaluate CT-CHAT's ability to retrieve information from CT, we conduct an ablation study for the image provision, as shown in Supplementary Table **10a**. Since the model is expected to hallucinate pathologies when the input volume is not provided, we implement a guardrail for prevention with the system prompt in Supplementary Figure **9**. Extended Data Figure **9** provides an example of the guardrail.

Computational hardware and software

Experiments for CT-CLIP are conducted using Python 3.11.5 with PyTorch 2.0.1, CUDA 11.7, and other libraries like SciPy 1.11.1, Torchvision 0.15.2, Scikit-learn 1.3.0, Pandas 2.0.3, NumPy 1.24.4, Transformers 4.30.1, and Accelerate 0.21.1. For CT-CHAT, Python 3.12.4 is used with PyTorch v2.4.0, CUDA v12.4, SciPy v1.14.0, Torchvision v0.19.0, Scikit-learn v1.2.2, Pandas v2.2.2, and NumPy v1.26.4. We use four 80GB A100 GPUs, 500GB RAM, and 24 CPUs, mainly for CT-CLIP training with Distributed Data-Parallel and Fully Sharded Data Parallelism. Finetuning of CT-CLIP and other experiments is performed on a single A100 GPU. CT-CHAT with the Llama 3.1 70B model is trained on four A100

GPUs using DeepSpeed ZeRO-3, while other CT-CHAT models use two GPUs. Hyperparameters are in Supplementary Tables **3**, **4**. All software details are available on our respective GitHub repositories.

Data availability

Our CT-RATE dataset, comprising paired chest CT volumes and radiology text reports, is publicly accessible via https://huggingface.co/datasets/ibrahimhamamci/CT-RATE. The external validation set, RAD-ChestCT, is also openly accessible through https://zenodo.org/records/6406114.

Code availability

Our trained models and codebase are publicly available at https://github.com/ibrahimethemhamamci/CT-CLIP and https://github.com/ibrahimethemhamamci/CT-CHAT for further research.

Author contributions

I.E.H., S.E., and B.M. designed the study. I.E.H. and S.E. were responsible for data collection, data analysis, model construction, model validation, figure preparation, and manuscript writing. F.A., A.G.S., S.N.E., and I.D. managed data anonymization and annotation. M.F.D. contributed to model construction. O.F.D contributed to the graphical user interface and manuscript writing. B.W., T.A., E.S., M.S., E.B.E., A.A., A.S., B.L., C.B., and M.K.O. provided knowledge support. B.M. supervised the all study. All authors reviewed the manuscript and approved the final version.

Acknowledgments

We extend our gratitude to the Helmut Horten Foundation for their invaluable support to our research. Additionally, we would like to express our sincere appreciation to Istanbul Medipol University Mega Hospital for their support and provision of data. Figures 1, 3, 5, and 6 are created with *BioRender.com*.

Methods References

[14] Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Melissa Zhao, Aaron K Chow, Kenji Ikemura, Ahrong Kim, Dimitra Pouli, Ankush Patel, Amr Soliman, Chengkuan Chen, Tong Ding, Judy J. Wang, Georg Gerber, Ivy Liang, Long Phi Le, Anil V. Parwani, Luca L Weishaupt, and Faisal Mahmood. "A Multimodal Generative AI Copilot for Human Pathology." In: *Nature* (2024), pp. 1–3.

[48] Ella A Kazerooni and Barry H Gross. *Cardiopulmonary imaging*. Vol. 4. Lippincott Williams & Wilkins, 2004.

[49] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J. Montine, and James Y Zou. "A visual–language foundation model for pathology image analysis using medical Twitter". In: *Nature Medicine* 29 (2023), pp. 2307–2316.

[50] Zhilu Zhang and Mert Rory Sabuncu. "Generalized cross entropy loss for training deep neural networks with noisy labels". In: *ArXiv* abs/1805.07836 (2018).

[51] Guillaume Alain and Yoshua Bengio. "Understanding intermediate layers using linear classifier probes". In: *ArXiv* abs/1610.01644 (2016).

[52] Laurens van der Maaten and Geoffrey E. Hinton. "Visualizing data using t-SNE". In: *Journal of Machine Learning Research* 9 (2008), pp. 2579–2605.

[53] Richard J Chen, Tong Ding, Ming Y Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L Weishaupt, Judy J. Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. "Towards a general-purpose foundation model for computational pathology." In: *Nature Medicine* 30.3 (2024), pp. 850–862.

[54] Mohammed Lamine Kherfi, Djemel Ziou, and Alan Bernardi. "Image retrieval from the world wide web: Issues, techniques, and systems". In: *ACM Computing Surveys* 36 (2004), pp. 35–67.

[55] Narayan Hegde, Jason D. Hipp, Yun Liu, Michael R. Emmert-Buck, Emily Reif, Daniel Smilkov, Michael Terry, Carrie J. Cai, Mahul B. Amin, Craig H. Mermel, Phil Q. Nelson, Lily H. Peng, Greg S Corrado, and Martin C. Stumpe. "Similar image search for histopathology: SMILY". In: *NPJ Digital Medicine* 2.1 (2019), p. 56.

[56] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Jiayu Lei, Ya Zhang, Yanfeng Wang, and Weidi Xie. "RadGenome-Chest CT: A Grounded Vision-Language Dataset for Chest CT Analysis". In: *ArXiv* abs/2404.16754 (2024).

[57] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. "Visual Instruction Tuning". In: *ArXiv* abs/2304.08485 (2023).

[58] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. "Bleu: a Method for Automatic Evaluation of Machine Translation". In: *Annual Meeting of the Association for Computational Linguistics*. 2002, pp. 311–318.

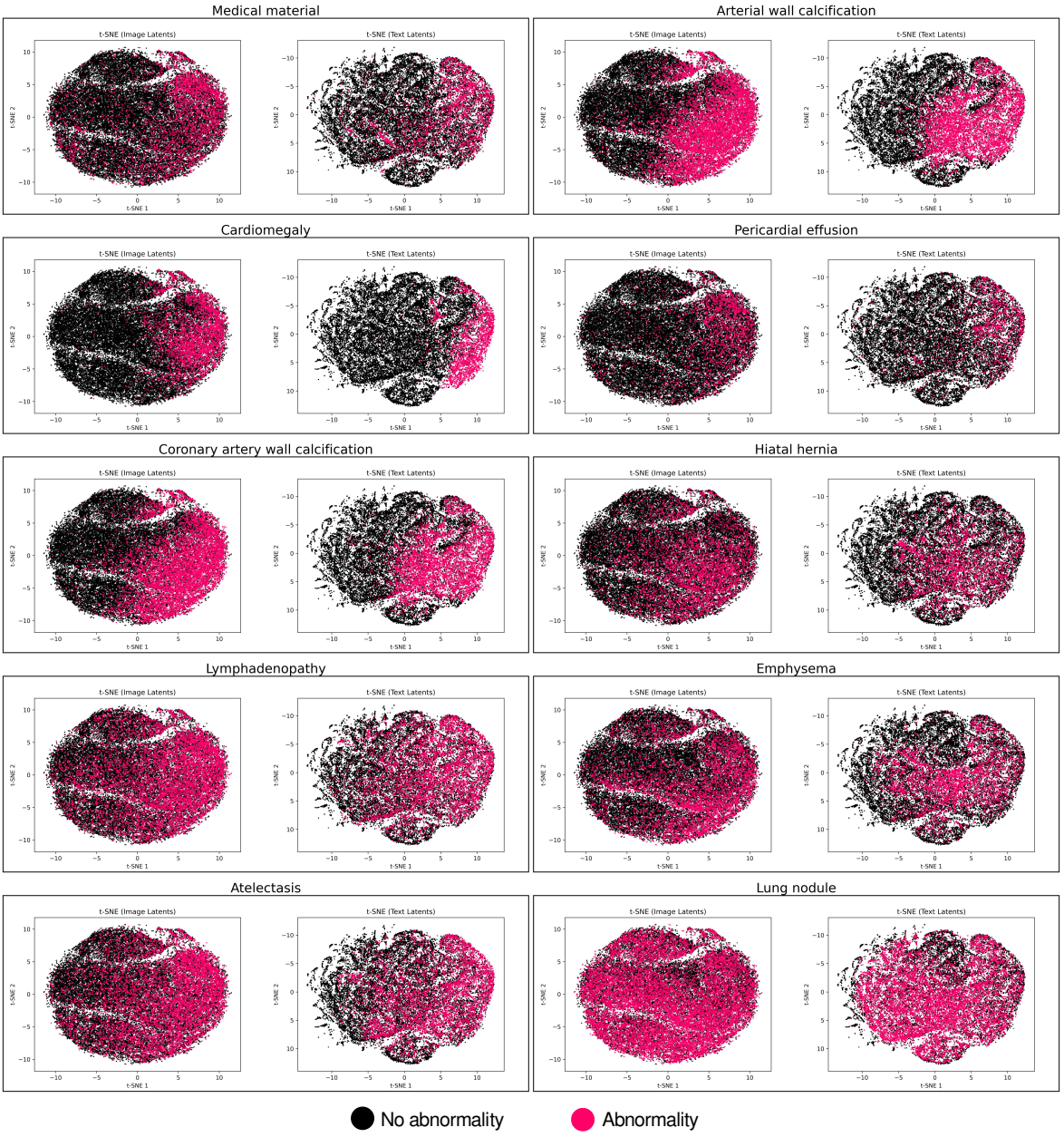
[59] Satanjeev Banerjee and Alon Lavie. "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments". In: *IEEvaluation@ACL*. 2005, pp. 65–72.

[60] Chin-Yew Lin. "ROUGE: A Package for Automatic Evaluation of Summaries". In: *Annual Meeting of the Association for Computational Linguistics*. 2004, pp. 74–81.

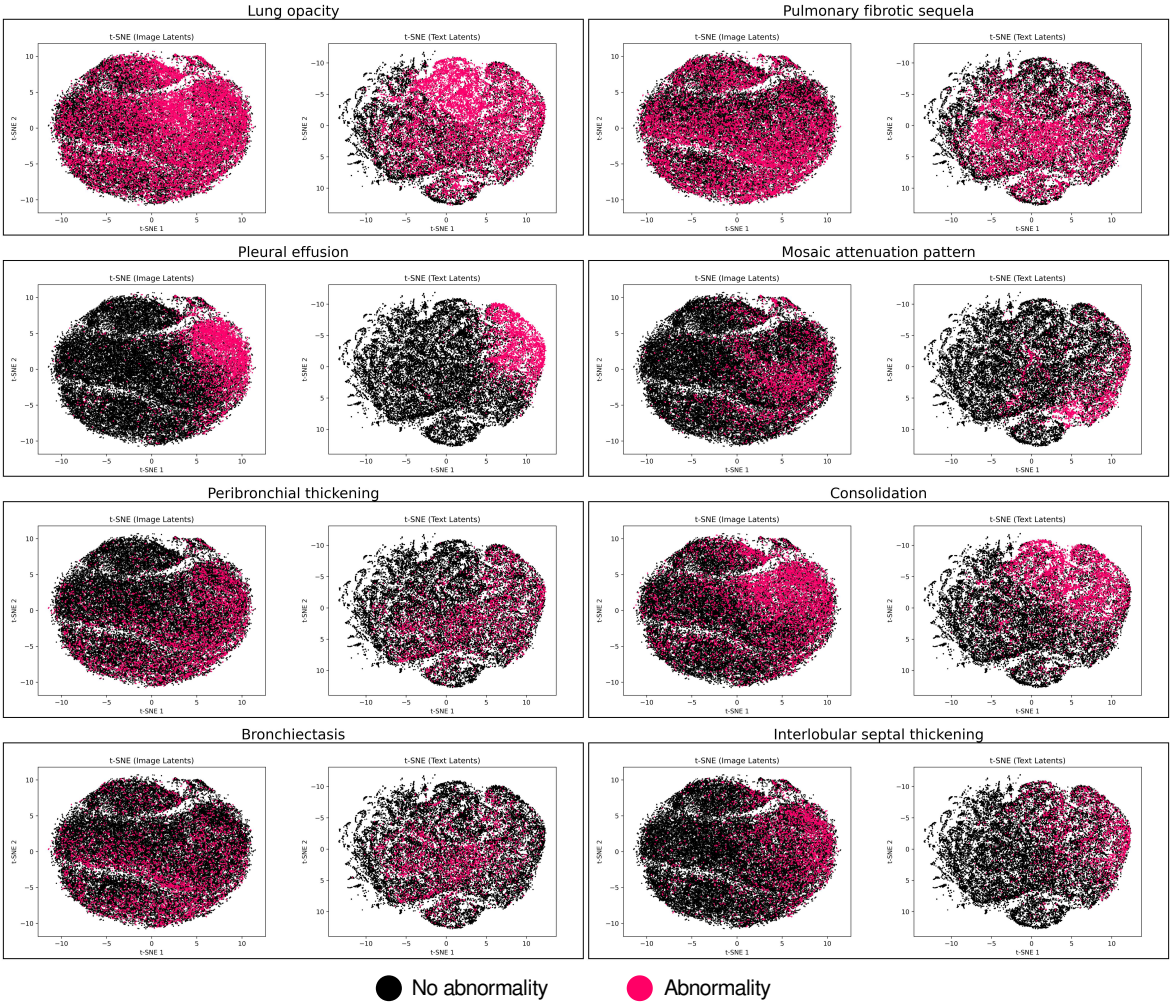
[61] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. "CIDEr: Consensus-based image description evaluation". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2014), pp. 4566–4575.

Attribute	Explanation
VolumeName	The file name.
Manufacturer	Manufacturer of the equipment that produced the Composite Instances.
SeriesDescription	Description of the Series.
ManufacturerModelName	Manufacturer's model name of the equipment that is to be used for beam delivery.
PatientSex	Sex of the named Patient.
PatientAge	Age of the Patient.
ReconstructionDiameter	Diameter, in mm, of the region from within which the data was used in creating the reconstruction of the image. Data may exist outside this region and portions of the patient may exist outside this region.
DistanceSourceToDetector	Distance in mm from source to detector center.
DistanceSourceToPatient	Distance in mm from source to the table, support or bucky side that is closest to the Imaging Subject, as measured along the central ray of the X-Ray beam.
GantryDetectorTilt	Nominal angle of tilt in degrees of the scanning gantry. Not intended for mathematical computations.
TableHeight	The height of the patient table in mm. The range and values of this element are determined by the manufacturer.
RotationDirection	Direction of rotation of the source when relevant, about nearest principal axis of equipment.
ExposureTime	Duration of X-Ray exposure in msec.
XRayTubeCurrent	X-Ray Tube Current in mA.
Exposure	The exposure expressed in mAs, for example calculated from Exposure Time and X-Ray Tube Current.
FilterType	Type of filter(s) inserted into the X-Ray beam (e.g., wedges).
GeneratorPower	Power in kW to the X-Ray generator.
FocalSpots	Nominal focal spot size in mm used to acquire this image.
ConvolutionKernel	A label describing the convolution kernel or algorithm used to reconstruct the data. A single value shall be present.
PatientPosition	Patient position descriptor relative to the equipment.
RevolutionTime	The time in seconds of a complete revolution of the source around the gantry orbit.
SingleCollimationWidth	Adjacent physical detector rows may have been combined to form a single effective acquisition row.
TotalCollimationWidth	This will be equal the number of effective detector rows multiplied by single collimation width.
TableSpeed	The distance in mm that the table moves in one second during the gathering of data that resulted in this image.
TableFeedPerRotation	Motion of the table (in mm) during a complete revolution of the source around the gantry orbit.
SpiralPitchFactor	Ratio of the Table Feed per Rotation to the Total Collimation Width.
DataCollectionCenterPatient	The x, y, and z coordinates (in the Patient-Based Coordinate System) in mm of the center of the region in which data were collected.
ReconstructionTargetCenterPatient	If the reconstructed image is not magnified or panned the value corresponds with the Data Collection Center Attribute.
ExposureModulationType	A multi-valued label describing the type of current modulation used for the purpose of limiting the dose.
CTDIvol	It describes the average dose for this image for the selected CT conditions of operation.
ImagePositionPatient	Image Position specifies the x, y, and z coordinates of the upper left hand corner of the image; it is the center of the first voxel transmitted.
ImageOrientationPatient	Image Orientation specifies the direction cosines of the first row and the first column with respect to the patient.
SliceLocation	Slice Location is defined as the relative position of the image plane expressed in mm. This information is relative to an unspecified implementation specific reference point.
SamplesPerPixel	Samples per Pixel is the number of separate planes in this image.
PhotometricInterpretation	The value of Photometric Interpretation specifies the intended interpretation of the image pixel data.
Rows	Number of rows in the image.
Columns	Number of columns in the image.
XYSpacing	Physical distance between the center of each pixel along the x and y axes, respectively, in the patient.
RescaleIntercept	The value b in relationship between stored values (SV) and the output units.
RescaleSlope	m in the equation specified by Rescale Intercept.
RescaleType	Specifies the output units of Rescale Slope and Rescale Intercept.
NumberofSlices	Number of slices.
ZSpacing	Physical distance between the center of each pixel along the z axis in the patient.
StudyDate	Date the Study started.

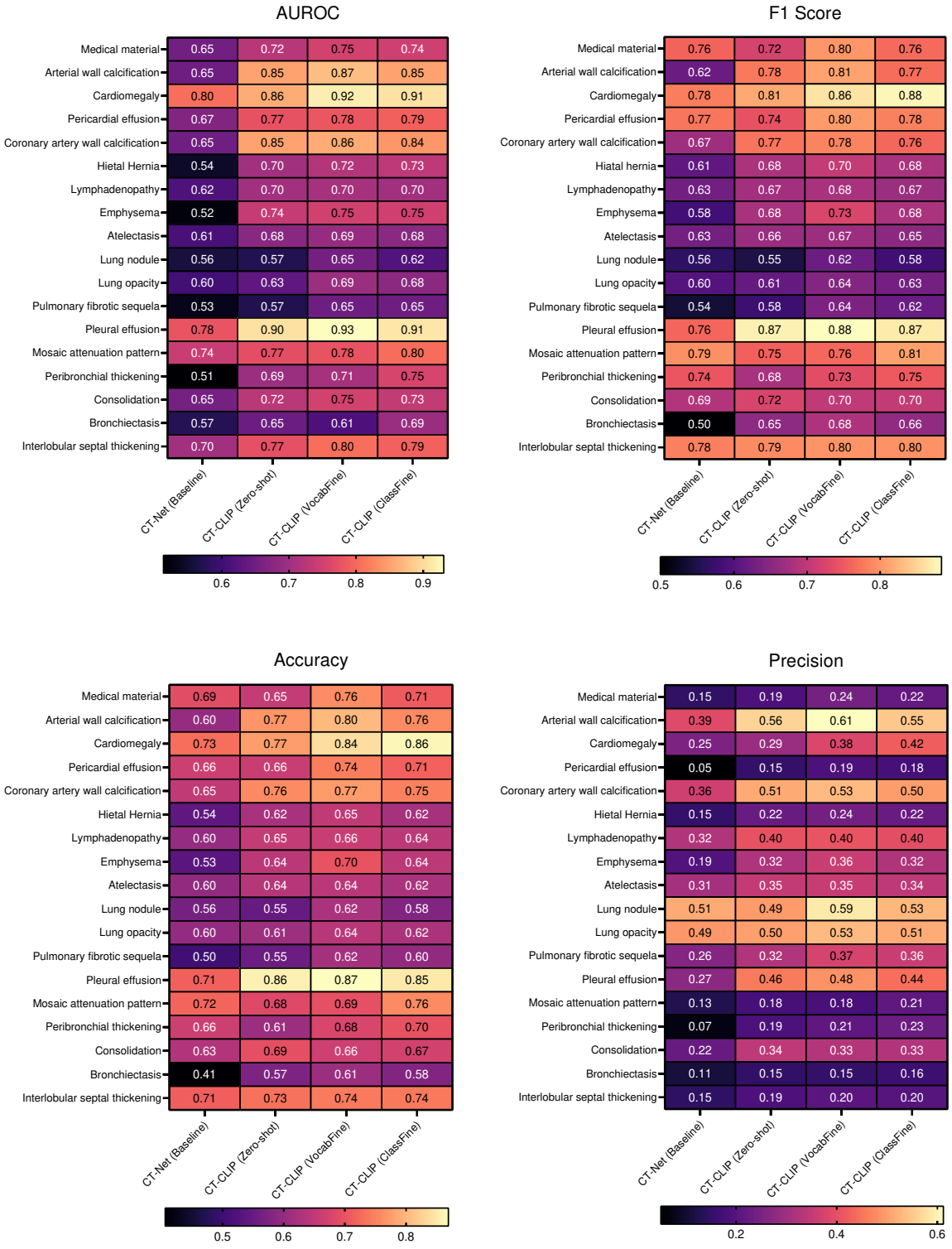
Extended Data Table 1: Metadata attributes. Detailed breakdown of the metadata attributes accessible within the CT-RATE dataset.



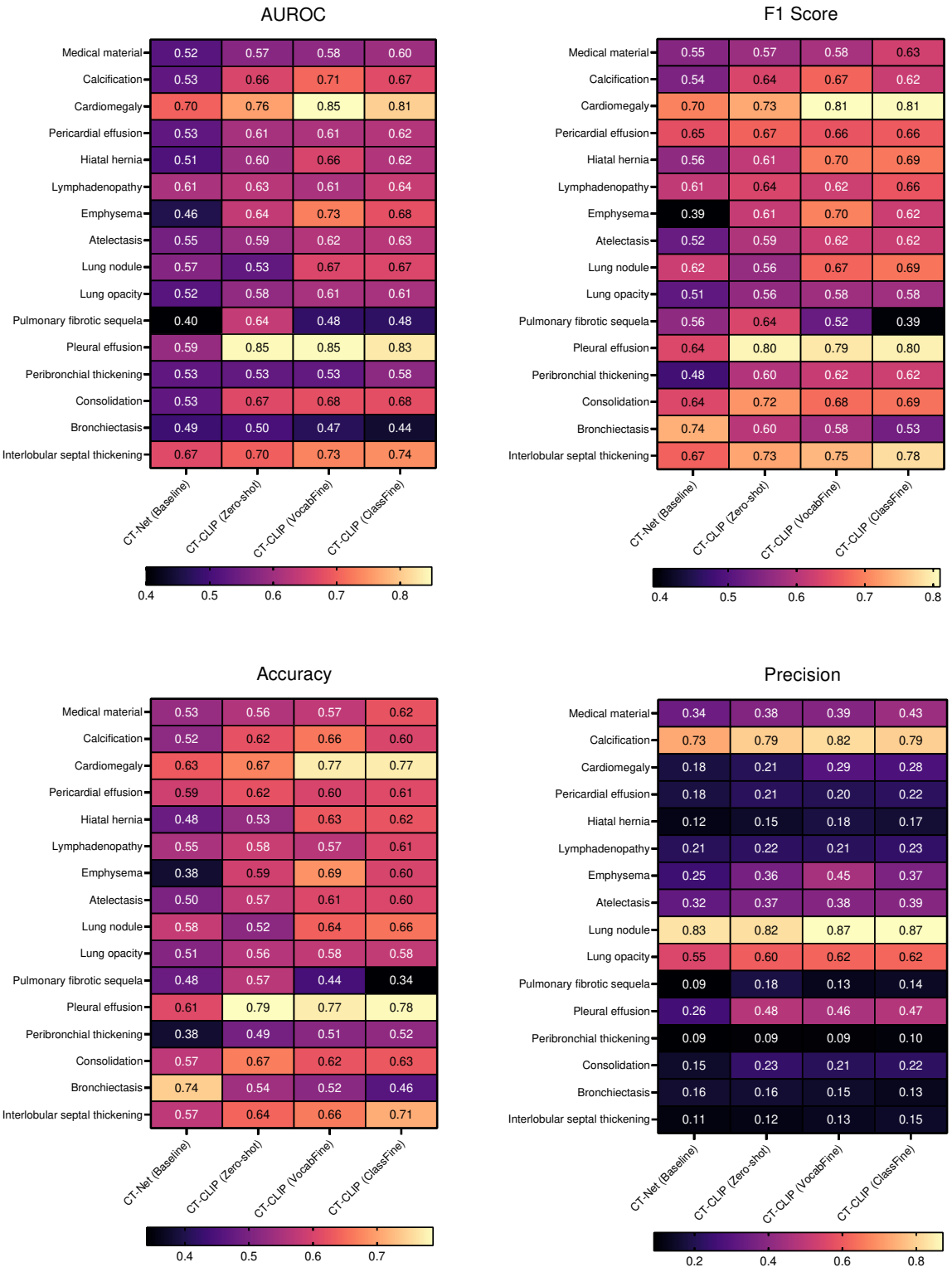
Extended Data Figure 1: Abnormality-based t-SNE projections for chest CT volumes and radiology text reports.



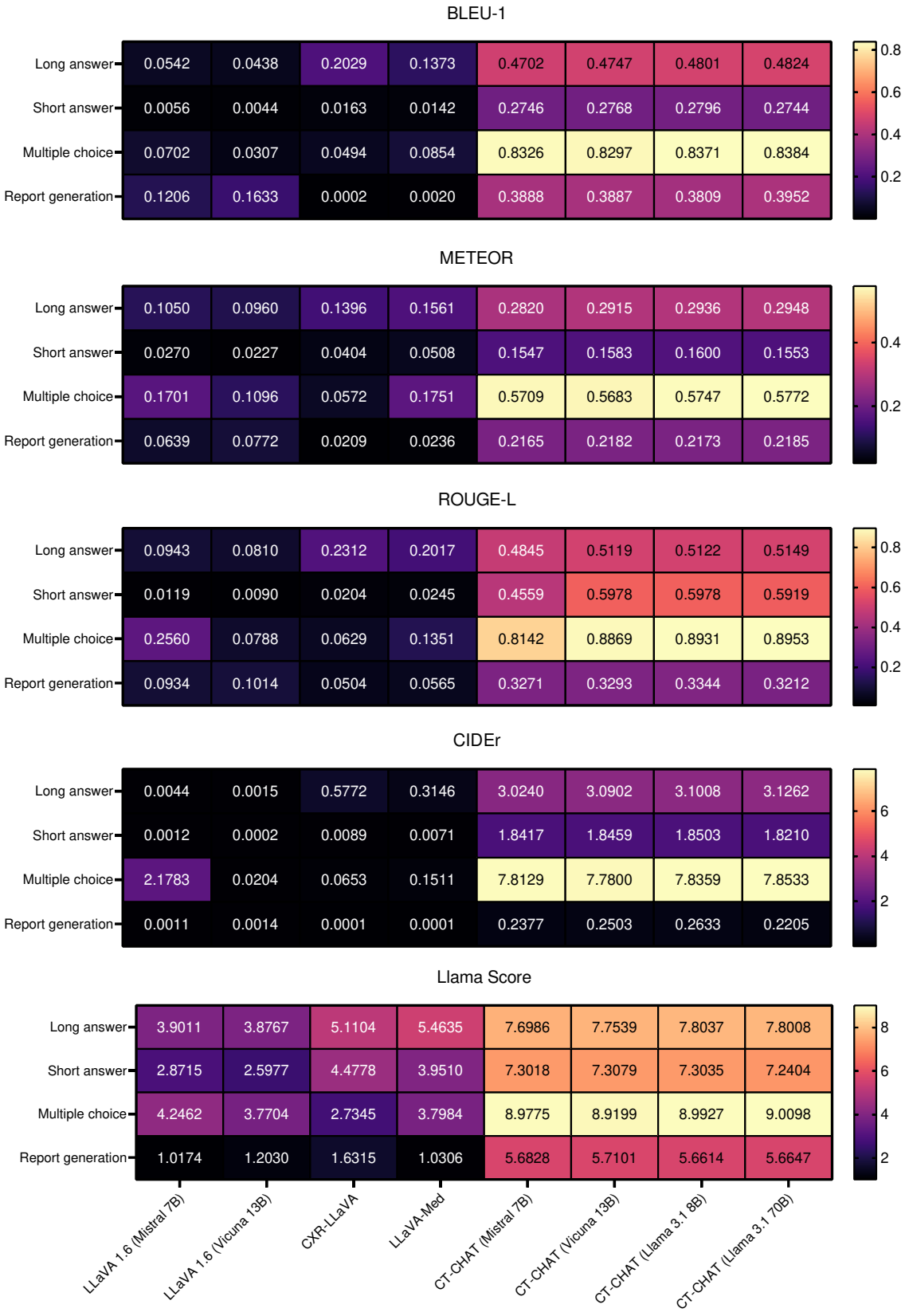
Extended Data Figure 2: Abnormality-based t-SNE projections for chest CT volumes and radiology text reports (continued).



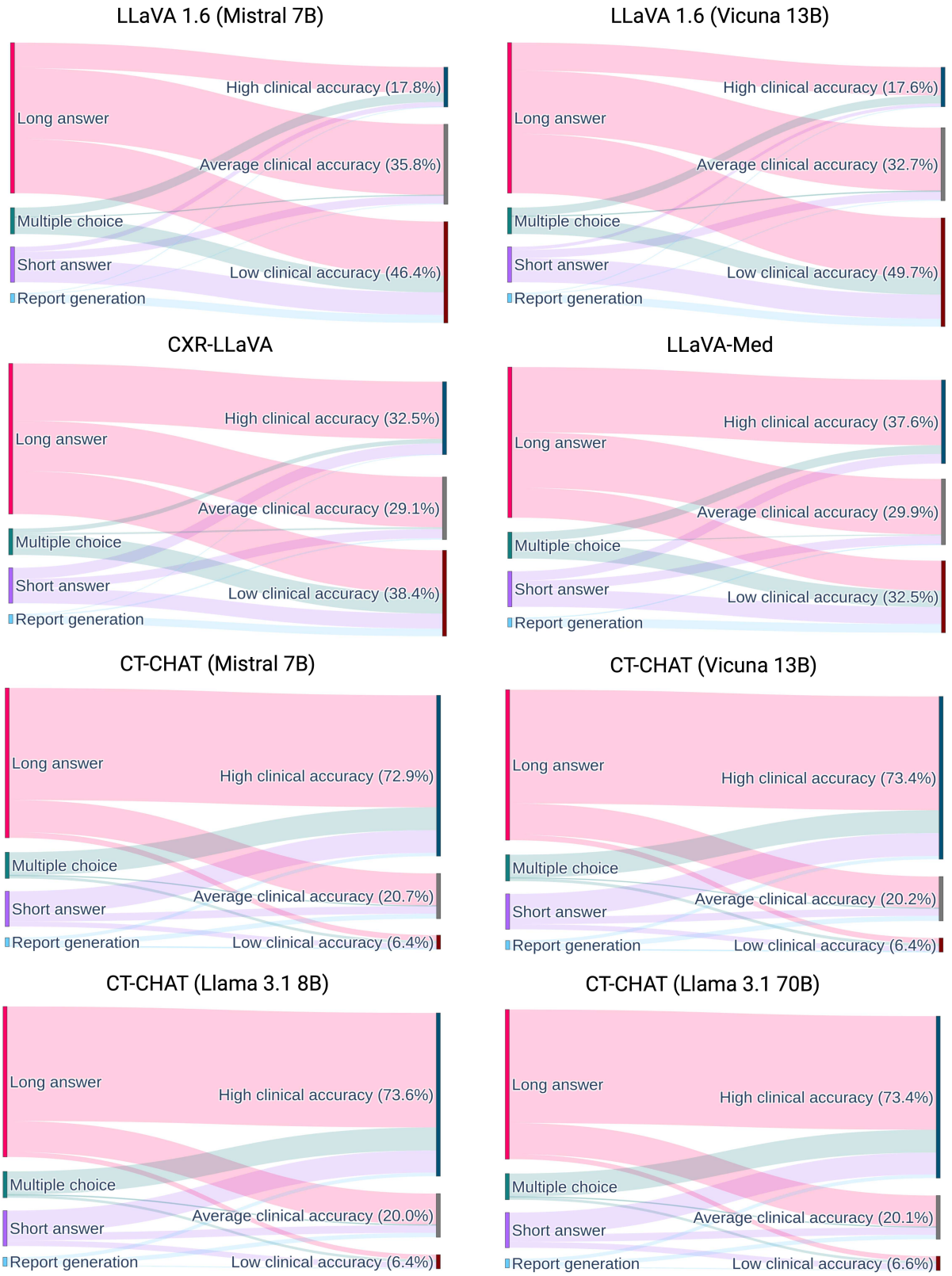
Extended Data Figure 3: Comparison of abnormality-based performance metrics in the internal validation set. This figure provides a detailed analysis of performance metrics, including AUROC, accuracy, precision, and F1 scores, for detecting various abnormalities with our models in the internal validation set, compared to the fully supervised baseline model. The proposed CT-CLIP based models demonstrate improvement over the baseline in almost all comparisons.



Extended Data Figure 4: Comparison of abnormality-based performance metrics in the external validation set. This figure provides a detailed analysis of performance metrics, including AUROC, accuracy, precision, and F1 scores, for detecting various abnormalities with our models in external validation set, compared to the fully supervised baseline model. It highlights the models' remarkable adaptability and superior effectiveness with distribution shifts, setting a new standard in performance compared to a fully supervised baseline model.



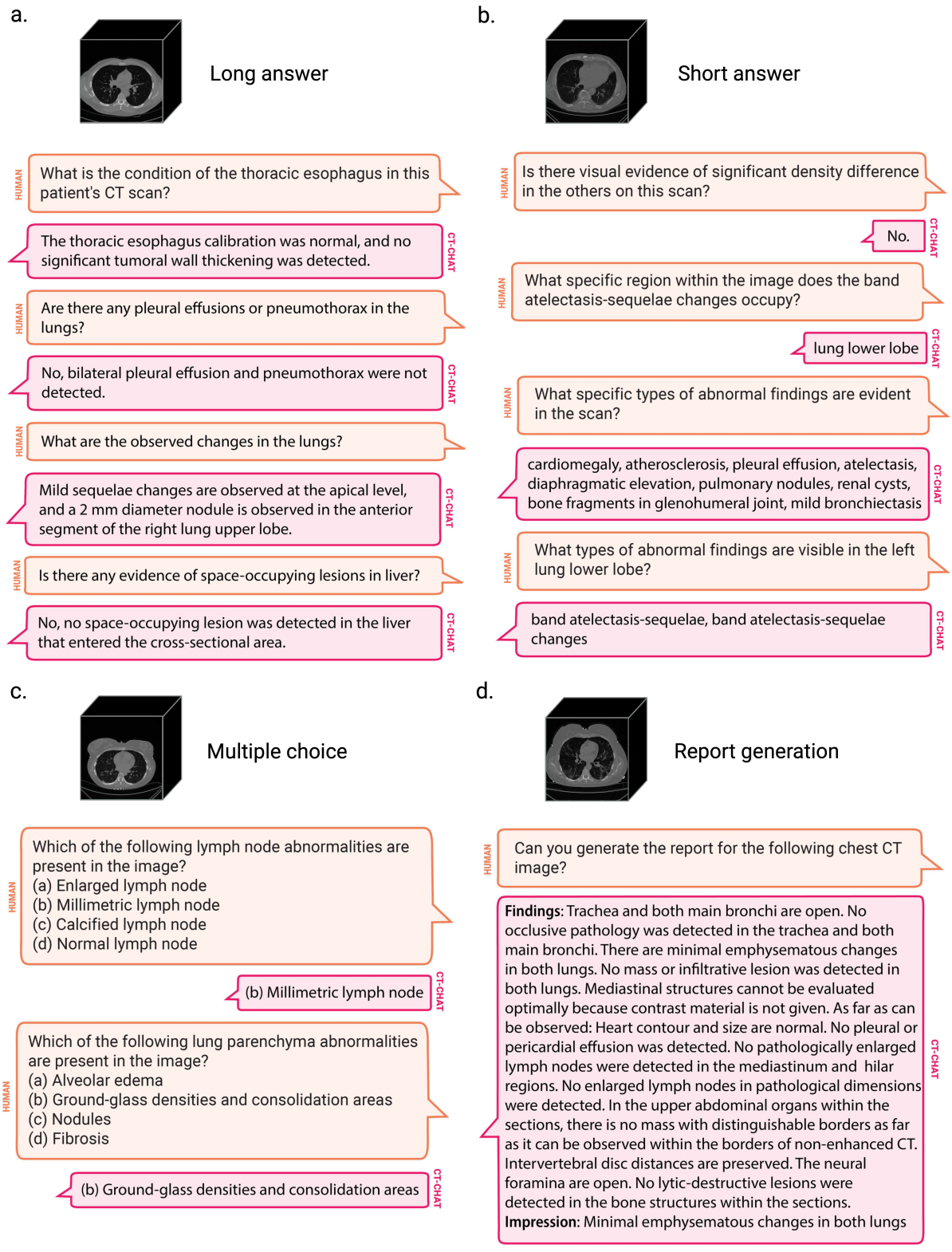
Extended Data Figure 5: Comparative metrics for CT-CHAT and baselines. This figure illustrates the performance of CT-CHAT and baselines across various tasks, demonstrating that CT-CHAT consistently outperforms the baselines, achieving the highest accuracy in every metric for each task.



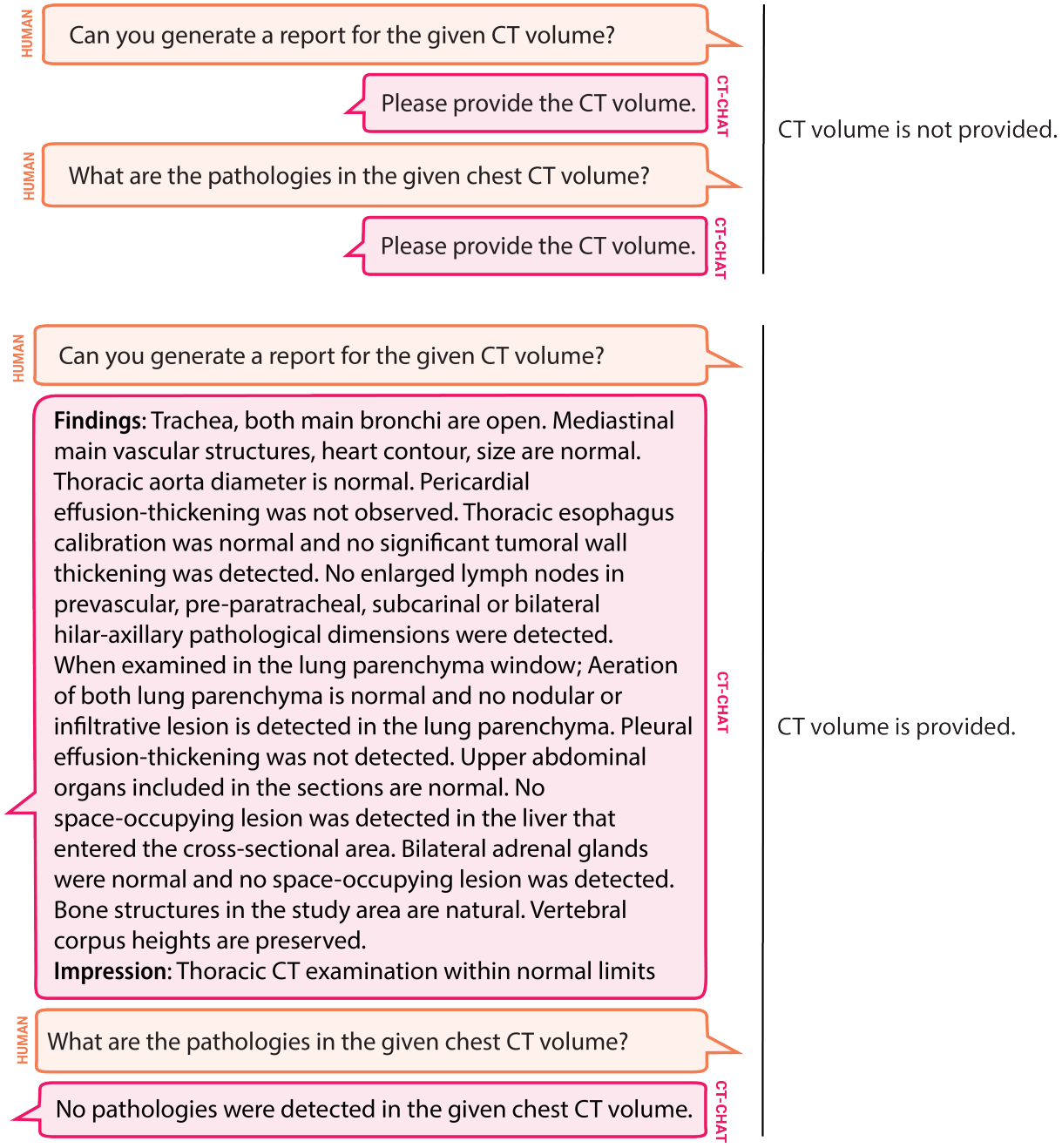
Extended Data Figure 6: Clinical accuracy comparison of VQA models. This figure provides a detailed assessment of clinical accuracy across different models, showing that CT-CHAT produces a higher proportion of outputs in the high clinical accuracy category, indicating superior performance.



Extended Data Figure 7: Representative examples from VQA. This dataset includes various question types, such as long-answer (which encompass conversational, descriptive, free-response, and text-only questions), multiple-choice, short-answer, and report-generation questions.



Extended Data Figure 8: Example outputs of CT-CHAT for (a) long answer, (b) short answer, (c) multiple-choice, and (d) report generation questions.



Extended Data Figure 9: Guardrail example to prevent hallucination of CT-CHAT when the CT volume is not provided.

a

Abnormality	Count in train set	Count in internal validation set	Train set ratio	Internal validation set ratio
Medical material	5818	313	0.123	0.103
Arterial wall calcification	13377	867	0.284	0.285
Cardiomegaly	5308	325	0.113	0.107
Pericardial effusion	3412	226	0.072	0.074
Coronary artery wall calcification	12025	765	0.255	0.252
Hiatal hernia	6751	417	0.143	0.137
Lymphadenopathy	12221	789	0.259	0.260
Emphysema	9122	600	0.193	0.197
Atelectasis	12263	713	0.260	0.235
Lung nodule	21382	1361	0.453	0.448
Lung opacity	17420	1184	0.369	0.390
Pulmonary fibrotic sequela	12589	831	0.267	0.273
Pleural effusion	5705	376	0.121	0.124
Mosaic attenuation pattern	3638	253	0.077	0.083
Peribronchial thickening	4973	355	0.105	0.117
Consolidation	8319	581	0.176	0.191
Bronchiectasis	4732	330	0.100	0.109
Interlobular septal thickening	3745	249	0.079	0.082

b

Manufacturer	Count in train set	Count in internal validation set	Train set ratio	Internal validation set ratio
Philips	28966	1928	0.614	0.634
Siemens	15946	971	0.338	0.320
PNMS	2237	140	0.048	0.046

c

Sex	Count in train set	Count in internal validation set	Train set ratio	Internal validation set ratio
Male	27462	1776	0.583	0.584
Female	19681	1263	0.417	0.416

Supplementary Table 1: Detailed overview of the CT-RATE dataset. a. The counts and proportions of various abnormalities identified within the training and validation subsets, offer insights into dataset composition. **b.** Details of the distribution of chest CT volumes according to manufacturer across both subsets, reflecting the dataset's diversity. **c.** The breakdown of chest CT volumes by sex in both subsets, provides a demographic perspective of the data.

Abnormality	Precision	Recall	F1 Score
Medical material	1.00	0.94	0.97
Arterial wall calcification	0.96	0.97	0.96
Cardiomegaly	1.00	0.96	0.98
Pericardial effusion	1.00	1.00	1.00
Coronary artery wall calcificatior	0.96	1.00	0.98
Hiatal hernia	0.96	0.96	0.96
Lymphadenopathy	0.94	1.00	0.97
Emphysema	1.00	0.96	0.98
Atelectasis	0.98	0.96	0.97
Lung nodule	0.95	1.00	0.98
Lung opacity	0.94	0.96	0.95
Pulmonary fibrotic sequela	0.97	0.96	0.96
Pleural effusion	1.00	1.00	1.00
Mosaic attenuation pattern	1.00	0.96	0.98
Peribronchial thickening	1.00	1.00	1.00
Consolidation	0.95	1.00	0.97
Bronchiectasis	1.00	0.91	0.95
Interlobular septal thickening	1.00	1.00	1.00

Supplementary Table 2: Finetuned text encoder's performance in automatic label extraction. This table presents a comprehensive assessment of performance metrics—including precision, recall, and F1 scores—for the finetuned text encoder across various abnormalities. The evaluation underscores the encoder's accuracy and efficiency in automating the extraction of labels for each abnormality.

	CT-Net (Baseline)	CT-CLIP (Zero-shot)	CT-CLIP (VocabFine)	CT-CLIP (ClassFine)
Optimizer	Adam	Adam	Adam	Adam
$\beta 1$	0.9	0.9	0.9	0.9
$\beta 2$	0.99	0.99	0.99	0.99
Learning rate	0.001	0.00000125	0.000	0.00001
Batch size	4	8	1	1
Schedular	N/A	N/A	Cosine annealing	Cosine annealing
Warmup steps	N/A	N/A	10,000	10,000
Max steps	N/A	N/A	20,000	20,000
GPUs utilized	1	4	1	1
Train duration	4 days	3 days	2 days	2 days
Iterations	100,000	100,000	20,000	20,000

Supplementary Table 3: Hyperparameters for training each model of the multi-abnormality detection experiments.

a

	CT-CHAT (Mistral 7B)	CT-CHAT (Vicuna 13B)	CT-CHAT (Llama 3.1 8B)	CT-CHAT (Llama 3.1 70B)
Automatic mixed precision	FP16	FP16	FP16	FP16
Per device batch size	12	12	12	12
Gradient accumulation steps	1	1	1	1
Learning rate scheduler	Cosine	Cosine	Cosine	Cosine
Warmup ratio	0.03	0.03	0.03	0.03
Peak learning rate	0.001	0.001	0.001	0.001
AdamW β	(0.9,0.999)	(0.9,0.999)	(0.9,0.999)	(0.9,0.999)
AdamW ϵ	1e-8	1e-8	1e-8	1e-8
Epochs	1	1	1	1

b

	CT-CHAT (Mistral 7B)	CT-CHAT (Vicuna 13B)	CT-CHAT (Llama 3.1 8B)	CT-CHAT (Llama 3.1 70B)
Automatic mixed precision	FP16	FP16	FP16	FP16
Per device batch size	3	3	3	3
Gradient accumulation steps	1	1	1	1
Learning rate scheduler	Cosine	Cosine	Cosine	Cosine
Warmup ratio	0.03	0.03	0.03	0.03
Peak learning rate	2e-5	2e-5	2e-5	2e-5
AdamW β	(0.9,0.999)	(0.9,0.999)	(0.9,0.999)	(0.9,0.999)
AdamW ϵ	1e-8	1e-8	1e-8	1e-8
Epochs	1	1	1	1

Supplementary Table 4: Hyperparameters for (a) pretraining and (b) finetuning during each CT-CHAT ablation.

Comparison	AUROC	F1 Score	Accuracy	Precision
CT-Net (Baseline) vs CT-CLIP (Zero-shot)	0.0002	0.004	0.0016	0.062
CT-Net (Baseline) vs CT-CLIP (VocabFine)	0.0002	0.0002	0.0002	0.015
CT-Net (Baseline) vs CT-CLIP (ClassFine)	0.0002	0.001	0.0002	0.025
CT-CLIP (Zero-shot) vs CT-CLIP (VocabFine)	0.431	0.252	0.167	0.509
CT-CLIP (Zero-shot) vs CT-CLIP (ClassFine)	0.403	0.538	0.436	0.710
CT-CLIP (VocabFine) vs CT-CLIP (ClassFine)	0.993	0.593	0.554	0.753

Supplementary Table 5: Statistical significance. Utilizing a two-sided unpaired permutation test with 10,000 permutations, we observe statistically significant differences ($p < 0.05$) in AUROC, F1 score, and accuracy between the supervised baseline model and each of the CT-CLIP models. For precision, significant differences are noted between the supervised baseline model and both finetuning.

Report generation questions

- Can you generate the report for the following chest CT image?
- Please provide the radiology report for the chest CT image mentioned.
- I need the radiology report for the given chest CT image.
- Could you create a report for this chest CT scan?
- Would you mind generating the radiology report for the specified chest CT image?
- Please generate the report for the chest CT image provided.
- Can you produce the radiology report for the attached chest CT image?
- I need a detailed report for the given chest CT image.
- Could you write the radiology report for this chest CT scan?
- Please give the radiology report for the specified chest CT image.
- Generate radiology report for the CT.
- Produce the report for this CT image.
- Write a radiology report for the following CT scan.
- Create a report for this chest CT.
- Provide the radiology report for this CT image.
- Can you generate the report for the following chest CT volume?
- Please provide the radiology report for the chest CT volume mentioned.
- I need the radiology report for the given chest CT volume.
- Could you create a report for this chest CT volume?
- Would you mind generating the radiology report for the specified chest CT volume?
- Please generate the report for the chest CT volume provided.
- Can you produce the radiology report for the attached chest CT volume?
- I need a detailed report for the given chest CT volume.
- Could you write the radiology report for this chest CT volume?
- Please give the radiology report for the specified chest CT volume.
- Generate radiology report for the CT volume.
- Produce the report for this CT volume.
- Write a radiology report for the following CT volume.
- Create a report for this chest CT volume.
- Provide the radiology report for this CT volume.
- Can you generate the report for the following chest CT scan?
- Please provide the radiology report for the chest CT scan mentioned.
- I need the radiology report for the given chest CT scan.
- Could you create a report for this chest CT scan?
- Would you mind generating the radiology report for the specified chest CT scan?
- Please generate the report for the chest CT scan provided.
- Can you produce the radiology report for the attached chest CT scan?
- I need a detailed report for the given chest CT scan.
- Could you write the radiology report for this chest CT scan?
- Please give the radiology report for the specified chest CT scan.
- Generate radiology report for the CT scan.
- Produce the report for this CT scan.
- Write a radiology report for the following CT scan.
- Create a report for this chest CT scan.
- Provide the radiology report for this CT scan.

Supplementary Table 6: Predefined report generation questions. These questions are randomly sampled while generating the report generation conversations.

Prompt to generate conversation questions

messages = [{"role": "system", "content": "You are a radiologist specializing in creating VQA (Visual Question Answering) datasets for Chest CT volumes. Generate three sets of conversation-style Q&A pairs based on the CT volume information, using the provided report as background without referring directly to it.

Enclose each set within tags <1>...<1>, <2>...<2>, and <3>...<3>. Use <q>...</q> for questions and <a>... for answers, ensuring most answers are derived from the images.

Example:
<1><q>Question 1</q><a>Answer 1<q>Question 2</q><a>Answer 2...<1><2><q>Question 1</q><a>Answer 1<q>Question 2</q><a>Answer 2...<2><3><q>Question 1</q><a>Answer 1<q>Question 2</q><a>Answer 2...<3>.

You can include as many question and answer couples as you find appropriate. Here is the report for you to get information about CT volume:\n"},

{"role": "user", "content": f"{report}"}]

Prompt to generate free response, description, and multiple choice questions

messages = [{"role": "system", "content": "You are a radiologist VQA dataset creator specializing in Chest CT reports. I need you to generate three sets of questions based on the given radiology report: free response, description, and multiple-choice questions. Each set should contain 3 Q/A pairs.

The free response questions should be enclosed within <free_response>...</free_response>. The description questions should be enclosed within <description>...</description>. The multiple-choice questions should be enclosed within <multiple_choice>...</multiple_choice>, with each question having 4 choices labeled as (a), (b), (c), and (d).

Ensure that the questions are specific to the Chest CT image described in the report.

Example for free response: <q>What abnormalities are present in the lung fields of this Chest CT image?<q>
Example for description: <q>Describe the findings shown in this Chest CT image.<q>
Example for multiple-choice: <q>Which of the following abnormalities are present in this Chest CT image?
(a) Choice 1 (b) Choice 2 (c) Choice 3 (d) Choice 4<q><a>Choice 2<a>

The structure should look like this:
<free_response><q>Question 1<q><a>Answer 1<a><q>Question 2<q><a>Answer 2<a><q>Question 3<q><a>Answer 3<a></free_response>
<description><q>Question 1<q><a>Answer 1<a><q>Question 2<q><a>Answer 2<a><q>Question 3<q><a>Answer 3<a></description>
<multiple_choice><q>Question 1 (a) Choice 1 (b) Choice 2 (c) Choice 3 (d) Choice 4<q><a>(b) Choice 2<a><q>Question 2 (a) Choice 1 (b) Choice 2 (c) Choice 3 (d) Choice 4<q><a>(a) Choice 1<a><q>Question 3 (a) Choice 1 (b) Choice 2 (c) Choice 3 (d) Choice 4<q><a>(c) Choice 3<a></multiple_choice>

Please always follow this format to create the questions and answers."},

{"role": "user", "content": f"{report}"}]

Prompt to generate text only questions

messages = [{"role": "system","content": "You are a chest CT dataset creator specializing in generating question-and-answer sets based on radiology articles focusing on chest CTs. I need you to create 5 question-and-answer pairs for each article.

Each question-and-answer pair should be enclosed within tags like <q>...<q> for questions and <a>...<a> for answers. Also include all output in <1>...<1> (remember to put one at the beginning one at the end).

For example, the structure should look like this: <1><q>Question 1<q><a>Answer 1<a><q>Question 2<q><a>Answer 2<a><q>Question 3<q><a>Answer 3<a><q>Question 4<q><a>Answer 4<a><q>Question 5<q><a>Answer 5<a><1>.

Please follow this format to create the conversation questions and answers. Make sure the questions are variable, educative, and the answers are long and instructive enough.The questions should be those that a student or radiologist might ask about chest CT images in general. Do not generate questions specific to the content or research idea of the provided article. Focus solely on general knowledge about chest CT imaging. Examples of such questions include:

- What are the typical imaging findings of interstitial lung disease on a chest CT?
- How is ground-glass opacity on a chest CT typically interpreted?
- What are the common causes of lymphadenopathy seen in chest CT scans?
- How can chest CT help in the diagnosis of pulmonary embolism?
- What are the distinguishing features of lung nodules on a chest CT?
- How are lung nodules found in chest CT scans managed?

The answers must be factually accurate and should be retrieved from the given article (but not about the case just the given factual knowledge). Do not reference specific details, cases, or research findings from the provided article. Ensure that the information in the answers is universally applicable to chest CT imaging. "},

{"role": "user", "content": f"{article_title}\n\n{article_content}"}]

Supplementary Table 7: The prompts that are used to generate conversation, description, free response, multiple choice, and text-only questions.

Prompt for clinical accuracy evaluation with Llama 3.1 70B

messages = [{"**role":** "system", **"content":** "**Role:** You are an expert medical evaluator specializing in the assessment of clinical accuracy for AI-generated reports and answers derived from 3D chest CT volumes. Your task is to compare the AI-generated answers with the provided ground truth answers and assign a clinical accuracy score based on specific evaluation criteria. Your assessment should focus on precision, relevance, and clinical soundness.

Scoring:

• **10/10:** The generated response is fully aligned with the ground truth—completely accurate, relevant, comprehensive, clear, and clinically sound.
• **7-9/10:** The generated response is mostly accurate, with minor discrepancies or omissions that do not significantly affect clinical interpretation.
• **4-6/10:** The generated response contains noticeable errors or omissions that could impact clinical understanding, though some correct information is present.
• **1-3/10:** The generated response has significant inaccuracies, irrelevant content, or lacks essential information, potentially leading to clinical misunderstanding.
• **0/10:** The generated response is completely incorrect or irrelevant, offering no clinical value.

Question Types: For each evaluation, you will be provided with the question type, question, ground truth answer, and generated answer.

• **Report Generation:** Requires a detailed narrative report summarizing findings from the 3D chest CT volume, covering all relevant clinical observations, including pathology, location, and significant findings. The report should follow the CT report format. Consider the formatting of the output for this question in addition to clinical accuracy. Non-existent pathologies are not necessarily required to be mentioned in the generated report. Do not deduct points if this is the case (e.g., "no pleural effusion" might not be in the output, but if "pleural effusion" is present, it must be included). Deduct points if the model hallucinates pathologies that are not in the ground truth.
• **Long Answer Questions:** Require comprehensive, detailed responses that thoroughly address the question, including necessary explanations or justifications.
• **Short Answer Questions:** Require brief, direct answers addressing the question with essential information, such as identifying pathology or confirming a diagnosis.
• **Multiple Choice Questions:** Require the selection of the most clinically accurate option from predefined choices, matching the ground truth answer. There should be only one correct answer for this question type. If more than one answer is provided, assign a very low score. If the answer is incorrect, assign a very low score as well.

Instructions: Please review the question type, question, ground truth answer, and generated answer. Assign a score out of 10 based on the evaluation criteria above. Provide only the score without additional commentary or explanation."},

{"**role":** "user", **"content":** f"

Question type: {question_type}

Question: {question}

Ground truth answer: {ground_truth}

Generated answer: {model_output}"}]

Supplementary Table 8: The prompt that is used to evaluate the clinical accuracy of the outputs generated by the VQA models using the LLM.

System prompt for CT-CHAT

messages = [{"**role":** "system", **"content":** " You are CT-CHAT, an AI assistant specializing in Chest CT imaging, dedicated to providing accurate and relevant information exclusively related to Chest CT scans and associated medical topics. You are equipped to answer questions and offer detailed analyses only when the CT volume/scan/image is provided, indicated by the <provided> token. If this token is not present and users inquire about specific findings, pathologies, or request descriptions related to a Chest CT, respond by requesting the necessary data with the phrase: "Please provide the CT volume." Once the <provided> token is present in the question, you are authorized to address questions about pathologies, anatomical or clinical findings, diagnostic descriptions, report generation, comparisons, or any other questions regarding the image. If it does not appear in the question, even when special tokens <multiple_choice>, <report_generation>, <long_answer>, and <short_answer> are given, ignore the question and ask for the CT volume. Always look for the <provided> token, even if there are special tokens. If there is a <provided> token in any question (including new and previous ones), never ask for the CT volume again and answer the question. You can ignore the <provided> token check and answer the question directly if and only if the question is about general medical knowledge, not about the provided CT volume (such as typical findings on a Chest CT or management of the patient). For example, "What are the typical imaging findings of acute respiratory distress syndrome (ARDS) on a chest CT?" is a general question not specific. If user asks a CT specific question after non-spesific question, look for the <provided> token as well even if the special tokens are given. It is crucial to avoid discussing topics outside of Chest CT imaging and directly related medical information, ensuring that all responses are clear, concise, and focused on the provided Chest CT data for the highest level of accuracy and relevance. If the user greets you with something like "hello," respond appropriately."},
{"**role":** "user", **"content":** f"{question}"}}]

Supplementary Table 9: The system prompt for CT-CHAT.

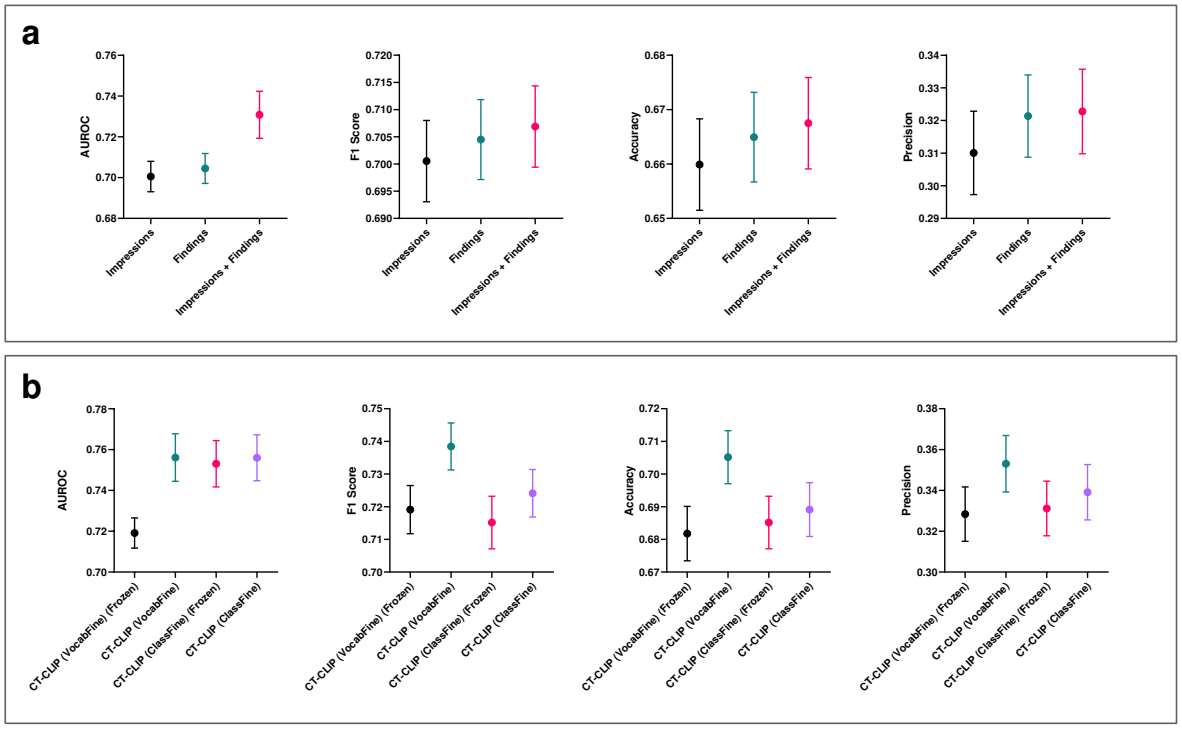
a

Metric	Long answer		Short answer		Multiple choice		Report generation	
	without image	with image	without image	with image	without image	with image	without image	with image
↑ BLEU-1	0.383	0.482	0.248	0.274	0.827	0.838	0.125	0.395
↑ METEOR	0.259	0.295	0.140	0.155	0.568	0.577	0.085	0.218
↑ ROUGE-L	0.434	0.515	0.565	0.592	0.886	0.895	0.129	0.321
↑ CIDEr	2.234	3.126	1.715	1.821	7.776	7.853	0.003	0.220
↑ Llama Score	7.448	7.801	6.902	7.240	8.903	9.010	2.130	5.665

b

Metric	Long answer		Short answer		Multiple choice		Report generation	
	without pretrain	with pretrain	without pretrain	with pretrain	without pretrain	with pretrain	without pretrain	with pretrain
↑ BLEU-1	0.482	0.482	0.277	0.274	0.837	0.838	0.390	0.395
↑ METEOR	0.294	0.295	0.156	0.155	0.576	0.577	0.218	0.218
↑ ROUGE-L	0.514	0.515	0.596	0.592	0.893	0.895	0.323	0.321
↑ CIDEr	3.115	3.126	1.839	1.821	7.839	7.853	0.222	0.220
↑ Llama Score	7.816	7.801	7.271	7.240	9.009	9.010	5.785	5.665

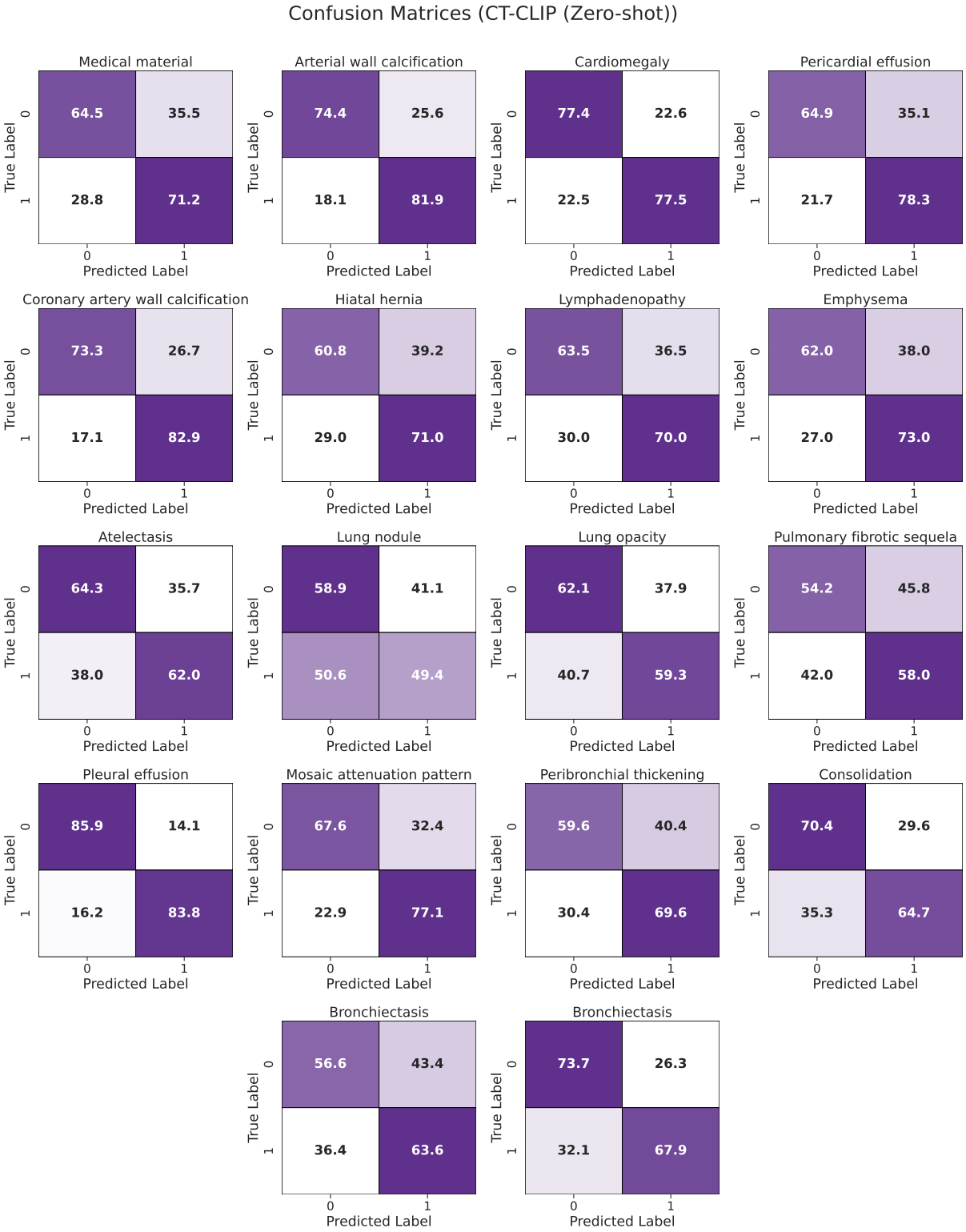
Supplementary Table 10: Ablations of CT-CHAT for (a) image provision and (b) medical concept alignment. The concept alignment has minimal impact on model accuracy while providing the volumes is essential for the model to accurately answer questions, especially in report generation.



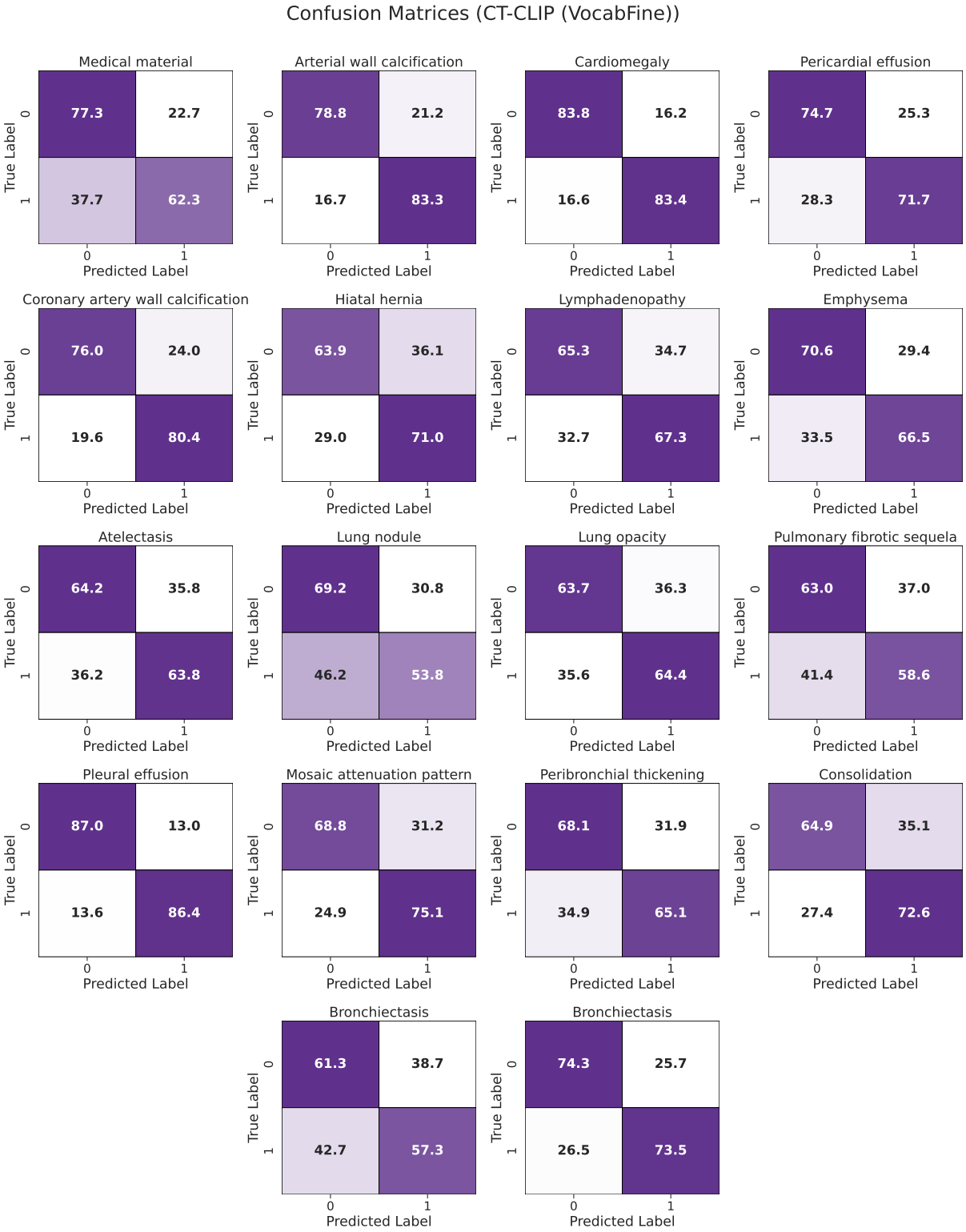
Supplementary Figure 1: Detailed ablation study. a. Compares the zero-shot abnormality detection performance of CT-CLIP models, each utilizing different sections of radiology text reports during training. **b.** Compares various finetuning techniques applied to CT-CLIP models: *frozen* models involve finetuning only the latent layers for the Vocabfine approach, while for the ClassFine method, only the new linear layer is finetuned, leaving all other layers unchanged.



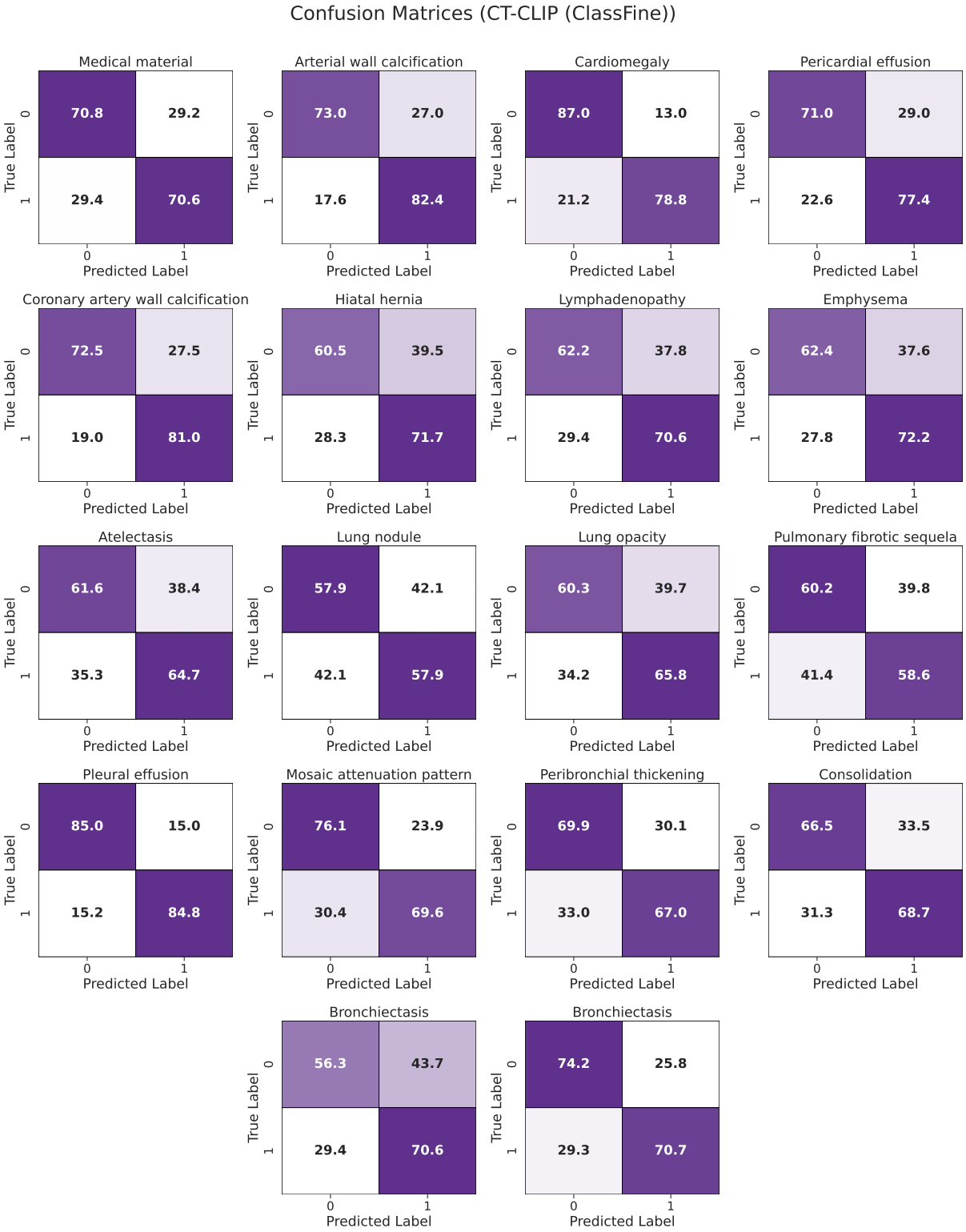
Supplementary Figure 2: Illustration of confusion matrices for the fully supervised baseline model in the internal validation set.



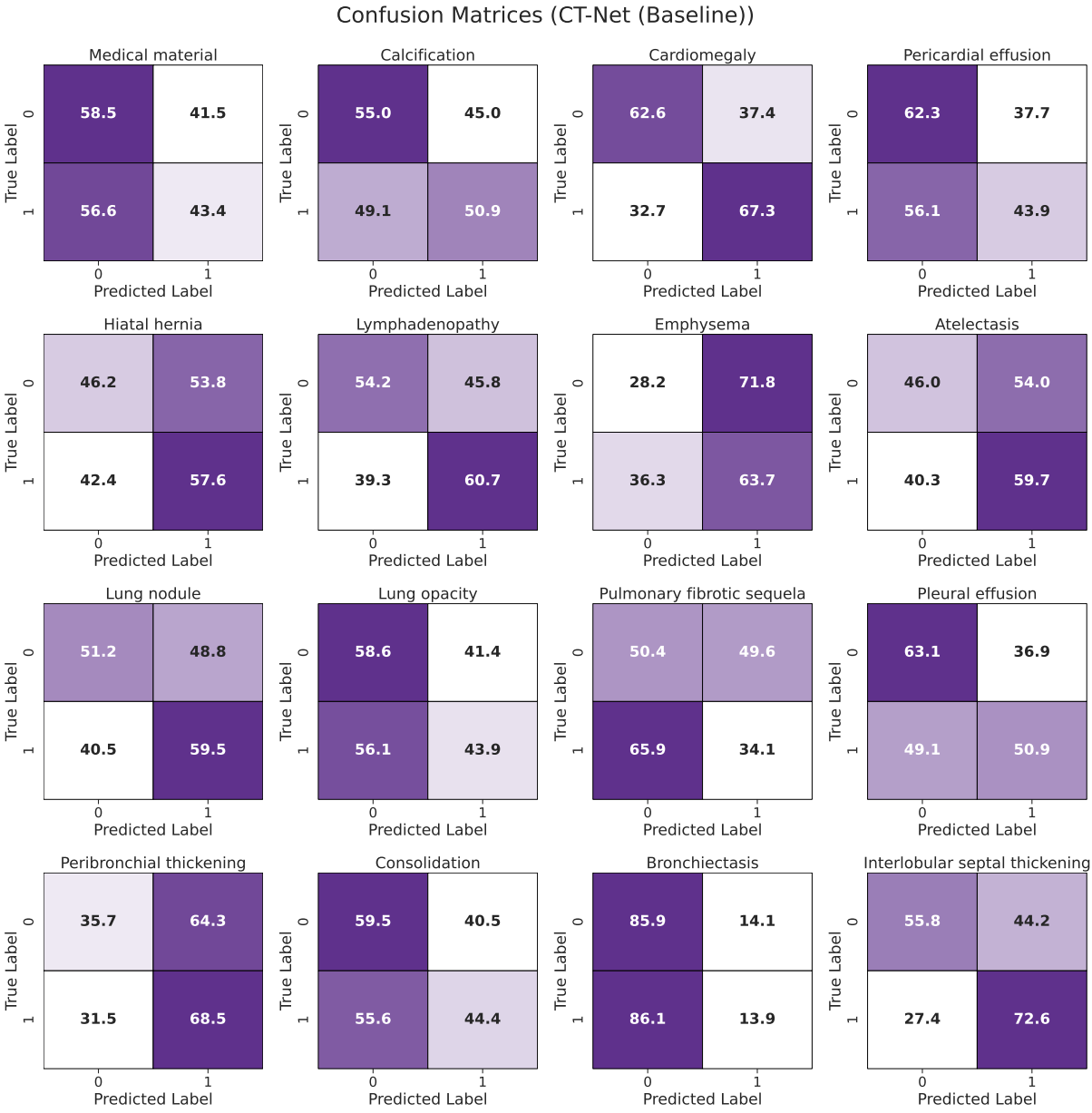
Supplementary Figure 3: Illustration of confusion matrices for the CT-CLIP model in the internal validation set.



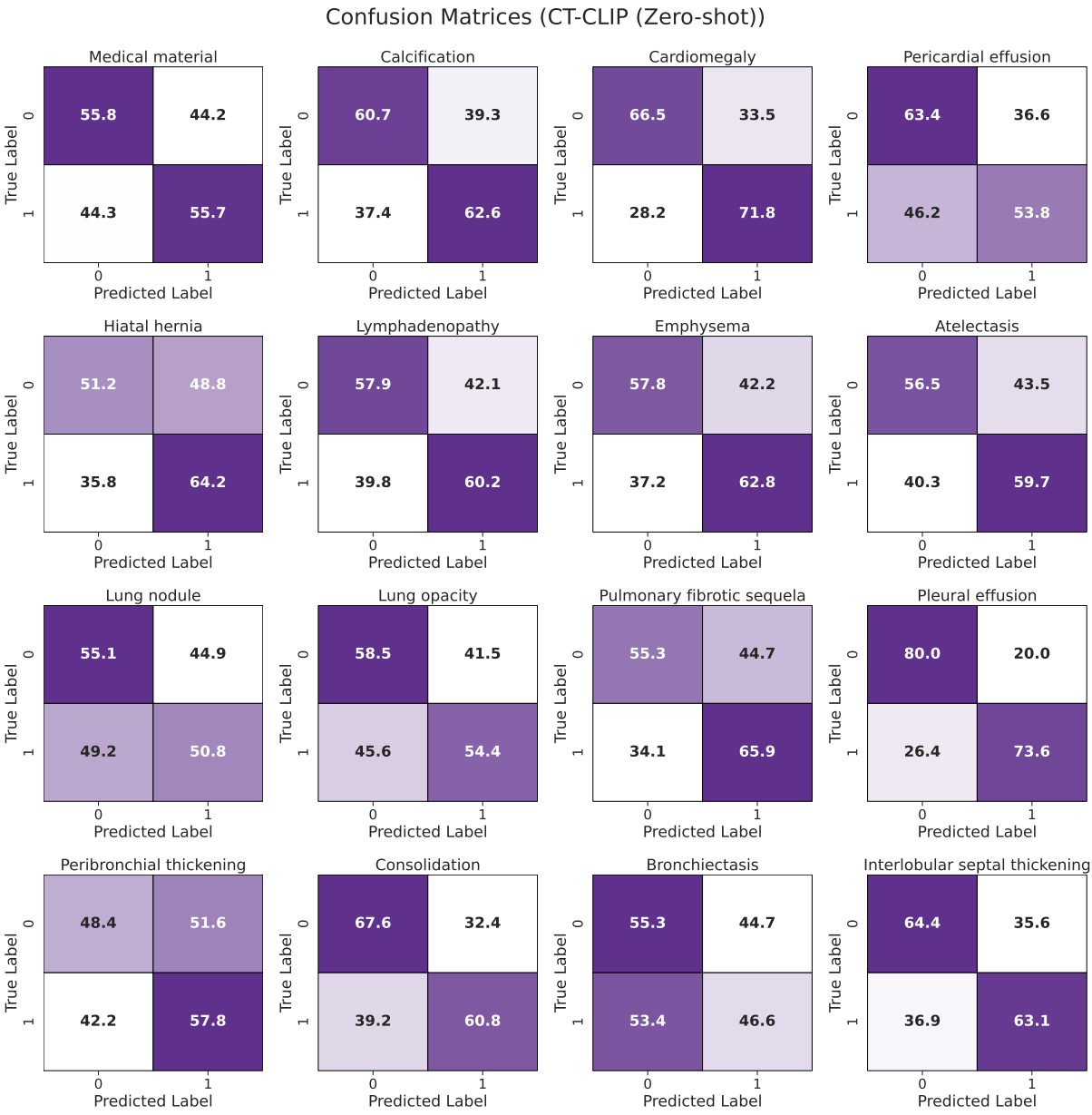
Supplementary Figure 4: Illustration of confusion matrices for the CT-CLIP model fine-tuned with VocabFine in the internal validation set.



Supplementary Figure 5: Illustration of confusion matrices for the CT-CLIP model fine-tuned with ClassFine in the internal validation set.



Supplementary Figure 6: Illustration of confusion matrices for the fully supervised baseline model in the external validation set.



Supplementary Figure 7: Illustration of confusion matrices for the CT-CLIP model in the external validation set.



Supplementary Figure 8: Illustration of confusion matrices for the CT-CLIP model fine-tuned with VocabFine in the external validation set.



Supplementary Figure 9: Illustration of confusion matrices for the CT-CLIP model fine-tuned with ClassFine in the external validation set.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- CTCHATdemo.mov