nature communications

Article

https://doi.org/10.1038/s41467-025-66620-z

A lung CT vision foundation model facilitating disease diagnosis and medical imaging

Received: 23 January 2025

Accepted: 11 November 2025

Published online: 03 December 2025

Check for updates

Zebin Gao[1,2,10], Guoxun Zhang[1,3,10], Hengrui Liang[4,10], Jiaxin Liu[5], Liangdi Ma[6], Tianyun Wang[2], Yanchen Guo[5], YuJia Chen[5], Zeping Yan[4], Xiangru Chen[7], Jianxing He[4] ✉, Feng Xu[1,6] ✉, Tien Yin Wong[8,9] ✉, Yuchen Guo[1] ✉ & Qionghai Dai[1,2,3] ✉

The concomitant development and evolution of lung computed tomography (CT) and artificial intelligence (AI) have made non-invasive lung imaging a key component of clinical care of patients. However, the scarcity of labeled CT data and the limited generative capacity of existing models have constrained their clinical utility. Here, we present LCTfound, a large-scale vision foundation model designed to overcome these limitations. Trained on a multi-center dataset comprising 105,184 CT scans, LCTfound leverages diffusion-based pretraining and joint encoding of imaging and clinical information to support 8 tasks, including CT enhancement, virtual computed tomography angiography (CTA), sparse-view reconstruction, lesion segmentation, diagnosis, prognosis, cancer pathological response prediction, and three-dimensional surgical navigation. In comprehensive multicenter evaluations, LCTfound consistently outperforms leading baseline models, delivering a unified, broadly deployable solution that both augments clinical decision-making and elevates CT image quality across diverse practice settings. LCTfound establishes a scalable foundation for next-generation clinical imaging intelligence, uniting large AI model with precision healthcare.

As an integral part of clinical imaging, lung computed tomography (CT) is crucial for doctors to examine the structure and function of key anatomical regions within the chest, including the heart, lungs, related system organs, including blood vessels, and the spine[1–5]. Annually, hundreds of millions of lung CTs are performed globally. Lung CT is indispensable for a multitude of clinical tasks, ranging from disease detection[6] and continuous monitoring of disease states to diagnostic evaluation of common and rare lung conditions, to preoperative strategic planning. Concurrently, the development, evolution, and application of medical AI systems for lung CT have substantially eased the workload in areas such as surgical navigation for lung operations, staging of lung cancer[7], and prognosis for lung

[1]Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing, China. [2]School of Information Science and Technology, Fudan University, Shanghai, China. [3]Department of Automation, Tsinghua University, Beijing, China. [4]Department of Thoracic Surgery, China State Key Laboratory of Respiratory Disease & National Clinical Research Center for Respiratory Disease, The First Affiliated Hospital of Guangzhou Medical University, Guangzhou, China. [5]Tsinghua Shenzhen International Graduate School, Tsinghua University, ShenZhen, China. [6]School of Software, Tsinghua University, Beijing, China. [7]Hangzhou Zhuoxi Institute of Brain and Intelligence, Hangzhou, China. [8]Tsinghua Medicine, Tsinghua University, Beijing, China. [9]Beijing Visual Science and Translational Eye Institute, Beijing Tsinghua Changgung Hospital, Tsinghua University, Beijing, China. [10]These authors contributed equally: Zebin Gao, Guoxun Zhang, Hengrui Liang. ✉e-mail: drjianxing.he@gmail.com; feng-xu@tsinghua.edu.cn; wongtienyin@tsinghua.edu.cn; yuchen.w.guo@gmail.com; qhdai@tsinghua.edu.cn

cancer[8,9], while also augmenting the imaging capabilities of lung CT systems.

However, the creation of newer and more sophisticated medical AI systems typically relies on massive and meticulously curated datasets that require laborious and time-intensive labeling that are tailored for specific clinical downstream tasks. The lack of doctors with domain knowledge in lung CT and the prolonged process of gathering sufficient data render the compilation of large, annotated datasets for individual lung CT scans and related medical tasks costly and challenging[10–12]. This is particularly true for rare lung diseases, in which obtaining enough clinical cases to train comprehensive AI models is impractical, leading to limitations in the performance of the currently trained AI lung CT models.

With breakthroughs in large language and vision models like ChatGPT[13–15], CLIP[16], SimCLR[17], and DINO[18], medical foundation models are now providing novel solutions to address these challenges. Foundation models have been developed in computational pathology[1,19], ophthalmic disease diagnosis[20,21], ultrasonography, X-Ray[22–24], and cancer biomarker innovation[25,26], improving diagnostic accuracy, facilitating knowledge sharing[27], and furthering medical education. Foundation models are trained via self-supervised learning (SSL) on large datasets without annotated labels. Foundation models are proficient at learning knowledge representations with the ability for zero-shot or few-shot learning, enabling these models to be fine-tuned for a multitude of downstream applications[28].

Despite rapid progress in medical imaging AI, vision foundation models tailored for diverse CT image processing tasks remain insufficiently explored and validated. For example, AI integrates the outcomes of lung CT examination with clinical symptoms, exposure history, and laboratory tests to quickly diagnose COVID-19 in patients[8,29,30]. Three-dimensional deep learning neural networks, leveraging low-dose lung CT imaging, enable lung cancer screening, with diagnostic accuracy that exceeds that of radiologists[7]. A modularized neural network has been engineered for reconstructing low-dose lung CT images, achieving a reconstruction speed that greatly outpaces commercialized iterative approaches, while maintaining the quality of the reconstruction[31]. Thus, foundation models in lung CT imaging can be expected to serve a dual purpose: offering clinical insights to provide diagnostic and therapeutic support and enhancing the imaging technology of CT scanners to improve practical operations.

In this study, we present LCTfound, a lung CT vision foundation model integrating images with correlated fundamental clinical information into the neural network, which is pre-trained by the diffusion policy on LungCT-28M (Fig. 1a) to address the dual purpose role of lung CT. The LungCT-28M dataset is an extensive collection tailored for lung CT, incorporating 105,184 lung CT scans from 5 centers, and comprising over 28 million images that encompass 14 distinct diseases (Fig. 1b). To our knowledge, this represents the largest lung CT dataset to date, covering the widest range of comprehensive disease states. By capitalizing on the LungCT-28M, we developed a feature encoder and decoder employing SSL through a denoising diffusion probabilistic model[32] (DDPM, Fig. 1c), which is adept at few-shot generalization across multiple downstream tasks leveraging the extracted features. We validated the clinical and technical significance of LCTfound across scanning-level to pixel-level downstream tasks. These tasks include the diagnosis of low-incidence diseases (e.g., mediastinal neoplasm and pulmonary alveolar proteinosis (PAP)), the prognosis prediction of non-small cell lung cancer (NSCLC), the prediction of major pathological response (MPR) to neoadjuvant chemoimmunotherapy, three-dimensional whole lung modeling virtual lung CTA imaging, the reconstruction from sparse views lung CT and the enhancement of low-dose lung CT (Fig. 1d). When benchmarked against conventional vision pre-trained models such as MAE[33] and MedSAM[34] in these tasks, LCTfound consistently outperformed these counterparts.

Results

Curation of LungCT-28M and rationale of LCTfound

We assembled a large national lung CT dataset, with 485,885 scans, from patients enrolled between 2009 and 2023 over 15 years (276,755 males and 209,130 females,) from five medical centers across China (Fig. 1a, Supplementary Fig. 1 and Supplementary Table 1). We refined this dataset based on the quality of images and the completeness of diagnostic information provided in the accompanying medical reports. The dataset refinement process consists of three main steps: completeness check of the reports, classification using an natural language processing (NLP) model, and image quality control (Supplementary Fig. 2). The NLP model was trained on manually annotated reports and used to filter data based on the confidence of the classification results (Supplementary Figs. 3–5 and Supplementary Table 2). This process culminated in the creation of the pre-training dataset, LungCT-28M (Fig. 1a, Methods), which encompasses 105,184 scans (59,935 from males and 45,249 from females). Excluding the cohort of 14,756 Lung CT scans with no evidence of disease, the final LungCT-28M dataset for training the model comprised CT scans for 14 common lung diseases, with the following distribution: Diffuse parenchymal lung diseases (DPLD, 49,133 scans), pulmonary calcification (PCal, 36,379 scans), Pulmonary bullae (PB, 19,029 scans), Lymph node enlargement (LNE, 16,967 scans), Pulmonary fibrosis (PF, 16,945 scans), bronchiectasis (BCH, 16,302 scans), Emphysema (13,961 scans), Pleural effusion (PE, 12,326 scans), Pulmonary cavity (PCav, 5544 scans), atelectasis (5169 scans), Pneumothorax (PTX, 2798 scans), Tracheal mass (TM, 1535 scans), Mediastinal mass (MM, 1444 scans), and Rib Fracture (RF, 649 scans) (Fig. 1b). The primary scanning protocol of LungCT-28M is chiefly guided by clinical application requirements, specifically choosing the window width and window level.

During the training phase, we employed a U-net architecture enhanced by the integration of transformer blocks[35], encompassing roughly 200 M trainable parameters (Supplementary Fig. 6). Paired clinical information (such as window width, window level, etc.) was randomly selected and, after Bert encoding, was coupled with the corresponding lung CT images and input into the main backbone network of LCTfound. The self-supervised pre-training LCTfound is grounded in the core principles of DDPMs: A two-dimensional lung CT image and its basic information were randomly selected from the LungCT-28M dataset and subjected to data augmentation. During the forward propagation, we progressively added Gaussian noise with a specific intensity into the image. After 1000 steps, this process resulted in the transformation of the image into pure noise (Fig. 1c). During the backward propagation, the neural network learned to perform denoising, thereby acquiring representation learning abilities (Methods). The basic information of the lung CT images was used as guidance, being input into the model simultaneously. To strengthen the robustness of the main backbone network to this basic information, we randomly occluded part or all of the basic information as input. For the inference stage, the pre-trained LCTfound was employed as a robust image encoder-decoder. For localization of mediastinal neoplasms and whole lung segmentation, we applied a trainable multilayer perceptron (MLP) during the fine-tuning process. This MLP served as a fusion layer for features integrated from the decoding path to obtain the final segmentation mask (Supplementary Fig. 7a). For tasks like PAP diagnosis and NSCLC prognosis prediction, we employed a trainable MLP to transform the bottom output features of the encoder into diagnostic labels (Supplementary Fig. 7b). For pixel-level tasks such as sparse-view lung CT reconstruction, LCTfound was used as a feature dictionary to guide the restoration of low-quality lung CT images (Supplementary Fig. 8a). For low-dose lung CT enhancement, LCTfound was fine-tuned on downstream task data through cold-diffusion manner[36] (Supplementary Fig. 8b).

https://doi.org/10.1038/s41467-025-66620-z

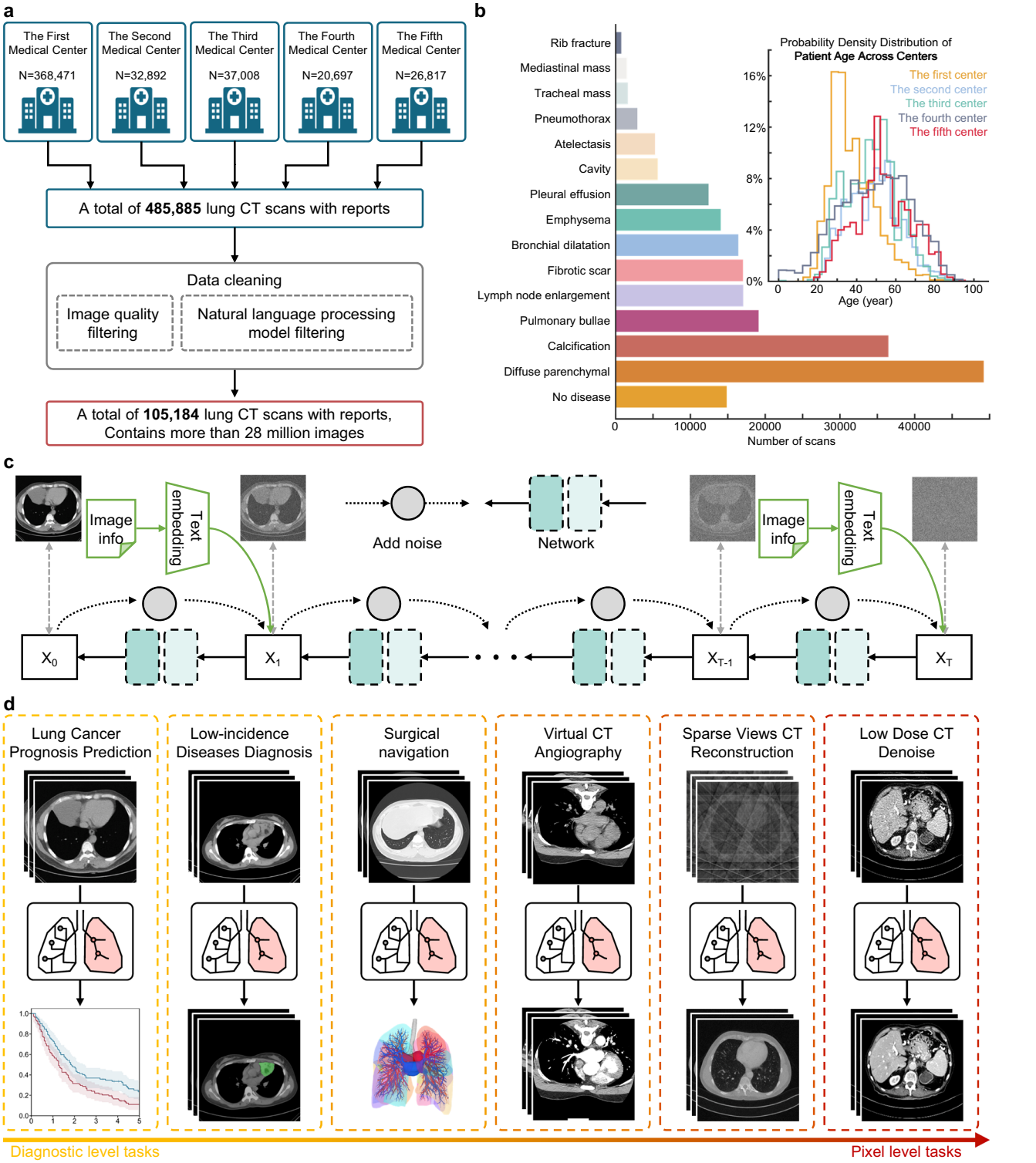


Fig. 1 | Profile of LCTfound. a Data preprocessing procedure. 485,885 lung CT scans and corresponding reports were collected from five medical centers. Then, they were screened and cleaned by the image quality and a natural language processing (NLP) model combined with reports (Methods), yielding the ultimate dataset LungCT-28M. LungCT-28M contains 105,184 lung CT scans, with a total of 28 million images. **b** Image compendium of LungCT-28M, a vast and varied pre-training dataset sampled over 100,000 lung CT stacks,encapsulating 14 different types of diseases. **c** The pre-training procedure for LCTfound. LCTfound incrementally introduces noise to the image, then learns the reverse denoising process based on a 200M-parameter U-net with an attention mechanism (Methods). Relevant information such as window width, window level, and types of diseases from lung CT images are embedded into the model as conditions guiding the image generation. To enhance the robustness of the model, this information will be randomly selected for embedding or not embedding. **d** Downstream task evaluation for LCTfound. The tasks for evaluation involve the diagnosis of low-incidence disease, the prognosis for NSCLC, the whole lung modeling for the surgical navigation, virtual CT angiography, the reconstruction of lung CT from sparse views, and the enhancement of low-dose CT. During the fine-tuning process, parameters that are frozen or modified are indicated in the panel. We used different adapters for different downstream tasks (Methods).

LCTfound facilitates disease diagnosis and prognosis evaluation

Disease identification and diagnosis based on medical imaging play a critical role in clinical practice, while imaging-driven prognosis prediction offers valuable guidance for treatment planning and decision-making. As a lung CT foundation model, LCTfound has demonstrated robust performance across multiple clinical tasks, including prognosis prediction for NSCLC, response prediction to neoadjuvant therapy, mediastinal tumor classification, and screening for PAP. Lung cancer remains the leading cause of cancer-related mortality globally, with NSCLC accounting for ~85% of cases[37,38]. The predominant subtypes of NSCLC are lung adenocarcinoma and lung squamous cell carcinoma[39]. The choice of treatment for NSCLC is multifaceted, tailored according to the stage of the disease, histopathological findings, genetic aberrations, and the overall health of patients[40]. A comprehensive treatment strategy typically includes surgery, radiation therapy, chemotherapy, immunotherapy, and molecularly targeted treatments[41]. For patients with localized NSCLC, early surgical intervention is available, yet the 5-year survival rate stands at a mere 59%. Prognostically valuable features extracted from lung CT scans are critically important for developing precise, non-invasive imaging-based NSCLC treatment protocols.

Consequently, we evaluated the efficacy of LCTfound on the prognostic imaging of NSCLC on the LUNG1 dataset (420 scans). We dedicated approximately half of the LUNG1 dataset[25] (220 scans) to fine-tune the model, and the other (200 scans), which remained unseen during training, was utilized for testing the model. The model incorporates both Lung CT images and tumor localization information as simultaneous inputs to improve prognostic accuracy. We conducted a comparative analysis of LCTfound, Foundation[25], Med3D[42], and MG[43] to investigate the performance disparities between models that were fully fine-tuned and those that only had their terminal linear classifier fine-tuned maintaining the feature extraction part frozen (Fig. 2a–h and Supplementary Fig. 9). The Kaplan–Meier survival estimates indicated that the fine-tuned LCTfound model provided superior stratification ($P < 0.001$), confirming its efficacy in accurately identifying mortality-based risk groups(Fig. 2a–h and Supplementary Fig. 10). By integrating the Grad-CAM, we generated saliency maps for LCTfound in the process of NSCLC prognosis prediction, revealing the attention of the model focuses on diseased regions which underscores its sensitivity to clinical problems (Supplementary Fig. 11). We also introduced perturbations to the input images to demonstrate the stability of LCTfound (Supplementary Fig. 12).

Meanwhile, we validated the proficiency of LCTfound in few-shot learning to prognosticate the MPR elicited by neoadjuvant chemoimmunotherapy in lung cancer[44–47]. We gathered a data cohort from four centers (the first affiliated hospital of Guangzhou Medical University, Liaoning Cancer Hospital & Institute, Shanghai Chest Hospital, and the first affiliated hospital of Xi'an Jiaotong University), composed of lung CT scans and the associated outcomes of MPR after treatment. We fine-tuned three pre-trained models (LCTfound, MAE[33], and RadImageNet[48]) using 876 images from 90 lung CT scans; the internal testing dataset included 74 images from 74 lung CT scans; the external testing dataset contained 82 images from 82 lung CT scans (Supplementary Fig. 13). LCTfound attained the highest area under the receiver operating characteristic curve (AUROC) on both the internal and external test sets, whether using lung window or mediastinal window CT images (Fig. 2i–p, Supplementary Figs. 14 and 15 and Supplementary Data 1).

The construction of AI models targeting uncommon lung diseases frequently encounters challenges stemming from the limited availability of the training data. Mediastinal neoplasms are classic low-incidence thoracic diseases with a variety of types, ranging from benign to malignant growth[49]. The global incidence of mediastinal neoplasms is ~0.77–1.68%, with about 60 million to 130 million patients worldwide[50]. The preceding decade has witnessed a steady increase in the incidence of mediastinal neoplasms, associated with a poor prognosis for patients with malignant forms. Currently, CT-based diagnosis of mediastinal neoplasms is fraught with substantial practical difficulties. The structural similarities and close proximity of mediastinal neoplasms to normal mediastinal anatomy frequently result in diagnostic ambiguity, thus complicating the comprehensive detection and precise localization of mediastinal neoplasms in regular CT imaging (Fig. 3a).

By focusing on the diagnosis of Mediastinal neoplasms, we demonstrated the few-shot learning ability of LCTfound, which is instrumental in the diagnosis of such low-incidence diseases. We standardized data collection from seven medical centers by the same protocol and collaborated with seasoned physicians for meticulous pixel-level tumor annotations (Supplementary Table 3 and Supplementary Fig. 16). We compared the LCTfound pre-trained on LungCT-28M with publicly accessible pre-trained models for lung CT: MedSAM[34], MAE[33], InternImage[51], Swin3D[52], Universal[53]. Additionally, we incorporated a U-net model that had been trained specifically on our assembled and annotated dataset of mediastinal neoplasms. The results from our investigation conclusively indicate that LCTfound surpasses its counterparts (Fig. 3b and Supplementary Data 1). Within the internal datasets, LCTfound achieved the Dice score apex of 0.7895, surpassing the subsequent leading model (MedSAM) by a margin of 5.08%. This trend of superiority was consistent across five external datasets (Fig. 3a). Overall, LCTfound has demonstrated unparalleled effectiveness in the segmentation of low-incidence mediastinal neoplasms, showcasing its adeptness in few-shot learning (Supplementary Movie 1). We visualized the saliency map of LCTfound using lung CT images with mediastinal neoplasms to highlight the significance of individual pixels in the segmentation process. The alignment between the visualization of saliency maps and clinical features underlines the efficacy of LCTfound in extracting clinically relevant information (Supplementary Fig. 17).

Furthermore, we verified the effectiveness of the pre-trained LCTfound in diagnosing PAP[54,55], leveraging a small dataset for fine-tuning. PAP, a rare lung disease, is a syndrome characterized by surfactant accumulation on alveoli and impaired alveolar macrophages, leading to insidious progressive respiratory difficulty, hypoxic respiratory failure, secondary infections, and pulmonary fibrosis[56]. Prevalence rates range from 3.7 to 40 cases per million (depending on the country/region), with an incidence estimated at 0.2 cases per million[57]. We compiled a dataset from Guangzhou First People's Hospital, comprising 270 cases of PAP-positive CT scan results (Supplementary Fig. 18). On the test dataset, we assessed the performance of LCTfound, MAE[33], and RadImageNet, all of which were fully pre-trained on the LungCT-28M dataset and subsequently fine-tuned on the training dataset. LCTfound achieved an AUROC of 0.9532, surpassing the second-place MAE by 6.01%. We halved the training dataset and fine-tuned the three pre-trained models. LCTfound maintained the lead with an AUROC of 0.9130 (Supplementary Data 1), outperforming the second-place MAE by 5.54%. This demonstrates the efficacy of the pre-trained LCTfound in diagnosing rare lung diseases through few-shot learning (Supplementary Fig. 19). The saliency map of the PAP image reveals that the fine-tuned LCTfound successfully extracted the clinical features of PAP, such as the accumulation of anomalous substances in the alveoli (Supplementary Fig. 20).

LCTfound enhances 3D modeling and AI-assisted surgical navigation

Building upon imaging-based diagnostic capabilities, intraoperative support has become increasingly important in the clinical management of lung diseases. For conditions such as lung cancer and pulmonary nodules, surgical resection is often a critical step following diagnosis. In this context, intraoperative 3D modeling and navigation based on lung anatomical structures offer essential guidance for

https://doi.org/10.1038/s41467-025-66620-z

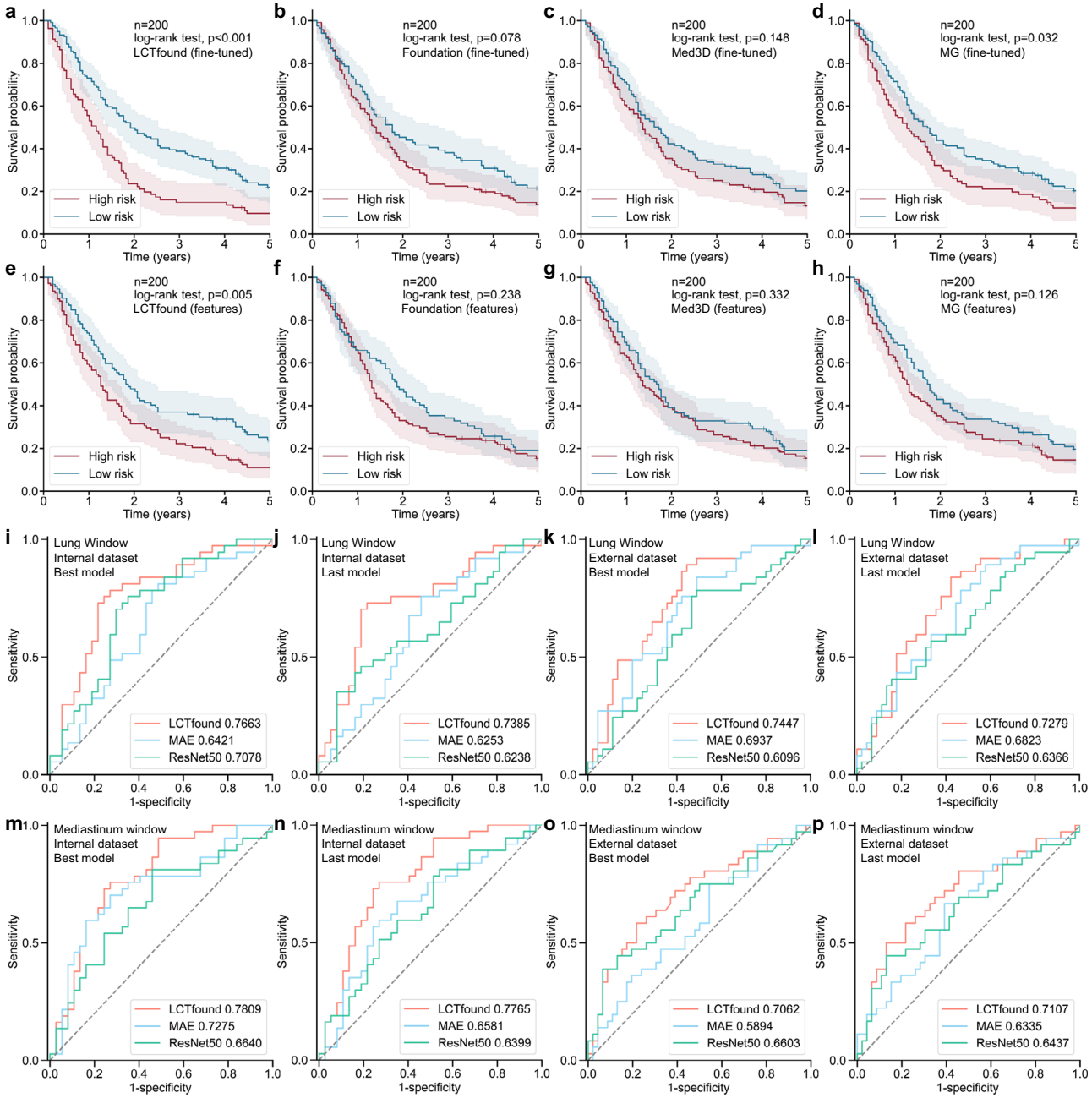


Fig. 2 | Performance of LCTfound on the prediction of the NSCLC prognosis and the pathological response after treatment. a–h The Kaplan–Meier survival curves on LUNG1 test datasets are exhibited for the groups stratified according to the best model predictions out of the eight methods. The midpoint of the error bands represents the Kaplan–Meier estimated survival function. The shaded area represents the 95% confidence interval. a-d represent the outcomes of full parameter fine-tuning for the four foundation models. **e–h** The results of employing fixed foundation model parameters for feature extraction, and use extracted features to train a linear classifier for prediction. **i–p** Performance of different models in predicting the major pathological response to neoadjuvant chemoimmunotherapy. **i** Comparison of the results using best-trained parameters of three models when using the lung window on the internal testing dataset. **j** Comparison of results using the last-trained parameters of three models, with the rest of the conditions mirroring those in (**i**). **k** Comparison of results on the external testing dataset, with all other conditions the same as in (**i**). **l** Comparison of results on the external testing dataset, keeping all other testing conditions consistent with (**j**). **m–p** Comparison of results using the mediastinal window to prediction, with all other conditions the same as in (**i–l**).

surgical planning and precise execution. Whole-lung 3D modeling enables the digital transformation of CT imaging data into anatomically accurate representations, which play a pivotal role not only in routine clinical workflows but also in medical research. These digital lung models are invaluable for preoperative planning[58], intraoperative surgical navigation[59], and the formation of postoperative treatment protocols[60]. Furthermore, with the advancement of medical AI, these digitized structures provide foundational data for various support systems and are instrumental in detailed longitudinal disease analyses, including tumor progression. The clinical applications of whole lung 3D modeling are vast and present numerous prospects for further investigation and application (Fig. 4a).

We delineated 21 anatomical structures in the lung through 3D modeling, encompassing the bronchi, arterial and venous networks, and specific segments such as the left apicoposterior, left anterior, inferior and superior lingula, multiple basal and dorsal segments, as

https://doi.org/10.1038/s41467-025-66620-z

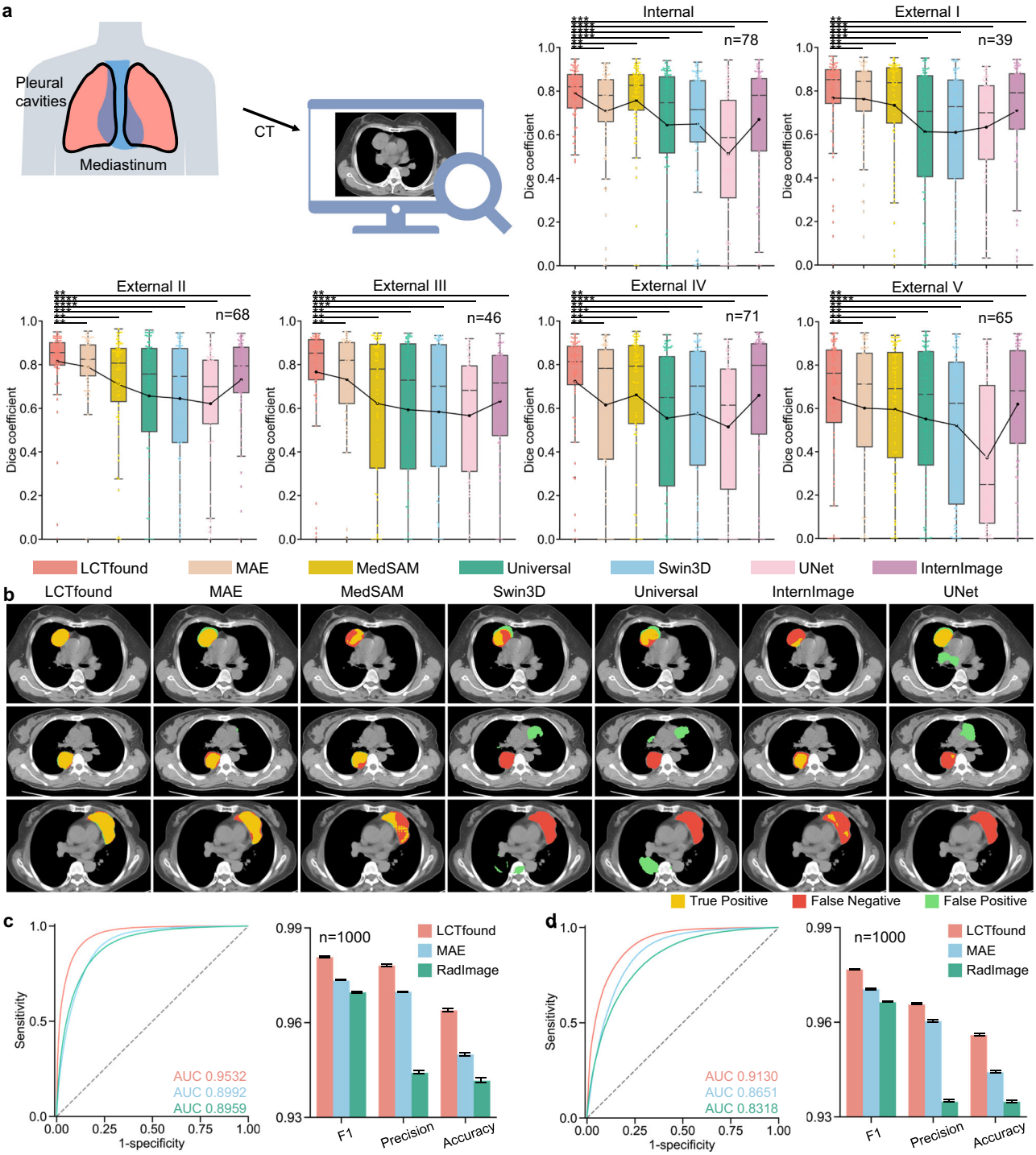


Fig. 3 | Evaluating the performance for diagnosis of low-incidence lung diseases. a Validation results of mediastinal neoplasms segmentation on internal and five external datasets. We demonstrated the accuracy of segmentation for mediastinal neoplasms with self-supervised pre-trained learning strategies, including MedSAM, MAE, InternImage, Swin3D, Universal, UNet, and LCTfound. The *P*-value is computed through two-sided *t*-tests and presented in the figure. Statistical significance is indicated as follows: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$, $^{****}p < 0.0001$; results without asterisks are not statistically significant ($p \geq 0.05$). In external test set 1, LCTfound exceeded the runner-up MAE by 0.25%; in external test set 2, LCTfound surpassed the second-best MAE by 1.58%; in external test set 3, LCTfound outperformed the second-place Universal by 2.30%; in external test set 4, LCTfound surpassed the second-place MAE by 7.90%; in external test set 5, LCTfound outpaced the second-place MAE by 3.64%. **b** Visualize the typical examples of the mediastinal neoplasms segmentation by the aforementioned methods. Red indicates tumor tissue. Fine-tuned LCTfound achieves more accurate segmentation of the mediastinal neoplasm regions. **c**, **d** Compare the diagnostic results of LCTfound, MAE, and RadImageNet on the PAP after fine-tuning with a small dataset (train = 297, validation = 2700). **c** The ROC curve of the three pre-trained models fine-tuned using the full training set ($n = 297$, positive case = 27). **d** The receiver operating characteristic (ROC) curve of the three pre-trained models fine-tuned using 50% of the training set ($n = 148$, positive case = 15). Solid bars on top of each column represent the standard deviation, the midpoint of each bar represents the mean calculated from 1000 bootstrap iterations. In the box plots, each point represents the score of an individual sample. The box indicates the interquartile range (IQR), with the middle line marking the median and the black dot representing the mean. Whiskers extend to 1.5 × IQR, and any points beyond this range are plotted as individual outliers.

https://doi.org/10.1038/s41467-025-66620-z

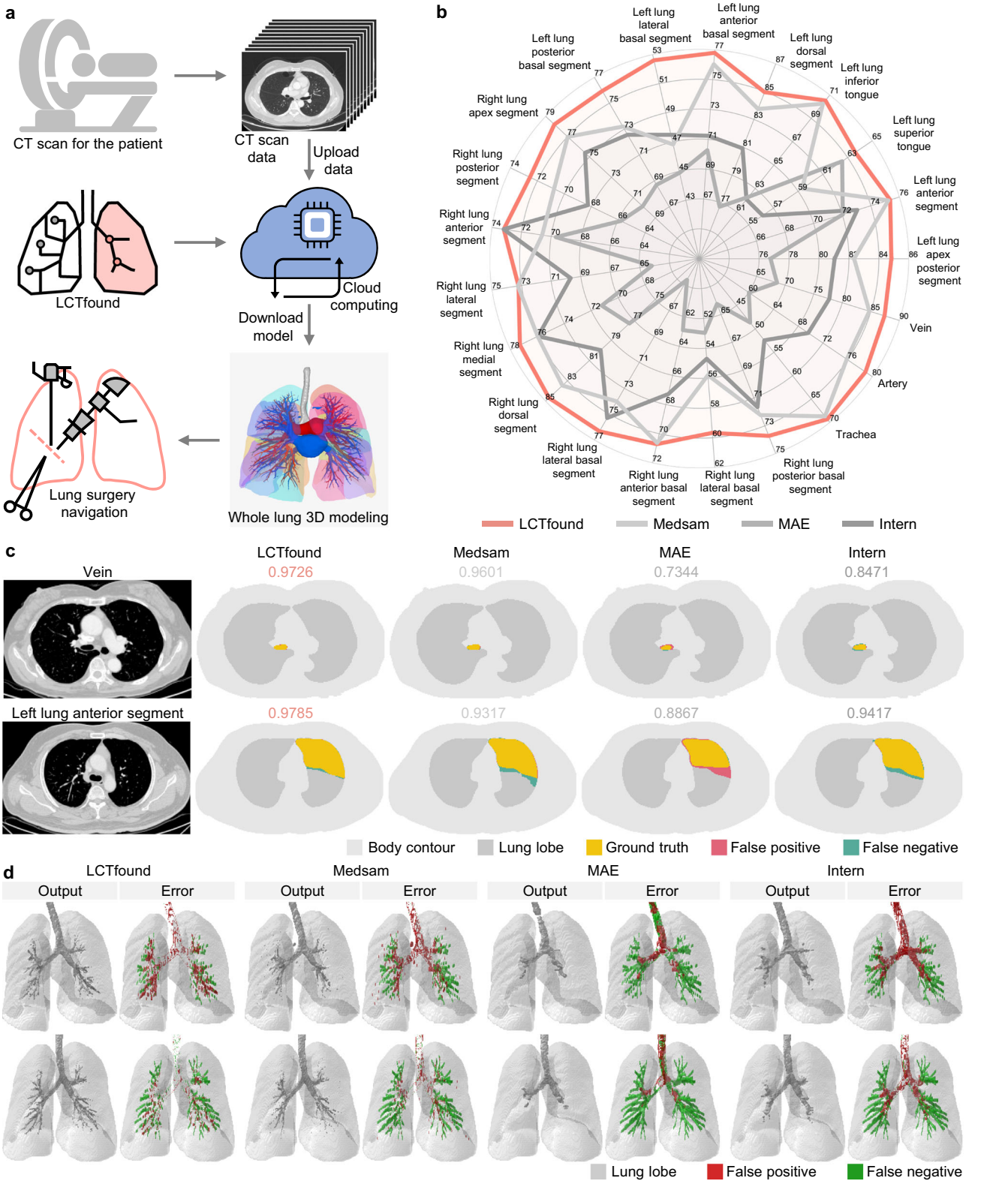
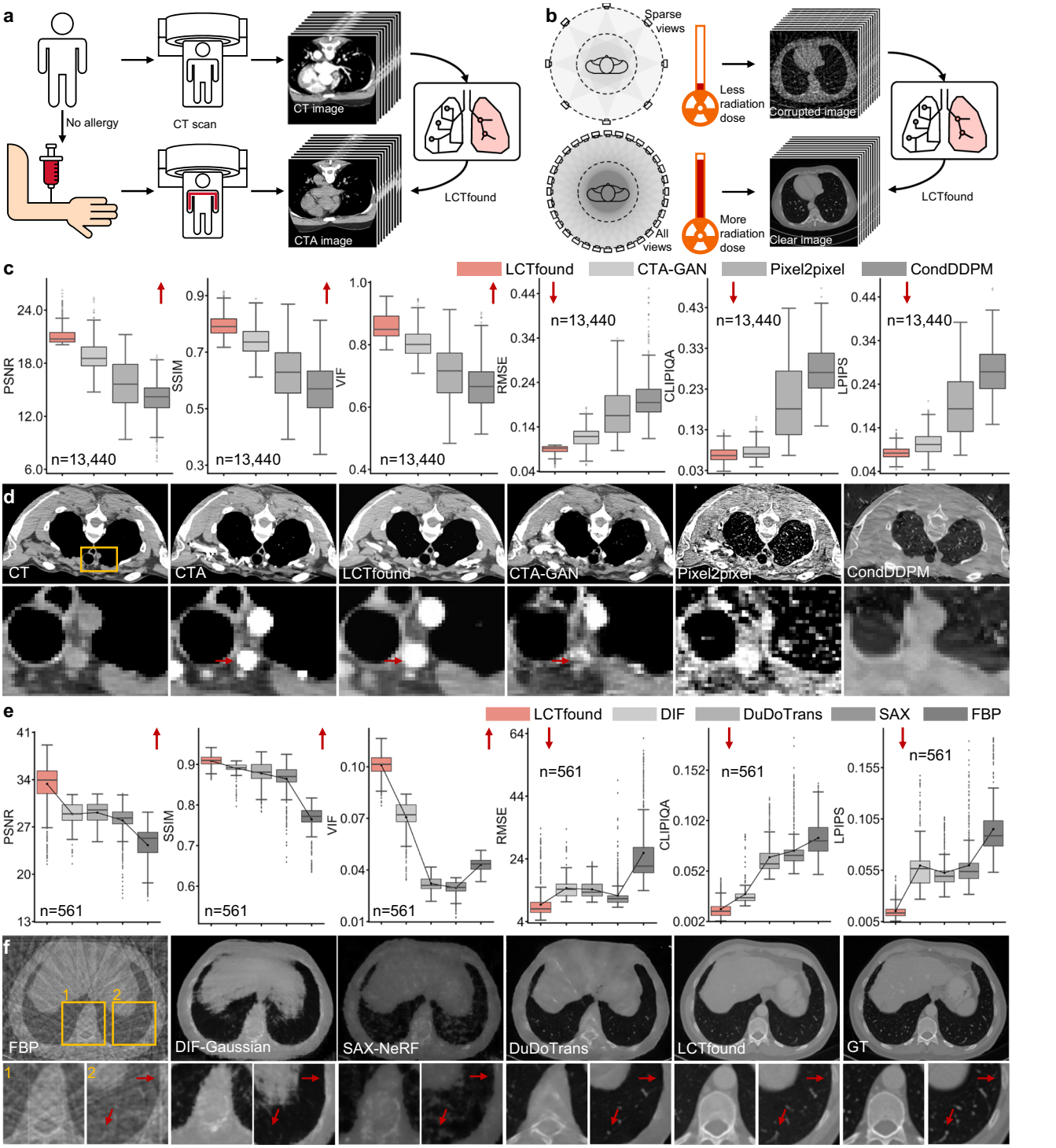


Fig. 4 | Comparison of lung three-dimensional modeling related to minimally invasive surgical navigation. a Once patients undergo lung CT scans, the data is uploaded to the cloud. LCTfound then completes three-dimensional modeling of 21 anatomical structures of the entire lung, which is used for navigational guidance in minimally invasive lung surgeries. **b** Detailed few-shot performance of lung three-dimensional modeling of 21 anatomical structures. LCTfound outperforms the current leading lung CT foundation models in segmentation results, with its maximum being over 10% higher than that of the second rank ($n = 20$). **c** 2D visualization of the segmentation results of four different methods. The first row is the segmentation result of the vein, the second row is the segmentation result of the Left lung anterior segment, and the third row is the segmentation result of the Left lung upper lobe. From left to right are: the original lung CT image, the segmentation result of LCTfound, the segmentation result of Medsam, the segmentation result of MAE, and the segmentation result of InternImage. Dice scores associated with the segmentations are annotated above the images. Different colors in the images represent true positive, false positive, and false negative. **d** 3D visualization of the tracheal segmentation outcome. Red indicates false positives, green indicates false negative, and gray indicates lung nodes.

https://doi.org/10.1038/s41467-025-66620-z



well as the right lung's posterior, anterior, apical, lateral, and medial segments (Fig. 4a). We constructed a dataset consisting of 35 lung CT scans, with ground truths manually annotated by experienced radiologists. This dataset was partitioned into subsets: 10 scans (comprising 2959 images) for training, 5 scans (1399 images) for validation, and 20 scans (6596 images) for testing. We conducted comparisons of whole lung segmentation between LCTfound and MedSAM[34], MAE[33], and InternImage[51]. In all 21 semantic segmentation tasks, LCTfound achieved the best segmentation performance, followed by MedSAM in second place for most tasks (Fig. 4b, Supplementary Fig. 21 and Supplementary Data 1). Our two-dimensional visualization of the segmentation results for 21 kinds of whole-lung anatomical structures across four methods (Fig. 4c, Supplementary Figs. 22, and 23), reveals that LCTfound consistently generated fewer false positives and false negatives. This is evident in both bulky structures (e.g., the Left lung anterior segment) and more intricate structures (e.g., the trachea, arteries, and veins). The three-dimensional visualization of the whole lung segmentation provides an in-depth, intuitive comparison of the four methods (Fig. 4d), highlighting the precise segmentation capability of LCTfound as a critical tool for lung surgery navigation (Supplementary Fig. 24). In clinical lung nodule resection surgeries, the utilization of whole lung segmentation and reconstruction by LCTfound has significantly simplified the process of precise nodule localization and removal (Supplementary Movies 2 and 3). The

Fig. 5 | performance of LCTfound on pixel level tasks. a Differences in clinical processes for acquiring non-contrast lung CT images and lung CTA images. Lung CTA involves administering iodinated contrast agents intravenously and using CT for rapid imaging, requiring time for contrast agents to circulate in the body of patient. **b** Conventional lung CT scans generally necessitate reconstructing images from hundreds to thousands of views. Sparse-view CT imaging requires only a few angles, thereby substantially reducing the radiation exposure to patients. **c** Evaluation of virtual CTA results for LCTfound and three other methods. The metrics, from left to right, are sequentially: PSNR, SSIM, VIF, RMSE, CLIPIQA, LPIPS. For the first three metrics, their numerical values have a positive correlation with the quality of the image; for the latter three, the correlation is negative. **d** A case of virtual CTA results. Sequentially from left to right, the non-contrast lung CT image, the paired lung CTA image, the outcome of LCTfound, the outcome of CTA-GAN, the outcome of Pixel2pixel, and the outcome of CondDDPM. The red arrow indicates the virtual CTA imaging result of a tiny blood vessel. **e** Assessing of few-shot learning results for sparse-view reconstruction of lung CT. With 16 projection views, the pre-trained LCTfound obtained improved reconstruction outcomes. **f** Two cases of Lung CT for 16 projection views. The sequence from left to right includes the FBP result, the result from DIF-Gaussian, the result from SAX-NeRF, the result from DuDoTrans, the result from LCTfound, and the ground truth image. The images in the second row are enlargements of the areas within the yellow boxes in the first row of images. Fewer artifacts are evidently present at the site pointed by the yellow arrow in the LCTfound results. Statistical comparisons were performed using the two-sided Wilcoxon signed-rank test. Statistical significance is indicated as follows: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$, $^{****}p < 0.0001$; results without asterisks are not statistically significant ($p \geq 0.05$). In the box plots, each point represents the score of an individual sample. The box indicates the interquartile range (IQR), with the middle line marking the median and the black dot representing the mean. Whiskers extend to 1.5 × IQR, and any points beyond this range are plotted as individual outliers.

LCTfound model, aimed at lung surgery navigation, has been deployed in the cloud and is paired with a fully interactive website (https://demo.lctfound.com/chest3d/records/97?access_token=93217b5d616b47218ea65aec83fc472f). By leveraging cloud computing resources, we aspire for LCTfound to serve as a practical tool in lung surgery (Supplementary Fig. 25).

LCTfound supports image enhancement across diverse clinical scenarios

As a diagnostic modality involving ionizing radiation, CT imaging carries inherent risks to patients. Therefore, enhancing CT acquisition and reconstruction processes holds significant value for reducing radiation exposure and improving diagnostic efficiency. As a generative model, LCTfound demonstrates strong capabilities in supporting such improvements. We validated its performance across multiple image-centric tasks, including virtual CTA synthesis, denoising, sparse-view reconstruction, and super-resolution. Building upon these results, we further trained segmentation models using the reconstructed images to evaluate their impact on downstream clinical tasks. Moreover, we leveraged LCTfound's generative capacity to synthesize large-scale imaging datasets for model distillation, highlighting its potential utility in data augmentation and sharing scenarios.

AI-enhanced CT imaging strives to mitigate the negative impacts of CT scans on patients (Fig. 5a, b). Given the potential harm and carcinogenic potential associated with X-ray exposure, the adoption of low-dose CT protocols has become critical in clinical settings, especially for widespread disease screening through CT[7,61]. The minimization of CT scan radiation is attainable primarily through two strategies: the reduction of the X-ray tube current (or voltage) and the curtailment of the number of scanning angles, with the latter also facilitating expedited scanning procedures (Fig. 5b). However, CT images reconstructed by these methods are plagued with considerable noise and artifacts, hence the elimination of such noise and artifacts is a crucial challenge in low-dose CT imaging[62]. Among numerous clinical applications, CTA stands as a non-invasive vascular imaging method extensively utilized for the diagnosis of vascular abnormalities, such as aneurysms and dissections. However, CTA necessitates the intravascular administration of iodinated contrast agents (ICA), which is both expensive and time-consuming and presents risks to patients with iodine allergies, renal insufficiency, or multiple myeloma (Fig. 5a). Thus, devising a technique to circumvent ICA and derive CTA-like images from non-contrast CT images could lower the adverse effects on patients. As LCTfound is engineered with the pixel-level pre-training strategy, it is naturally adept at pixel-level tasks (Supplementary Figs. 26–33). Accordingly, we validated the effectiveness of LCTfound for virtual CTA imaging, sparse-view reconstruction and lung CT low-dose enhancement tasks.

For virtual CTA imaging, we collected paired lung CT and CTA data from 102 patients at the First Affiliated Hospital of Guangzhou Medical University, with 17 pairs used as the training set (2720 paired images), and 85 pairs used as the test set (13,440 paired images). In the context of few-shot learning, we compared the virtual CTA imaging performance of LCTfound, CTA-GAN[63], Pixel2pixel[64], and Conddiff[65]. For the metrics that are positively correlated with image quality (Fig. 5c), LCTfound achieved the highest scores: LCTfound achieved superior performance across multiple image quality metrics, including a PSNR of 21.14, structural similarity (SSIM) of 0.794, and visual information fidelity (VIF) of 0.860, outperforming the second-best method by 12.9%, 7.3%, and 6.7%, respectively. For metrics inversely related to image quality (Fig. 5c), LCTfound consistently obtained the lowest error scores, with a 33.5% lower root mean square error (RMSE), 9.9% lower CLIPIQA, and 26.1% lower VIF. These results highlight LCTfound's strong few-shot learning capability at the pixel level, enabling high-fidelity virtual lung CTA synthesis from limited data.

For the sparse view reconstruction in lung CT imaging, our training dataset consisted of 9 lung CT scans (5377 images), and our test dataset included 1 lung CT scan (561 images). The Radon transform simulation was utilized to generate the acquisition results of the unreconstructed lung CT signal, which are then used for training or fine-tuning the model. We conducted comparisons between the LCTfound with pre-training, DIF[66], DuDoTrans[67], SAX[68], and filtered back projection (FBP). In the 16-views reconstruction of lung CT images, the LCTfound with pre-training yielded superior results (Fig. 5c and Supplementary Data 1). It achieved a PSNR of 33.38 and learned perceptual image patch similarity (LPIPS) of 0.0153, significantly surpassing DuDoTrans (PSNR 29.12, LPIPS 0.0525). LCTfound reported a lower CLIPIQA (0.0142 vs. 0.0292), and a higher VIF (0.1010 vs. 0.0706) and SSIM (0.909 vs 0.891) compared to the DIF model. In terms of error metrics, LCTfound exhibited a lower RMSE (46.94 vs. 61.29) than FBP, further highlighting its superior reconstruction fidelity. ßPre-training eradicated the artifacts that are easily generated in pixel-level tasks (Fig. 5d). We arrived at comparable conclusions for the reconstruction tasks at 8 and 32 views (Supplementary Fig. 28). We also conducted experiments on a larger public dataset, LUNA. Specifically, we randomly selected 16 cases (comprising 4032 images) from the official subset0 for testing, while the remaining 72 cases (comprising 18,884 images) were used for training. The results consistently demonstrated that LCTfound outperformed other methods across all evaluation metrics (Supplementary Fig. 29).

Regarding the enhancement of low-dose lung CT image quality, we utilized the Mayo 2016 dataset, which contains 10 pairs of matched low-dose and full-dose lung CT scans, totaling 5916 images. We designated a single scan (524 images) for the training dataset and the other nine scans (5392 images) for the testing dataset. We conducted comparisons between the LCTfound with pre-training, WGAN[69], and

DualGAN[70] about their performance on the low-dose CT image enhancement task (5% dose). To rigorously evaluate the enhancement outcomes, we employed six established metrics from the field of computer vision: PSNR, SSIM, VIF, RMSE, CLIPIQA[71], and LPIPS[72] (Methods). Across all six metrics, LCTfound achieved the best scores (Supplementary Fig. 30).

To evaluate the impact of reconstructed images on downstream clinical tasks, we trained and tested mediastinal tumor segmentation models using three types of input data: (1) sparse-view CT reconstructed with FBP, (2) sparse-view CT reconstructed with LCTfound, and (3) original full-view CT scans. Results showed that models trained on LCTfound-reconstructed images achieved performance closely matching those trained on original CT data across the validation set and multiple test sets (Supplementary Fig. 31). We also conducted additional experiments to evaluate the zero-shot capability of LCTfound, specifically on lung CT image super-resolution tasks with scaling factors of 4× and 8×. The results demonstrate that LCTfound generates higher-quality super-resolved CT images compared to conventional baselines (Supplementary Figs. 32 and 33).

Discussion

In this study, we collated a large multicenter dataset of over 100 million lung CT images from 485,885 lung CT scans, and by combining diagnostic reports and image quality, we established the LungCT-28M pre-training dataset which includes 105,184 lung CT scans (over 20 million images) covering the common 14 lung diseases as well as healthy lungs. Using a self-supervised training strategy, we developed a vision foundation model for lung CT, LCTfound, and validated this model with multiple high-level and low-level clinical tasks. Our goal is to develop a foundation model capable of supporting a wider spectrum of tasks. Leveraging diffusion-based pre-training, we construct LCTfound, which not only supports disease diagnosis and lesion localization but also generalizes to a broader set of CT imaging tasks. By open-sourcing its model weights, we empower the research community to leverage its generative capabilities for diverse downstream applications and scientific exploration. We demonstrated the following outcomes. First, to assess the efficacy of LCTfound in diagnosing uncommon lung diseases, we showed that LCTfound outperformed other pre-training models, achieving the highest accuracy in both localizing mediastinal neoplasms and diagnosing PAP. The diagnosis of uncommon diseases is clinically significant, and these experiments highlight LCTfound's potential for generalization in tasks related to diagnosing such conditions. Second, we showed that LCTfound improves neoadjuvant response and NSCLC prognostication prediction. We further validated LCTfound on additional disease types and tasks. On the LUNA dataset, the model outperformed state-of-the-art baselines such as nnU-Net in lung nodule segmentation. For the COVIDx CT dataset, it achieved an AUC of over 99% using a limited training set of ~3000 images and evaluated on a much larger test set of more than 30,000 images (Supplementary Figs. 34 and 35). These results provide further evidence of the model's generalizability across diverse diseases and clinical tasks. Third, as an important technique for surgical navigation, we showed that LCTfound has superior performance for whole lung segmentation. Fourth, LCTfound demonstrated significant improvements in lung CT imaging tasks, including virtual lung CTA imaging, optimizing reconstructions from sparse-view CT scans, and denoising low-dose lung CT images to enhance overall image quality. These various tasks that LCTfound could perform are useful in many clinical settings and disease conditions.

Compared with other vision-centric encoders, we showed that LCTfound pre-trained with the DDPM strategy could be employed for a wider variety of clinical tasks related to lung CT. LCTfound was more adept at pixel-level tasks like CT image denoising and reconstruction than pre-trained models such as MAE[33] and MedSAM[34], enabling ordinary lung CT imaging machines to attain superior imaging capabilities. LCTfound goes beyond the capabilities of GANs by infusing pixel-level tasks with richer image details, effectively enhancing lung CT images regardless of their initial quality. Its DDPM-driven training enables it to extract key semantic elements, capable of delineating the lung anatomy even from images plagued with substantial noise. This ensures that in high-level semantic tasks such as lung disease diagnosis or localization, LCTfound also achieved superior results. Especially in the task of full lung segmentation, whether it is for anatomical structures with many details such as Trachea, Vein, and artery, or for other mass structures, LCTfound demonstrates consistent performance. Furthermore, the special strategy of DDPM has expanded the diversity of the model. With the input of different time steps, different time embeddings endow LCTfound with the ability to extract features at various levels. We are convinced that LCTfound will revolutionize the foundation vision-based models for lung CT images and advance the progress of large models in medicine.

As a vision foundation model, pre-trained on over 28 million lung CT images, LCTfound can perform a wide array of clinical applications, particularly in scenarios with limited data availability. In developing data-driven AI approaches, acquiring actual clinical data typically involves addressing the issue of patient privacy[73]; otherwise, the data cannot be exported from the hospital. With LCTfound, it is possible to generate synthetic lung CT images in both lung window (window width: 1500, window level: −500) and mediastinal window (window width: 400, window level: 0) settings, aiding AI model training and validation. Designing various adapters based on LCTfound enables fine-tuning of the model with multicenter hospital data, which allows the data to remain within the hospital, but the model to be shared among multiple medical centers, achieving effective federated learning. LCTfound is a model pre-trained based on the DDPM strategy, thus it possesses inherent advantages in low-level tasks, such as converting lung CT images into CT pulmonary angiography images. CT pulmonary angiography[74], which images the pulmonary arteries, is a crucial tool for diagnosing pulmonary embolism, but it requires the traumatic insertion of an intravenous catheter into the patient[75]. Developing a reliable method for virtually transforming from ordinary lung CT scans to CT pulmonary angiography images would be beneficial in alleviating patient discomfort and advancing large-scale disease screening. From the perspective of data compression, we posit that the adequately pre-trained LCTfound embodies extensive information from numerous lung CT images, facilitating the substantial compression of large datasets that were previously cumbersome to duplicate and disseminate. In essence, LCTfound encapsulates a vast repository of lung CT scans into a fundamental model, ready to tackle diverse tasks in lung CT imaging and position itself as the intelligent brain of lung CT imaging.

Our study had several limitations. For example, for practical clinical applications, the operational speed of the model is crucial for doctors, particularly in the case of intraoperative lung CT imaging, which may demand the real-time processing of captured data. The contradiction to the demand for real-time processing is the limited computing resources of hospitals, hence, undertaking a series of AI acceleration techniques like pruning[76], distillation[77], and quantization[78] on LCTfound is the forthcoming task. Another major limitation of current work lies in the insufficient utilization of textual information for tasks like report generation. Medical clinical issues are inherently multimodal. In addition to lung CT scans, there exist complementary data sources including diagnostic reports, electronic medical records[79] and biochemical indicators. Patients who undergo regular medical examinations and follow-ups can also contribute longitudinal multimodal data[80]. While LCTfound serves as a robust vision foundation model for lung CT imaging, it currently focuses on visual representation learning and does not yet incorporate multimodal information for joint reasoning or diagnosis. Fully exploiting textual resources would require integrating large language models capable of understanding

and reasoning over multimodal data[81]. Integrating multimodal information is crucial for developing clinically relevant diagnostic tools, as demonstrated by recent works such as MIMO[82] and MedPLIB[83]. We acknowledge this as a major limitation and are actively extending our model toward a vision-language foundation model that aligns CT imaging with textual supervision from reports. Such integration is expected to further enhance clinical interpretability, enable report generation, and bridge the gap between imaging and decision-making. As there is a diversity in the *z*-axis slice thickness within the ChestCT-100K, the current LCTfound is predominantly pre-trained on 2D lung CT images. Given sufficient computational resources, we believe that a lung CT foundation model based on 3D images will exhibit enhanced diagnostic potential (Supplementary Fig. 36), especially for small structures that are more continuous along the *z*-axis, such as lung nodules. Future efforts will be dedicated to leveraging LCTfound as a foundational component for constructing multimodal lung CT AI models, with an ongoing commitment to enhance its applicability in pulmonary medicine[84].

Methods

Inclusion and ethics

This study was conducted using retrospective clinical data and complies with all relevant ethical regulations. The study was approved by the Ethics Committee of the National Center for Respiratory Medicine/ The First Affiliated Hospital of Guangzhou Medical University. The need for informed consent was waived due to the retrospective nature of the study. The study protocol was reviewed and approved internally by the First Affiliated Hospital of Guangzhou Medical University. This study follows the TRIPOD+AI reporting guideline for the development and validation of artificial intelligence models in medical research. The completed TRIPOD+AI checklist is provided in the Supplementary Information.

Network architecture of LCTfound

The main architecture of LCTfound is a CrossAttention-based U-shape network (Supplementary Fig. 5). LCTfound employs an innovative image self-supervised approach, specifically utilizing diffusion models for pre-training to obtain robust feature representations. Diffusion models are capable of generating realistic and diverse images, and numerous studies have demonstrated that well-trained diffusion models contain rich prior knowledge, which can significantly aid in small sample tasks[85]. The generative capability of diffusion models broadens their range of applications. Compared to models such as DINO and MAE, Diffusion models show strong potential for low-level tasks in CT imaging, including enhancement, segmentation, and lesion localization. The data generated by diffusion models can be regarded as an indirect form of dataset sharing (Supplementary Fig. 37).

To better acquire feature representations across multiple levels, we opted for a pixel-to-pixel space diffusion model. For LCTfound, we designed a UNet model architecture enhanced with a cross-attention mechanism, enabling the model to generate images with supplementary textual information. Specifically, LCTfound features five downsampling modules, five upsampling modules, and one deep feature extraction module. Each of these modules incorporates residual structures to ensure effective gradient optimization during training. Moreover, certain downsampling and upsampling modules are enhanced with a cross-attention Transformer module, further improving the model's ability to capture detailed and contextually relevant features. This mechanism enables the encoded textual information to directly influence the image generation process. Such textual information includes, but is not limited to, parameters like the image's window width and window level, as well as disease category information extracted from reports using NLP models. In visual tasks, convolutional neural networks (CNNs) have good efficiency, speed, and generalization capabilities. However, owing to their inherent inductive biases, they struggle to effectively capture global contextual information. Incorporating Transformer-based attention modules addresses this limitation by enabling the model to achieve superior global feature capture[86]. This hybrid architecture leverages the strengths of both CNNs and Transformers, resulting in a more robust model. Additionally, the upsampling, downsampling, and deep feature extraction modules not only accept image features as inputs but also receive encoded time step information. This integration allows the model to better estimate the noise present at each point in the diffusion process, thereby improving the overall accuracy and reliability of the extracted feature maps. A more detailed description of the structural information, including the specific designs and implementation details, will be provided in the supplementary methods section.

Data cleaning of LungCT-28M

In the development of LCTfound, a massive lung CT imaging dataset was initially compiled. This dataset, named LungCT-100M, includes 485,885 instances of lung CT scans gathered from the First Affiliated Hospital of Guangzhou Medical University, with the data ranging from the year 2009 to 2023 and encompassing patient ages from 2 to 88 years. The LungCT-100M was refined by integrating image quality with diagnostic details from associated reports to ensure the high quality of pre-training lung CT scan data.

First, we focused on lung CT scan data that had complete diagnosis reports. A total of 263,483 scans were retained due to the availability of corresponding diagnostic reports. Then, we employed a trained NLP model to categorize the lung diagnostic reports, targeting 14 significant lung diseases as the main classification results. We only retained lung CT scans whose prediction results by the NLP model with a confidence level higher than 0.9. Subsequently, we screened out images that were corrupted due to improper storage. Ultimately, we discarded images with substantial noise. Specifically, to filter out noisy images, we applied the Canny edge detector and calculated the ratio of pixels containing edge information to the total number of pixels. If this ratio exceeded 0.2, the image was included in the candidate set. From this set, images with acceptable quality were further selected through manual review, while the remaining images were discarded (Supplementary Fig. 1). As a final result, we acquired a dataset with 105,184 lung CT scans (more than 20 million images), designated as LungCT-28M.

The training for the NLP classifier was conducted with image reports annotated manually. It takes diagnostic reports as input and yields disease categories as output. The training of the NLP model was segmented into three phases. This NLP model comprises three modules: an extractor, a relation model, and a classifier head. The extractor module (Supplementary Fig. 3) primarily extracts the diagnosis results, positions, and statuses from the diagnostic reports. The primary function of the relation model (Supplementary Fig. 4) is to determine whether there is a diagnosis-related semantic relationship between two identified entities in the medical text. The primary role of the pure model (Supplementary Fig. 5) is to categorize diagnostic entities within medical texts, outputting their corresponding disease categories (e.g., emphysema, pulmonary tuberculosis, among 15 categories). The NLP model training process is as follows: For the first phase, we manually annotated 2000 reports along with their corresponding categories. Of these, 1000 reports were used as the training dataset, and 1000 reports were used as the final test set to evaluate the model's performance. Subsequently, we applied the trained model to categorize the rest of the unannotated reports. The 2000 reports with the lowest confidence scores were chosen for another round of manual annotation before being incorporated back into the training set for further model training. This entire process was conducted three times.

Pre-training of LCTfound

To adapt to a wider range of downstream tasks, we employed diffusion models for pretraining. Generally, diffusion models are capable of

generating realistic images from Gaussian noise. Recent studies have demonstrated that diffusion models can learn stable prior knowledge, which improves performance across various downstream tasks. However, there has been no prior work leveraging diffusion models for foundation model with CT images. In our work, we used the diffusion model as a self-supervised pretraining task. Prior to training, the images were preprocessed by converting them into lung window and mediastinal window formats. For the diffusion process, we adhered to the standard strategy settings[32]. The training process of diffusion models involves two primary phases: the forward diffusion process and the reverse diffusion process. In the forward diffusion process, noise is progressively added to the data. Let x_0 denote the original data sample, and x_t represent the data state at time step t. The forward diffusion process can be mathematically expressed as $x_t = \sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}\epsilon_t$. Here, α_t is a time-dependent variance parameter, and ϵ_t denotes noise sampled from a standard Gaussian distribution. The goal of DDPM is to train a model to recover the original data from noisy observations. This is achieved by learning the noise distribution through the minimization of the following loss function: $Loss = \mathrm{E}_{x_0,\epsilon_t,t}[||\epsilon_t - \epsilon_\theta(x_t,t)||^2]$, In this equation, $\epsilon_\theta(x_t,t)$ represents the noise predicted by the model, and ϵ_t is the actual noise added at the time step t. By minimizing this loss function, the model learns to estimate the noise at each time step effectively. The training process used a batch size of 36. Due to the large volume of data, we trained for 3 epochs, at which point realistic CT images could already be generated. The optimizer is AdamW, with a learning rate of 1e−4.

Training platform information

We utilized the Tianhe-2 supercomputing platform to conduct the training of the pre-training model alongside various downstream tasks. The platform boasts a peak computational speed of 10.07 petaflops per second and a sustained speed of 6.14 petaflops per second. Its total memory capacity is ~3PB, complemented by a global storage capacity of around 19PB. During the pre-training phase, we employed two nodes equipped with eight V100 GPUs, with each node housing 256GB of memory and two Xeon E5 series 12-core central processing units. The training duration for one epoch of the pre-training model was roughly 144 h, whereas the model with the longest training period took ~36 days to complete six epochs. The training code was implemented using Python version 3.8.12, PyTorch 2.10, Accelerator 0.24, and Diffusers 0.27.2.

Visualization of saliency maps

Grad-CAM[87] is used to generate the Saliency Map for the input image model. First, the activation feature maps of the convolutional layers are obtained through forward propagation, and the gradients of these feature maps relative to the target class are calculated through backpropagation. Then, these gradients are subjected to global average pooling to obtain the weights of each channel. These weights are used to weight the activation feature maps of the convolutional layers, producing a two-dimensional heat map of weighted sum, which indicates the contribution of different areas in the input image to the target category. Subsequently, the heat map is upscaled to the same size as the input image using bilinear interpolation, and finally, the heat map is visualized through color mapping to display the areas focused on by the model. The contour map represents lines of equal value within a saliency map.

Fine-tuning LCTfound to downstream tasks

To fully explore the potential of LCTfound across different tasks, we conducted adaptation designs and experiments on various downstream tasks by integrating multiple state-of-the-art deep learning techniques. As a result, LCTfound demonstrated strong competitiveness across six sub-tasks.

Mediastinal neoplasms segmentation. LCTfound contains rich prior knowledge of lung CT structures. To effectively leverage this prior knowledge for mediastinal neoplasm segmentation, a dense prediction task, we employed an MLP classifier to assign labels to individual pixels[85]. In summary, we used LCTfound to extract image features and trained an MLP classifier to classify the features extracted from each spatial location. Specifically, during training, we froze the parameters of LCTfound to prevent updates. Upon receiving input data, we used LCTfound to extract deep features corresponding to four different time steps. For each image, we obtained four features at different scales from LCTfound, which were then upsampled to match the input resolution and concatenated. The feature vector corresponding to each spatial location was then fed into the MLP classifier to predict the class of that pixel, and the loss was computed with the segmentation labels to update the MLP. During training, we used single-center data and split it into training, validation, and testing sets. The best model was selected based on the validation set, and results were reported on both the test set and an external test set. The AdamW optimizer was used with a weight decay of 1e−2 and an initial learning rate of 1e−3, which was gradually reduced to zero following a cosine decay schedule. The training lasted for 20 epochs.

The diagnosis of PAP. This task is a classification task, and DDPM is a UNet-based dense prediction task. To accomplish this task, we used features extracted from the deep feature learning module of the LCTfound for prediction. Specifically, during training, we input images corresponding to two different time steps and extracted features from the deep feature learning module, which had a dimension of $[h \times w \times c]$. We added an additional learnable convolutional module to reduce the resolution to $[h \times w \times 256]$, aiming to reduce the feature dimensions and better adapt the features for the classification task. The resulting feature vectors were averaged and normalized to obtain a $[1 \times 1 \times 256]$ feature vector. As we had input from different time steps, we concatenated the feature vectors and fed them into an MLP classifier to learn the classification. For this task, we used the stochastic gradient descent (SGD) optimizer with an initial learning rate of 0.001 and a weight decay of 1e−3. The model was trained for 10 epochs, with the learning rate reduced by a factor of 10 at the 7th epoch. Cross-entropy was used as the loss function.

NSCLC prognostication prediction. For this prognostication task, to better compare with existing work, we used two approaches to demonstrate LCTfound's capabilities. The first approach was linear adaptation, where we input images from four different time steps into LCTfound, collected the outputs from the deep feature module, flattened them, and concatenated the features to form a feature vector. This feature vector was trained and validated using a linear classifier, as in the literature, for comparison with other methods. The second approach was full parameters fine-tuning, which was similar to the PAP task. After obtaining and concatenating the deep features, we trained an MLP classifier to predict the patient's survival time. Since we used a 2D model, we obtained prediction results for each individual slice with disease. To determine whether a patient's overall survival exceeds 5 years, we averaged the predicted probabilities across all relevant slices and used the aggregated result for the final classification (Supplementary Fig. 38). For full fine-tuning, the SGD optimizer was used with a learning rate of 0.002, kept constant, and a weight decay of 1e−3. The model was trained for 4 epochs, and cross-entropy with label smoothing was used as the loss function.

Neoadjuvant response prediction. In this task, we employed a similar approach to that used in the diagnosis of PAP. During training, images corresponding to different time steps were inputted, and features were extracted from the deep feature learning module. The resulting

feature vectors were averaged and normalized before being fed into a MLP classifier for classification. For this task, we utilized the AdamW optimizer with an initial learning rate of 0.001 and a weight decay of 1e−3. The model was trained for 100 epochs, with the learning rate reduced by a factor of 10 at the 40th and 80th epochs. Binary cross-entropy was employed as the loss function.

Whole lung semantic segmentation. In this task, we fine-tuned all parameters of LCTfound to better support the segmentation of small objects, such as small airways and vessels. When training the segmentation model, we replaced the output layer of LCTfound to produce predictions corresponding to the number of classes. Only images at time step 0, with minimal Gaussian noise, were used as input during training. The AdamW optimizer was used with a constant learning rate of 1e−4 and a weight decay of 0.01. The training lasted for 40 epochs.

Low-dose CT enhancement. The low-dose CT enhancement task involves image-to-image mapping, which is highly suitable for diffusion models. Therefore, in this task, we adopted the approach from CoreDiff[88], which is a state-of-the-art model based on diffusion models for this type of task. Cold diffusion demonstrated that adding Gaussian noise is not the only method for image degradation in diffusion models. Thus, we treated low-dose images as a degraded version of full-dose images and used the approach cold-diffusion[36] to construct degraded images at different time steps, allowing the diffusion model to progressively learn to restore high-quality images. LCTfound, trained on large amounts of CT data, contains rich prior knowledge of different qualities and equipment, which aids in denoising with limited data. During training, we froze the parameters of LCTfound and only fine-tuned the added adjust sub-network. This approach helps mitigate overfitting and better utilize prior knowledge. The learning rate was set to 2e−4, and the Adam optimizer was used. Training was conducted for 15,000 steps.

Sparse view CT reconstruction. This task is also an image-to-image mapping task, where diffusion model-based methods have recently shown strong performance. In this task, we used a guided approach for fine-tuning[89]. Specifically, we fine-tuned our diffusion model on the downstream task dataset to generate data closer to that of the downstream task. After training, during inference, we guided the image generation process of the diffusion model using sparse view CT images from the validation set to produce images in the desired direction. Fine-tuning was conducted using a stochastic differential equation framework with a total of 2000 time steps and 5 epochs of iteration. The learning rate was set to 2e−4, and the Adam optimizer was used.

Segmentaion based on reconstructed CT image. In this task, we evaluated mediastinal tumor segmentation performance using CT images reconstructed by two different methods, and compared them with segmentation results obtained from original high-quality full-view CT images. The two reconstruction approaches were conventional FBP and the LCTfound-based reconstruction method. Segmentation performance was assessed using the Dice similarity coefficient. In the experimental setup, half of the mediastinal tumor training set was reserved for training a sparse-view CT reconstruction model based on LCTfound. This model was then used to reconstruct 32-view CT images for the remaining half of the training set, the internal validation set, the internal test set, and three external test sets. Subsequently, three segmentation models were trained on these reconstructed datasets: one on FBP-reconstructed images, one on LCTfound-reconstructed images, and one on the original full-view images. We then evaluated the performance of all three segmentation models across the internal validation set, internal test set, and the three external test sets.

Super-resolution. In the super-resolution task, we evaluated two upsampling settings with scaling factors of 4× and 8×. We designed a zero-shot super-resolution strategy for lung CT that leverages LCTfound to inject high-frequency details without any task-specific fine-tuning. Specifically, during inference, the model starts from pure noise and follows a standard reverse diffusion process, where each denoising step progressively refines the volume into a high-resolution estimate. This is immediately followed by a fidelity correction step, in which the generated image is merged with the original low-resolution CT scan. This two-stage refinement—first generating plausible high-frequency structures via the diffusion model, then fusing them with the original input—ensures that the final output remains faithful to the input while benefiting from the rich anatomical priors learned during pretraining, resulting in a substantial enhancement in image resolution.

Data collection and model training for virtual CTA imaging. For all CTA scans, the contrast agent used is iodine, at a concentration of 370 mg/mL and an injection speed of 4.5 mL/s. Each volume is treated separately, with intensity values from −200 to 200 pixels normalized to a scale of 0 to 1. The final data, organized by scans, is randomly split into training and testing sets for the purposes of model training and validation. Non-contrast CT images serve as conditional guidance for fine-tuning the outputs of LCTfound. The code used for training CTA-GAN is sourced from https://github.com/yml-bit/CTA-GAN, with the model retrained using default parameters in the code. For training Pixel2pixel, the code from https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix is used, retraining the model with default parameters. The training of conditional diffusion uses code from https://github.com/Janspiry/Palette-Image-to-Image-Diffusion-Models, with the model retrained using default parameters.

Evaluation metrics for classification and segmentation

Kaplan–Meier curves are used to illustrate the predictive outcomes of NSCLC prognosis. The survival probability S_t at any specific time is defined by the following formula:

$$S_t = \frac{N - D_t}{N} \tag{1}$$

Here, N represents the total number of patients at the start time, and D_t is the number of patients who died at time t.

Evaluation metrics for pixel-level tasks. Several different metrics are used to evaluate the results of lung low-dose CT enhancement and sparse view reconstruction. PSNR (Peak Signal to Noise Ratio) is a widely recognized metric to measure image quality. a higher PSNR indicates higher image quality. If the ground truth image is y, and the raw image is x, then the definition of PSNR is as follows:

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX^2}{MSE}\right) = 20 \cdot \log_{10}\left(\frac{MAX}{MSE}\right) \tag{2}$$

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}||x(i,j) - y(i,j)||^2 \tag{3}$$

Here, MAX the maximum pixel value, for normalized images $MAX = 1$; m and n are the two dimensions of the image.

RMSE quantifies the variance between two images directly and an RMSE approaching 0 indicates a greater preservation of visual

information between the raw image and the ground truth image, defined as follows:

$$RMSE = \sqrt{MSE} \tag{4}$$

VIF chiefly gauges the preservation level of visual information between the raw image and the ground truth image. The VIF It assesses image quality by evaluating the likeness in structural and textural information between two images. As VIF nears 1, it signifies a higher level of VIF between the raw and the ground truth image, correlating with improved image quality.

LPIPS quantifies image similarity by assessing perceptual differences between two images using a deep learning network. Images that are pixel-wise similar may still be distinguished as different by human observers. LPIPS utilizes features extracted by pre-trained CNNs (e.g., VGG, AlexNet) and calculates the feature distances to determine the perceptual likeness of images. Image quality is inversely proportional to the LPIPS. LPIPS is defined as follows:

$$LPIPS = \frac{1}{MN}\sum_{i=0}^{M-1}\sum_{j=0}^{M-1} || \cos \odot (X(i,j) - Y(i,j))||^2 \tag{5}$$

$$X = Net(x) \tag{6}$$

$$Y = Net(y) \tag{7}$$

Here X and Y represent the features of x and y extracted by the neural network; M and N represent the dimensions of the extracted features. The network used is the pre-trained VGG's output of the third layer features. cos represents the calculation of cosine distance.

SSIM is a metric that quantifies the resemblance between two images. SSIM assesses the images by comparing their luminance, contrast, and structural integrity separately, then applies weights to these three components and uses their product to represent the similarity. The calculation of the SSIM is carried out using a sliding window on the image. In this process, a window with the dimensions $a \times a$ is selected from the image for each calculation, and the SSIM is computed for that window. The overall SSIM for the image is the average of the values from all such windows after the image has been fully scanned. Higher SSIM corresponds to superior image quality. SSIM is defined as follows:

$$l(x,y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1} \tag{8}$$

$$c(x,y) = \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2} \tag{9}$$

$$s(x,y) = \frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3} \tag{10}$$

$$SSIM = l(x,y) \cdot c(x,y) \cdot s(x,y) \tag{11}$$

Here, μ_x and μ_y represent the mean values of x and y; σ_x and σ_y represent the variances of x and y, σ_{xy} represents the covariance between x and y. c_1, c_2, c_3 are three constants.

Feature similarity (FSIM) employs phase congruency (PC) to depict local structures. The gradient magnitude (GM) feature is utilized to offset the disadvantage of PC, which is relatively invariant to image alterations. The calculation of FSIM can be directly performed using available functions at (https://github.com/chaofengc/IQA-PyTorch). A higher CLIPIQA indicates a higher quality of the image. CLIPIQA assesses image quality utilizing the prior knowledge from the CLIP text-image model and can be directly executed by using CLIPImageQualityAssessment from pytorch. The lower the CLIPIQA, the higher the image quality. For a fair comparison of different methods, we used the same hyper parameters to calculate the aforementioned metrics.

Statistical analysis methods

To evaluate statistical significance and estimate confidence intervals, we employed different statistical methods depending on the nature of the metric. For metrics computed on a per-sample basis, such as segmentation Dice scores and image quality scores, we applied the Wilcoxon signed-rank test using the scipy.stats.wilcoxon function (SciPy version 1.15.3). This non-parametric test was chosen due to its robustness for paired, non-normally distributed data. For dataset level metrics, such as area under the ROC curve (AUC) and overall classification accuracy, we used a non-parametric bootstrap procedure. Specifically, we performed 1000 iterations of bootstrap sampling with replacement, each time drawing a sample of the same size as the original dataset. Then, we applied the Wilcoxon methods to estimate the P-values. P-values less than 0.05 were considered statistically significant.

Training process of contrastive methods

In the prognosis prediction of NSCLC, the comparison methods were implemented using the code from the author[25]. Both full-parameter fine-tuning and linear fine-tuning were performed using their default parameters. In the mediastinal neoplasms segmentation task, UNet, MedSAM, and InternImage use the same training script and settings as LCTfound, while UNI and SWIN are trained using their original code and default settings. MAE is trained using the training settings of the MMCV framework[90]. In the diagnosis of PAP disease and the prediction of the response prediction of neoadjuvant, the comparative methods are MAE and RadImagenet. The pre-trained weights for MAE are obtained using the MMCV framework and the same pre-training data as LCTfound, while RadImagenet uses the pre-trained weights released by the authors. The downstream tasks are trained with the same parameters as LCTfound. In the whole lung semantic segmentation, since the original MedSAM model outputs binary results, we trained 21 models. In the task of low-dose CT enhancement and sparse-view reconstruction, DUGAN, WGAN, DuDoTrans, SAX-NeRF, and DIF-Gaussian were performed using the authors' code.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All data supporting the findings of this study are included in the article and Supplementary Information files and Supplementary Data file. The data used to generate figures in this study have been deposited at [https://figshare.com/articles/figure/LCTfound/30343459]. The study utilized several publicly available datasets, including Mayo 2016 (available at https://www.cancerimagingarchive.net/collection/ldct-and-projection-data/), LUNA16, LUNG1 and NSCLC-Radiomics (https://www.cancerimagingarchive.net/collection/nsclc-radiomics/), AAPM, and COVIDx CT. The public data were accessed and utilized in accordance with the terms and conditions of each source, and no identifiable patient information was involved. The raw pretraining CT data are protected and are not publicly available due to data privacy laws. A anonymized subset of the pretraining data is available at: https://huggingface.co/datasets/GuoxunZhang1997/LCTfound. To access the data, users are required to complete a brief application form, primarily to verify their human identity and intended purpose of use.

Code availability

Our LCTfound can be found at https://github.com/gingerbread000/LCTfound. [DOI: 10.5281/zenodo.17310766].

References

1. Walsh, C. L. et al. Imaging intact human organs with local resolution of cellular structures using hierarchical phase-contrast tomography. *Nat. Methods* **18**, 1532–1541 (2021).
2. Kuan, A. T. et al. Dense neuronal reconstruction through X-ray holographic nano-tomography. *Nat. Neurosci.* **23**, 1637–1643 (2020).
3. Umetani, K., Okamoto, T., Saito, K., Kawata, Y. & Niki, N. 36M-pixel synchrotron radiation micro-CT for whole secondary pulmonary lobule visualization from a large human lung specimen. *Eur. J. Radiol. Open* **7**, 100262 (2020).
4. Eckermann, M. et al. 3D virtual pathohistology of lung tissue from Covid-19 patients based on phase contrast X-ray tomography. *eLife* **9**, e60408 (2020).
5. Maes, A. et al. Cryogenic contrast-enhanced microCT enables nondestructive 3D quantitative histopathology of soft biological tissues. *Nat. Commun.* **13**, 6207 (2022).
6. De Koning, H. J. et al. Reduced lung-cancer mortality with volume CT screening in a randomized trial. *N. Engl. J. Med.* **382**, 503–513 (2020).
7. Ardila, D. et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat. Med.* **25**, 954–961 (2019).
8. Primakov, S. P. et al. Automated detection and segmentation of non-small cell lung cancer computed tomography images. *Nat. Commun.* **13**, 3423 (2022).
9. Mei, X. et al. Interstitial lung disease diagnosis and prognosis using an AI system integrating longitudinal data. *Nat. Commun.* **14**, 2272 (2023).
10. He, Y. et al. Disorder-free data are all you need—inverse supervised learning for broad-spectrum head disorder detection. *NEJM AI* **1**, AIoa2300137 (2024).
11. Liu, A. et al. Automatic intracranial abnormality detection and localization in head CT scans by learning from free-text reports. *Cell Rep. Med.* **4**, 101164 (2023).
12. Liu, A., Guo, Y., Yong, J. & Xu, F. Multi-grained radiology report generation with sentence-level image-language contrastive Learning. *IEEE Trans. Med. Imaging* **43**, 2657–2669 (2024).
13. Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. *Improving Language Understanding by Generative Pre-training*. OpenAI blog (2018).
14. Brown, T. et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901 (2020).
15. Radford, A. et al. *Language Models are Unsupervised Multitask Learners*. OpenAI blog (2019).
16. Radford, A. et al. Learning transferable visual models from natural language supervision. *International conference on machine learning*, 8748–8763 (2021).
17. Chen, T. et al. A simple framework for contrastive learning of visual representations. *International conference on machine learning*, 1597–1607 (2020).
18. Caron, M. et al. Emerging properties in self-supervised vision transformers. In *Proc. IEEE/CVF international conference on computer vision*, 9650–9660 (2021).
19. Chen, R. J. et al. Towards a general-purpose foundation model for computational pathology. *Nat. Med.* **30**, 850–862 (2024).
20. Zhou, Y. et al. A foundation model for generalizable disease detection from retinal images. *Nature* **622**, 156–163 (2023).
21. Qiu, J. et al. Development and validation of a multimodal multitask vision foundation model for generalist ophthalmic artificial intelligence. *NEJM AI* **1**, AIoa2300221 (2024).
22. Zhang, X., Wu, C., Zhang, Y., Xie, W. & Wang, Y. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nat. Commun.* **14**, 4542 (2023).
23. Zhou, H.-Y. et al. A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. *Nat. Biomed. Eng.* **7**, 743–755 (2023).
24. Huang, W. et al. Enhancing representation in radiography-reports foundation model: a granular alignment algorithm using masked contrastive learning. *Nat. Commun.* **15**, 7620 (2024).
25. Pai, S. et al. Foundation model for cancer imaging biomarkers. *Nat. Mach. Intell.* **6**, 354–367 (2024).
26. Niu, C. et al. Medical multimodal multitask foundation model for lung cancer screening. *Nat. Commun.* **16**, 1523 (2025).
27. Yang, R. et al. Sharing massive biomedical data at magnitudes lower bandwidth using implicit neural function. *Proc. Natl. Acad. Sci. USA* **121**, e2320870121 (2024).
28. Xiong, Z. et al. How generalizable are foundation models when applied to different demographic groups and settings? *NEJM AI* **2**, AIcs2400497 (2024).
29. Feng, Z. et al. Early prediction of disease progression in COVID-19 pneumonia patients with chest CT and clinical characteristics. *Nat. Commun.* **11**, 4968 (2020).
30. Lassau, N. et al. Integrating deep learning CT-scan model, biological and clinical variables to predict severity of COVID-19 patients. *Nat. Commun.* **12**, 634 (2021).
31. Shan, H. et al. Competitive performance of a modularized deep neural network compared to commercial algorithms for low-dose CT image reconstruction. *Nat. Mach. Intell.* **1**, 269–276 (2019).
32. Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **33**, 6840–6851 (2020).
33. He, K. et al. Masked autoencoders are scalable vision learners. In *Proc. IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009 (2022).
34. Ma, J. et al. Segment anything in medical images. *Nat. Commun.* **15**, 654 (2024).
35. Vaswani, A. et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, 5998–6008 (2017).
36. Bansal, Arpit, et al. Cold diffusion: Inverting arbitrary image transforms without noise. *Adv. Neural Inf. Process. Syst.* **36**, 41259–41282 (2023).
37. Wang, M., Herbst, R. S. & Boshoff, C. Toward personalized treatment approaches for non-small-cell lung cancer. *Nat. Med.* **27**, 1345–1356 (2021).
38. Leiter, A., Veluswamy, R. R. & Wisnivesky, J. P. The global burden of lung cancer: current status and future trends. *Nat. Rev. Clin. Oncol.* **20**, 624–639 (2023).
39. Herbst, R. S., Morgensztern, D. & Boshoff, C. The biology and management of non-small cell lung cancer. *Nature* **553**, 446–454 (2018).
40. Chaft, J. E. et al. Evolution of systemic therapy for stages I–III non-metastatic non-small-cell lung cancer. *Nat. Rev. Clin. Oncol.* **18**, 547–557 (2021).
41. Gridelli, C. et al. Non-small-cell lung cancer. *Nat. Rev. Dis. Prim.* **1**, 15009 (2015).
42. Chen, S., Ma, K. & Zheng, Y. Med3D: Transfer learning for 3D medical image analysis. arXiv preprint arXiv:1904.00625 (2019).
43. Zhou, Z. et al. Models genesis: Generic autodidactic models for 3d medical image analysis. *International conference on medical image computing and computer-assisted intervention*, 384–393 (2019).
44. She, Y. Deep learning for predicting major pathological response to neoadjuvant chemoimmunotherapy in non-small cell lung cancer: a multicentre study. *eBioMedicine* **86**, 104364 (2022).

45. Deutsch, J. S. et al. Association between pathologic response and survival after neoadjuvant therapy in lung cancer. *Nat. Med.* **30**, 218–228 (2024).
46. Banna, G. L. et al. Neoadjuvant chemo-immunotherapy for early-stage non–small cell lung cancer: a systematic review and meta-analysis. *JAMA Netw. Open* **7**, e246837 (2024).
47. Sorin, M. et al. Neoadjuvant chemoimmunotherapy for NSCLC: a systematic review and meta-analysis. *JAMA Oncol.* **10**, 621 (2024).
48. Mei, X. et al. RadImageNet: an open radiologic deep learning research dataset for effective transfer learning. *Radiol. Artif. Intell.* **4**, e210315 (2022).
49. Carter, B. W. et al. ITMIG classification of mediastinal compartments and multidisciplinary approach to mediastinal masses. *RadioGraphics* **37**, 413–436 (2017).
50. Jiang, Y. et al. Spatiotemporal distribution of mediastinal neoplasms: A comprehensive multi-center study. *Lung Cancer* **191**, 107558 (2024).
51. Wang, W. et al. InternImage: exploring large-scale vision foundation models with deformable convolutions. In *Proc. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 14408–14419 (IEEE, Vancouver, BC, Canada, 2023).
52. Tang, Y. et al. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proc. IEEE/CVF conference on computer vision and pattern recognition*, 20730–20740 (2022).
53. Liu, J. et al. Clip-driven universal model for organ segmentation and tumor detection. In *Proc. IEEE/CVF international conference on computer vision*, 21152–21164 (2023).
54. Trapnell, B. C. et al. Pulmonary alveolar proteinosis. *Nat. Rev. Dis. Prim.* **5**, 1–17 (2019).
55. Wijsenbeek, M., Suzuki, A. & Maher, T. M. Interstitial lung diseases. *Lancet* **400**, 769–786 (2022).
56. Trapnell, B. C. et al. Inhaled molgramostim therapy in autoimmune pulmonary alveolar proteinosis. *N. Engl. J. Med.* **383.17**, 1635–1644 (2020).
57. Borie, R. et al. Pulmonary alveolar proteinosis. *Eur. Respir. Rev.* **20**, 98–107 (2011).
58. Yang, H.-X. et al. Long-term survival based on the surgical approach to lobectomy for clinical stage I nonsmall cell lung cancer: comparison of robotic, video-assisted thoracic surgery, and thoracotomy lobectomy. *Ann. Surg.* **265**, 431 (2017).
59. Gex, G. et al. Diagnostic yield and safety of electromagnetic navigation bronchoscopy for lung nodules: a systematic review and meta-analysis. *Respiration* **87**, 165–176 (2014).
60. Sato, M. et al. Use of virtual assisted lung mapping (VAL-MAP), a bronchoscopic multispot dye-marking technique using virtual images, for precise navigation of thoracoscopic sublobar lung resection. *J. Thorac. Cardiovasc. Surg.* **147**, 1813–1819 (2014).
61. Oudkerk, M., Liu, S., Heuvelmans, M. A., Walter, J. E. & Field, J. K. Lung cancer LDCT screening and mortality reduction — evidence, pitfalls and future perspectives. *Nat. Rev. Clin. Oncol.* **18**, 135–151 (2021).
62. Ruano-Ravina, A., Pérez-Ríos, M., Casàn-Clará, P. & Provencio-Pulla, M. Low-dose CT for lung cancer screening. *Lancet Oncol.* **19**, e131–e132 (2018).
63. Lyu, J. et al. Generative adversarial network–based noncontrast CT angiography for aorta and carotid arteries. *Radiology* **309**, e230681 (2023).
64. Zhu, J.-Y. et al. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. IEEE international conference on computer vision*, 2223–2232, (2017).
65. Saharia, C. et al. Palette: Image-to-image diffusion models. *ACM SIGGRAPH 2022 conference proceedings*, 1–10 (2022).
66. Lin, Y. et al. Learning 3d gaussians for extremely sparse-view cone-beam ct reconstruction. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 425–435 (2024).
67. Wang, C. et al. DuDoTrans: Dual-Domain Transformer for Sparse-View CT Reconstruction. In *Machine Learning for Medical Image Reconstruction*, Vol. 13587, 84–94 (Springer, Cham, 2022).
68. Cai, Y., Wang, J., Yuille, A., Zhou, Z. & Wang, A. Structure-aware sparse-view X-ray 3d reconstruction. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11174–11183 (2024).
69. Yang, Q. et al. Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss. *IEEE Trans. Med. Imaging* **37**, 1348–1357 (2018).
70. Yi, Z., Zhang, H., Tan, P. & Gong, M. DualGAN: unsupervised dual learning for image-to-image translation. In *Proc. 2017 IEEE International Conference on Computer Vision (ICCV)* 2868–2876 (IEEE, Venice, 2017).
71. Wang, J., Chan, K. C. & Loy, C. C. Exploring clip for assessing the look and feel of images. In *Proc. AAAI conference on artificial intelligence*. Vol. **37,** 2555–2563 (2023).
72. Zhang, R. et al. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. IEEE conference on computer vision and pattern recognition*, 586–595 (2018).
73. Yang, Y. et al. A digital mask to safeguard patient privacy. *Nat. Med.* **28**, 1883–1892 (2022).
74. Albrecht, M. H. et al. State-of-the-art pulmonary CT angiography for acute pulmonary embolism. *Am. J. Roentgenol.* https://doi.org/10.2214/AJR.16.17202 (2016).
75. Apfaltrer, P. et al. Value of monoenergetic low-kV dual energy CT datasets for improved image quality of CT pulmonary angiography. *Eur. J. Radiol.* **83**, 322–328 (2014).
76. Han, S. et al. Learning both weights and connections for efficient neural network. *Adv. Neural Inf. Process. Syst.* **28**, 1135–1143 (NIPS, 2015).
77. Hinton, G., Vinyals, O. & Dean, J. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015).
78. Lin, J. et al. "Mcunet: Tiny deep learning on iot devices". *Adv. Neural Inf. Process. Syst.* **33**, 11711–11722 (2020).
79. Wang, C. et al. Data-driven risk stratification and precision management of pulmonary nodules detected on chest computed tomography. *Nat. Med.* 1–12 https://doi.org/10.1038/s41591-024-03211-3 (2024).
80. Heumos, L. et al. "An open-source framework for end-to-end analysis of electronic health record data.". *Nat. Med.* **30.11**, 3369–3380 (2024).
81. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
82. Chen, Y. et al. MIMO: a medical vision language model with visual referring multimodal input and pixel grounding multimodal output. In *Proc. IEEE/CVF conference on computer vision and pattern recognition*, 24732–24741 (2025).
83. Huang, X. et al. Towards a multimodal large language model with pixel-level insight for biomedicine. *Proc. AAAI Conf. Artif. Intell.* **39**, 3779–3787 (2025).
84. Moor, M. et al. Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).
85. Baranchuk, D. et al. Label-efficient semantic segmentation with diffusion models. *International Conference on Learning Representations*, 15633–15647 (2021).
86. Liu, Z. et al. Swin transformer: hierarchical vision transformer using shifted windows. In *Proc. IEEE/CVF conference on computer vision and pattern recognition*, 10012–10022 (2021).
87. Selvaraju, R. R. et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proc. IEEE/CVF conference on computer vision and pattern recognition*, 618–626 (2017).
88. Gao, Q., Li, Z., Zhang, J., Zhang, Y. & Shan, H. CoreDiff: Contextual error-modulated generalized diffusion model for low-dose CT

denoising and generalization. *IEEE Trans. Med. Imaging* **43**, 745–759 (2023).
89. Chung, H., Ryu, D., McCann, M. T., Klasky, M. L. & Ye, J. C. Solving 3d inverse problems using pre-trained 2d diffusion models. In *Proc. IEEE/CVF conference on computer vision and pattern recognition*, 22542–22551 (2023).
90. MMCV: OpenMMLab Computer Vision Foundation. Version 2.0, MMCV Contributors, 2023, GitHub, https://github.com/open-mmlab/mmcv.

Acknowledgements

This work was supported by National Science and Technology Major Project No.2023ZD0506304 (Y.G.), National Natural Science Foundation of China No.82441013(Y.G.), No.62021002(F.X.) and No. 62088102 (Q.D.), the R&D Program of Guangzhou National Laboratory (grant SRPG22-017, H.L., J.H.).

Author contributions

Q.D., T.W., F.X., J.H., and Y.G. conceived the LCTfound project and revised the manuscript. G.Z. and Z.G. implemented the LCTfound pipeline, completed the fine-tuning of downstream tasks, organized the experimental results, and composed the manuscript. H.L. collected data and established the LungCT-20M dataset. J.L. and L.M. completed the saliency visualization of LCTfound attention. T.W., YC.G., X.C., Z.Y., and Y.C. accomplished the three-dimensional visualization for the whole lung segmentation.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-66620-z.

Correspondence and requests for materials should be addressed to Jianxing He, Feng Xu, Tien Yin Wong, Yuchen Guo or Qionghai Dai.

Peer review information *Nature Communications* thanks Yang Yang, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.