npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-02085-0

From pretraining to privacy: federated ultrasound foundation model with self-supervised learning

Yuncheng Jiang[1,2,3,4,19], Chun-Mei Feng[5,19], Jinke Ren[1,2], Jun Wei[6], Zixun Zhang[1,2], Yiwen Hu[7], Yunbi Liu[8], Rui Sun[1,2], Xuemei Tang[9], Juan Du[10], Xiang Wan[11], Yong Xu[12], Bo Du[13], Xin Gao[14,15], Guangyu Wang[16], Shaohua Zhou[17,18], Shuguang Cui[1,2] & Zhen Li[1,2] ✉

Ultrasound imaging is widely used in clinical diagnosis due to its non-invasive nature and real-time capabilities. However, traditional ultrasound diagnostics relies heavily on physician expertise and is often hampered by suboptimal image quality, leading to potential diagnostic errors. While artificial intelligence (AI) offers a promising solution to enhance clinical diagnosis by detecting abnormalities across various imaging modalities, existing AI methods for ultrasound face two major challenges. *First*, they typically require vast amounts of labeled medical data, raising serious concerns regarding patient privacy. *Second*, most models are designed for specific tasks, which restricts their broader clinical utility. To overcome these challenges, we present **UltraFedFM**, an innovative privacy-preserving ultrasound foundation model. UltraFedFM is collaboratively pre-trained using federated learning across 16 distributed medical institutions in 9 countries, leveraging a dataset of over 1 million ultrasound images covering 19 organs and 10 ultrasound modalities. This extensive and diverse data, combined with a secure training framework, enables UltraFedFM to exhibit strong generalization and diagnostic capabilities. It achieves an average area under the receiver operating characteristic curve (AUROC) of 0.927 for disease diagnosis and a dice similarity coefficient (DSC) of 0.878 for lesion segmentation. Notably, UltraFedFM surpasses the diagnostic accuracy of mid-level ultrasonographers (4−8 years of experience) and matches the performance of expert-level sonographers (10+ years of experience) in the joint diagnosis of 8 common systemic diseases.c These findings indicate that UltraFedFM can significantly enhance clinical diagnostics while safeguarding patient privacy, marking a significant advancement in AI-driven ultrasound imaging for future clinical applications.

Ultrasound is becoming increasingly important in clinical practice worldwide. It offers significant advantages over magnetic resonance imaging (MRI) and computed tomography (CT), including freedom from radiation, non-invasive nature, and cost-effectiveness. Thus, it is widely adopted as the primary imaging method for monitoring fetal growth during pregnancy[1], diagnosing internal organ pathology, and assisting in surgical decision-making[2]. However, ultrasound-based diagnosis relies heavily on the clinician's experience, while factors like noise and artifacts in the images can compromise quality and hinder the clinician's assessment of pathological regions, increasing the risk of missed or incorrect diagnoses[3,4]. Recent efforts have turned to artificial intelligence (AI) technologies to mitigate ultrasound-specific artifacts (e.g. speckle, false textures) and enhance diagnostic accuracy[5–15]. These contributions demonstrate that careful preprocessing, task-specific network design, and curated annotations can substantially improve performance for single-organ tasks. Despite notable successes, existing AI-based ultrasound models typically focus on very specific medical scenarios and require large amounts of high-quality labeled data, which restricts their scalability and generalizability across diverse medical applications.

Over the past two years, foundational models (FMs) have attracted much attention due to their generality and high performance. In the medical field, many efforts[16,17] have leveraged unlabeled ultrasound data to pre-train

A full list of affiliations appears at the end of the paper. ✉e-mail: lizhen@cuhk.edu.cn

FMs and fine-tuned them for specific tasks using labeled data. However, existing ultrasound foundational models (USFMs) face three key challenges: (1) **Data privacy**. Ultrasound data are distributed across multiple medical institutions and cannot be shared due to privacy regulations (e.g., GDPR[18]), restricting the volume of data available for pre-training; (2) **Limited modality**. Many USFMs are designed for particular ultrasound imaging modalities (e.g., echocardiograms), limiting their applicability to other imaging modalities and reducing their versatility; (3) **Imbalanced data distribution**. Existing USFMs often face an imbalance caused by the long-tailed distribution of the organ/lesion types represented in the dataset (e.g., 91% breast ultrasound in 3M-US[16]), leading to a biased performance in diagnosing uncommon conditions. These challenges highlight the need for new solutions that simultaneously address data privacy, scalability, and generalizability across various ultrasound imaging modalities and clinical scenarios.

In this work, we introduce UltraFedFM, a novel ultrasound foundation model pre-trained collaboratively by multiple medical institutions without exposing and aggregating all the data together. Specifically, we utilize a federated learning framework with one server and 16 clients from 9 countries, collectively possessing 1, 015, 754 unlabeled ultrasound images (Fig. 1a). These images cover 19 systemic organs and 10 ultrasound imaging modalities (Fig. 2a), providing an extensive and diverse representation for pre-training. By leveraging large-scale unlabeled datasets, UltraFedFM addresses key challenges in the medical field with the following solutions: **a**. When new modalities or organ data are continuously introduced, UltraFedFM can continuously update the model on new clients without accessing private data from other clients, thereby effectively safeguarding patient privacy; **b**. UltraFedFM minimizes the reliance on labor-intensive annotations by medical professionals, overcoming a critical bottleneck in medical AI development. This unsupervised approach ensures that valuable medical data can be efficiently utilized without requiring costly annotations from medical experts. The development of UltraFedFM consists of two stages: (1) Federated pre-training, in which the multiple clients collaboratively pre-train a shared model in a distributed, self-supervised manner. Throughout the pre-training process, the server periodically aggregates the local model parameters from each client without accessing their private data (Fig. 1b); (2) Downstream fine-tuning, where the pre-trained FM is fine-tuned using specific data to adapt to various clinical tasks, such as disease screening and diagnosis, sub-classification of disease phenotypes (e.g., tumor infiltration depth and type classification), prenatal maternal-fetal health analysis, and critical lesion identification and segmentation (Fig. 1c).

UltraFedFM is adapted to various ultrasound imaging modalities, modes, qualities, and clinical tasks. To accommodate the diverse features of different modalities, we propose a dynamic ultrasound image masking approach based on the specific texture features of organs and lesions. Additionally, we incorporate a random image corruption branch within the masked image modeling process to handle low-quality images commonly encountered in real-world scenarios. Furthermore, we use simple yet effective image transformations to generate simulated ultrasound images, aiming to address the uneven distribution of scan patterns in the pre-training dataset.

We conduct extensive experiments to evaluate the performance of UltraFedFM. To provide a fair and comprehensive evaluation, we collect and curate the largest ultrasound evaluation benchmark, covering the two most common ultrasound clinician tasks (i.e., disease diagnosis and lesion segmentation) with 11 sub-tasks from 19 ultrasound datasets. Several fully-supervised methods and a state-of-the-art USFM[16] are utilized for comparison. Experimental results demonstrate that UltraFedFM outperforms all baselines, achieving an average area under the curve (AUROC) of 0.927 for disease diagnosis and a dice similarity coefficient (DSC) of 0.878 for lesion segmentation. Notably, UltraFedFM outperforms ultrasonographer clinicians with intermediate levels (e.g., 4−8 years of clinical experience) and achieves comparable performance to high-level (e.g., more than 10 years of clinical experience) ultrasonographers in the joint diagnosis of 8 common systemic diseases. Furthermore, UltraFedFM leverages the principles of federated learning, enabling continuous model updates without the need for centralized data aggregation. This capability ensures that the model can be further trained using private data from different institutions or clients while preserving privacy and adhering to data protection regulations. By avoiding direct data sharing, UltraFedFM addresses critical privacy concerns, fostering trust and collaboration across institutions. With these capabilities, UltraFedFM provides a reliable model for clinical tasks, making it a pioneering solution for advancing ultrasound AI across institutions, regions, and clinical tasks.

Results

UltraFedFM enables systemic disease diagnosis and can assist clinicians in the diagnostic process

UltraFedFM aims to serve as a comprehensive FM in ultrasound imaging. To assess its effectiveness for disease diagnosis, 6 publicly available datasets and 2 private datasets (see Supplementary Table 2) are utilized, covering 8 kinds of organs (i.e., pancreas, gallbladder, liver, lung, colorectum, breast, heart, and fetal organs, see Supplementary Fig. 6) and 6 ultrasound imaging modalities (i.e., abdominal, lung, endorectal, superficial, echocardiogram, and fetal ultrasound). To provide an overall assessment for UltraFedFM, we average its performance on these datasets, and compare it with 4 baseline methods, including supervised training from scratch, ImageNet-21k centralized pre-training, USFM[16] centralized pre-training, and masked autoencoder (MAE)[19] federated pre-training. More details of the four methods are described in the Method section. The experimental results are shown in Fig. 3. We observe that UltraFedFM achieves an average AUROC of 0.927, which significantly ($p < 0.05$) outperforms its counterparts, surpassing the second-best USFM with an average AUROC of 0.894, by 0.033 ($p = 0.002$). Additionally, UltraFedFM performs well in data-limited situations (see line plot in Fig. 3b). As the amount of fine-tuning data is progressively reduced to 80%, 60%, 40%, and 20%, UltraFedFM remains robust, with only a modest decline in average AUROC of 0.124 (fine-tuning data from 100% to 20%), outperforming other methods. Notably, due to the federated pre-training with the large volume of unlabeled data, UltraFedFM possesses powerful feature extraction capabilities and can identify various types of lesions using a single organ-agnostic decoder, thus eliminating the need for task-specific classifiers utilized in other FMs. To demonstrate this, we constructed an organ-agnostic dataset by combining eight distinct datasets from different organs and fine-tuned UltraFedFM to recognize eight types of malignant tumors. It is observed that UltraFedFM accurately identifies most categories without requiring a separate classifier for each organ and the predicted scores of UltraFedFM concentrate in higher confidence intervals (Fig. 3c). Figure 3d illustrates the receiver operating characteristic (ROC) curves among eight different diseases, showing that UltraFedFM achieves superior efficiency in organ-agnostic disease diagnosis. More quantitative results for UltraFedFM, including accuracy, F1-score, and ROC, are provided in Supplementary Fig. 7, Supplementary Fig. 8, and Supplementary Fig. 9.

To evaluate the reliability of UltraFedFM's generalist intelligence in clinical practice, we compare it with ultrasonographers having different clinical levels. Seven ultrasonographers participated in this study, of whom two are intermediate-level (clinicians A, B: 4−8 years of clinical experience) and five are high-level (clinicians C-G: more than 10 years of clinical experience). A total of 80 ultrasound images containing 8 systemic malignant diseases were tested. As shown in Fig. 3e and Supplementary Table 7, UltraFedFM outperforms the ultrasonographers with intermediate-level and achieves comparable performance with high-level ultrasonographers. More specifically, while some specific organ diseases are easy for ultrasonographers to diagnose (e.g., average accuracy: 0.800 for breast and 0.871 for kidney), their diagnostic capabilities are limited when multiple ultrasound diseases are jointly diagnosed (e.g., average accuracy: 0.314 for gallbladder). In contrast, UltraFedFM can provide a consistent and accurate diagnosis of different ultrasound organ diseases (average accuracy: 0.900 for breast, 1.000 for kidney, and 0.800 for gallbladder). These results reveal that the UltraFedFM has the potential to serve as a reliable decision support tool to assist clinicians in prioritizing cases, reducing repetitive workloads, and minimizing missed diagnoses.

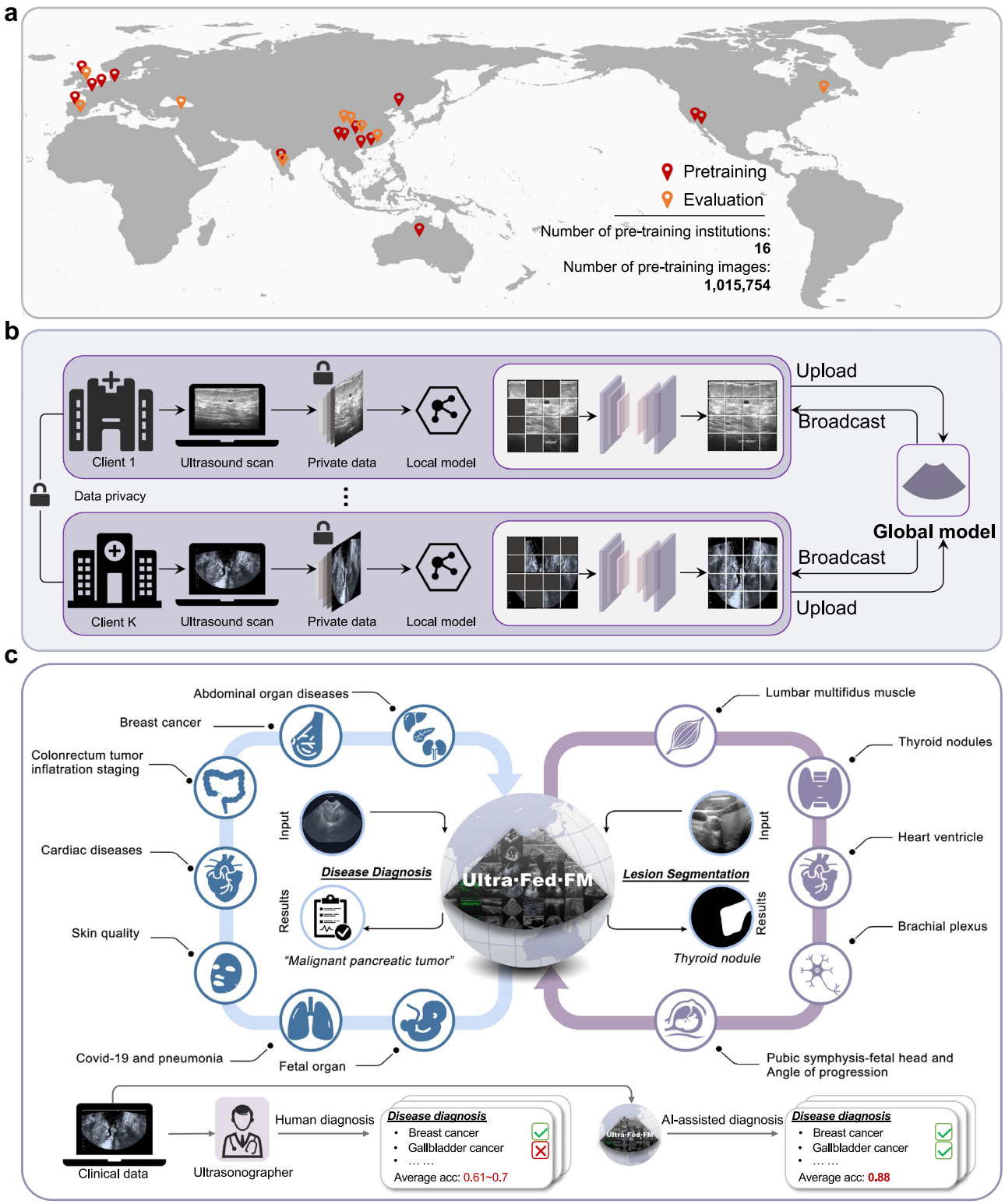

Fig. 1 | Overview of the study. a Medical data from 16 institutions and 9 countries are collected to pre-train and evaluate UltraFedFM, encompassing 1 million ultrasound images with extensive diversity. **b** The pre-training framework of UltraFedFM, where each client uses its private data to pre-train a local model through pixel-level reconstruction. During pre-training, only the local model parameters are uploaded for learning the global model, thus eliminating the risk of privacy breaches. Icons used are free to download from www.iconfont.cnand do not involve commercial use. **c** Clinical applications of UltraFedFM. UltraFedFM is a versatile ultrasound foundation model capable of handling multiple ultrasound scenarios, supporting multi-disease, multi-modal, and multi-task applications, and demonstrating superior performance compared with ultrasonographers in real clinical scenarios. Icons used are free to download from www.iconfont.cnand do not involve commercial use.

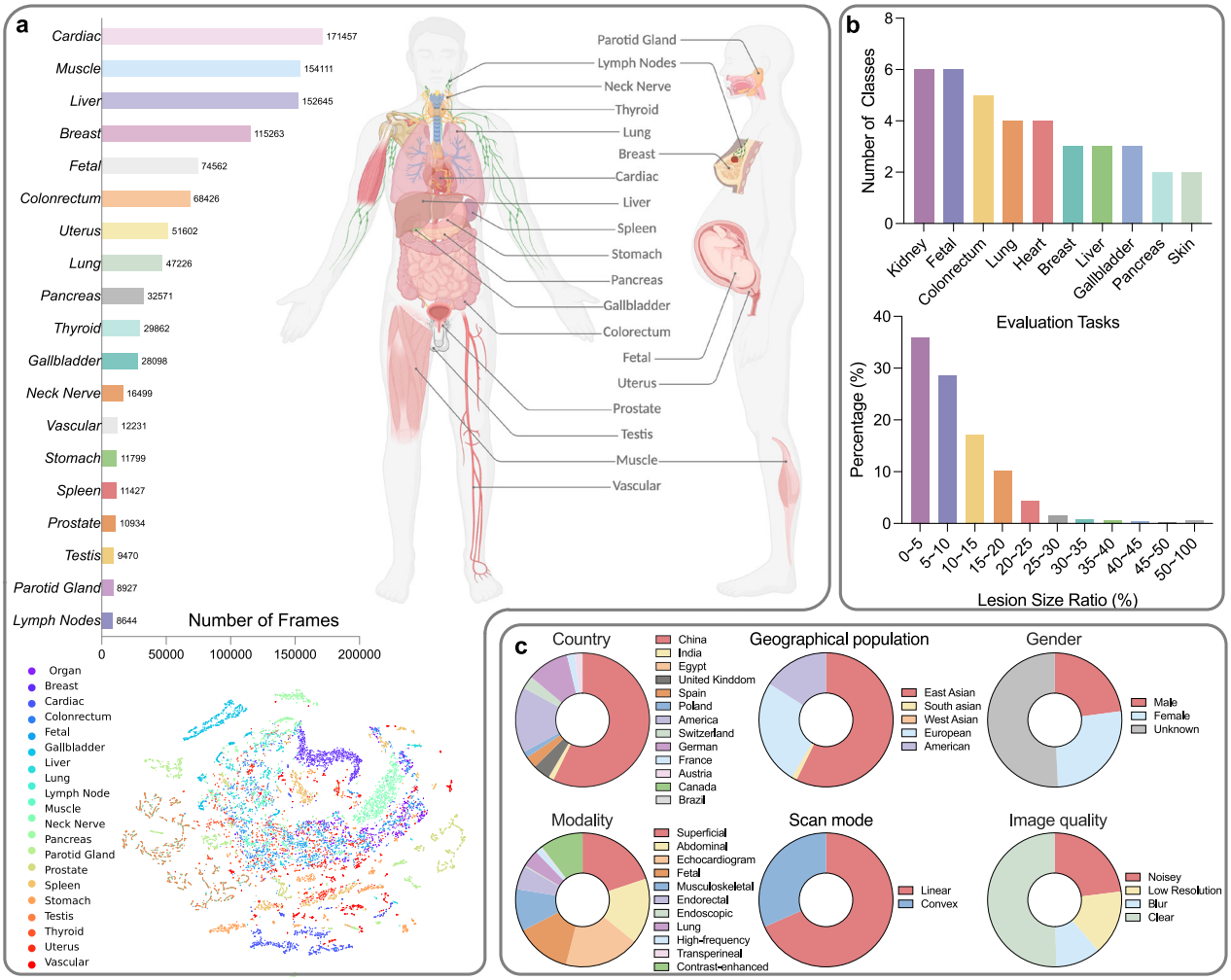

Fig. 2 | Statistics of the pre-training and downstream validation datasets. a The pre-training dataset covers 19 major organs across the entire body captured by various ultrasound imaging modalities. The figure are created in Biorender and have obtained publication license. **b** The distribution of class numbers for each downstream diagnosis dataset, ranging from basic binary classification to complex multi-class classification, and the distribution of the target size for organ and lesion segmentation tasks in the downstream validation dataset. Most segmentation targets occupy less than 1/10 of the entire image. **c** The distribution of sensitive information in the dataset across six attributes.

UltraFedFM facilitates organ and lesion segmentation

Organ and lesion segmentation for ultrasound images is crucial for clinical decision-making. To assess UltraFedFM's segmentation accuracy across different ultrasound imaging modalities, we evaluated it on four binary segmentation datasets (nerve[20], muscle[21], heart[22], and thyroid[23–26]) and one multi-class segmentation dataset (pubic symphysis-fetal head[27]). UltraFedFM consistently achieves high segmentation accuracy, successfully managing targets with diverse shapes and structures. In the binary segmentation task (Fig. 4a), UltraFedFM achieves the highest average dice similarity coefficient (DSC) score of 0.857 across the three binary segmentation datasets, significantly outperforming all other baselines ($p < 0.005$). In particular, USFM achieves a DSC score of 0.828, which is much lower than that of UltraFedFM ($p = 0.002$).

The multi-class segmentation task involves two steps, beginning by segmenting the pubic symphysis and fetal head, followed by measuring the angle between them. In this task, UltraFedFM achieves a DSC score of 0.842, significantly outperforming the second-best method (USFM[16]) with a DSC score of 0.810 ($p = 0.004$). Additionally, UltraFedFM excels in measuring the angle of progression (AoP), with a mean absolute error of 8.80, outperforming all baselines by a significant margin ($p < 0.005$) (Fig. 4a). Similar to the classification settings, we also evaluated UltraFedFM's effectiveness in scenarios with limited labeled data (Fig. 4b). Notably, even with 20% of the fine-tuning data, UltraFedFM still achieves an average DSC score of 0.772, outperforming the supervised method and USFM by 14.0% and 2.3%, respectively. To further assess UltraFedFM's generalization capability, we compiled an organ-agnostic segmentation dataset comprising five types of lesions. As shown in Fig. 4c, UltraFedFM demonstrates superior performance in locating and segmenting these lesions using a single unified segmentation model.

We also conducted cross-institutional validation (Fig. 4d) and imbalanced scanning mode validation (Fig. 4e). The former explores the segmentation generalization capability of UltraFedFM across different organ modalities, while the latter evaluates its performance under varying scanning modes. Across all cross-validation datasets, UltraFedFM consistently outperformed other baseline models ($p < 0.01$), demonstrating exceptional stability and balanced generalization capability. Figure 4e shows the fine-tuning performance of UltraFedFM under different data ratios of linear array and convex array scanning modes. When the data distribution was highly imbalanced (e.g., 0%:100% or 100%:0%), both models exhibited uncontrollable bias and overfitting during training, leading to a decline in prediction performance. In contrast, when the data proportions were more balanced, the models achieved optimal performance. This indicates that the feature distribution of images plays a crucial role in both pre-training and fine-tuning stages.

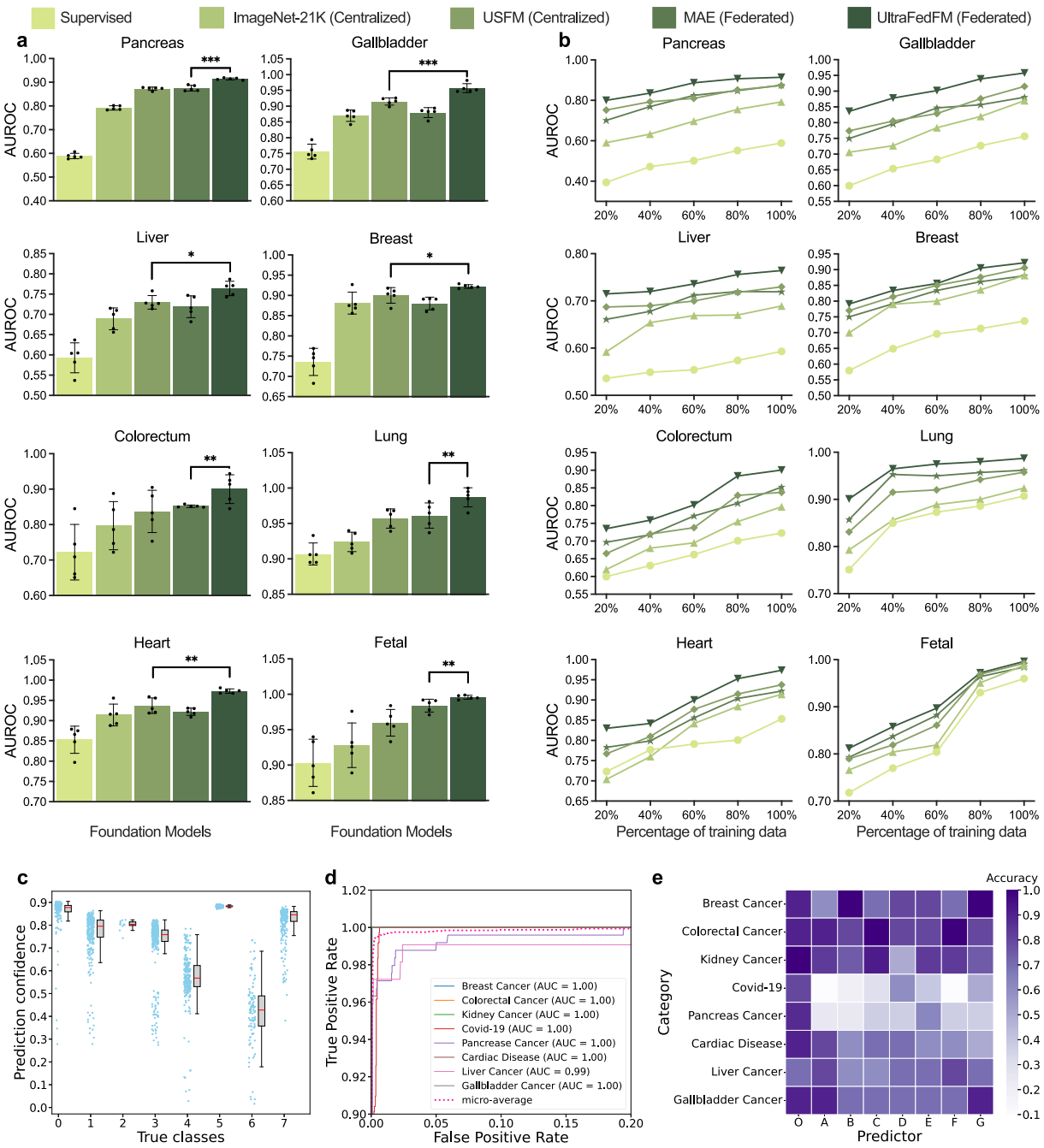

Fig. 3 | Diagnostic performance for systemic disease classification. a Internal validation of disease classification performance across eight diagnostic tasks. Comparative analysis shows model performance when fine-tuned on complete datasets. For each task, we fine-tune the model with five different random seeds. The error bars show 95% confidence intervals (CI) of the estimates, and the bar center is the mean estimate. We compare the performance using the area under the receiver operating characteristic curve (AUROC). *P*-value is calculated with the two-sided *t*-test between UltraFedFM and the most competitive comparison model. *, **, *** denotes $p < 0.05$, $p < 0.01$, and $p < 0.001$. **b** The experimental results of disease classification on limited labeled data subsets. **c**, **d** Performance of UltraFedFM on organ-agnostic fine-tuning setting. **c** Prediction confidence distribution over eight disease classes. The center line of the box denotes the median, while the box edges represent the first and third quartiles and the whiskers extend to 1.5 times the inter-quartile range. **d** Receiver Operating Characteristic (ROC) curves for distinct disease categories. **e** Generalist diagnostic accuracy of UltraFedFM across eight diseases and the comparison with seven experienced ultrasonographers.

UltraFedFM outperforms existing ultrasound task-specific methods

To further evaluate the excellence of UltraFedFM in medical image analysis tasks, we comprehensively compared it with existing task-specific methods in the ultrasound field. We chose four representative tasks, namely fetal plane classification, gallbladder cancer classification, breast nodule segmentation, and thyroid nodule segmentation, for evaluation. For each task, we chose both traditional models and the latest high-performing methods as comparisons. Detailed information about the datasets and comparative methods is presented in the following Supplementary Table 8 and Supplementary Table 9. The results are illustrated in Fig. 5a. The first two sub-figures display the results of classification tasks (evaluated by accuracy), while the last two sub-figures present the results of segmentation tasks (evaluated by the Dice similarity coefficient). Overall, UltraFedFM

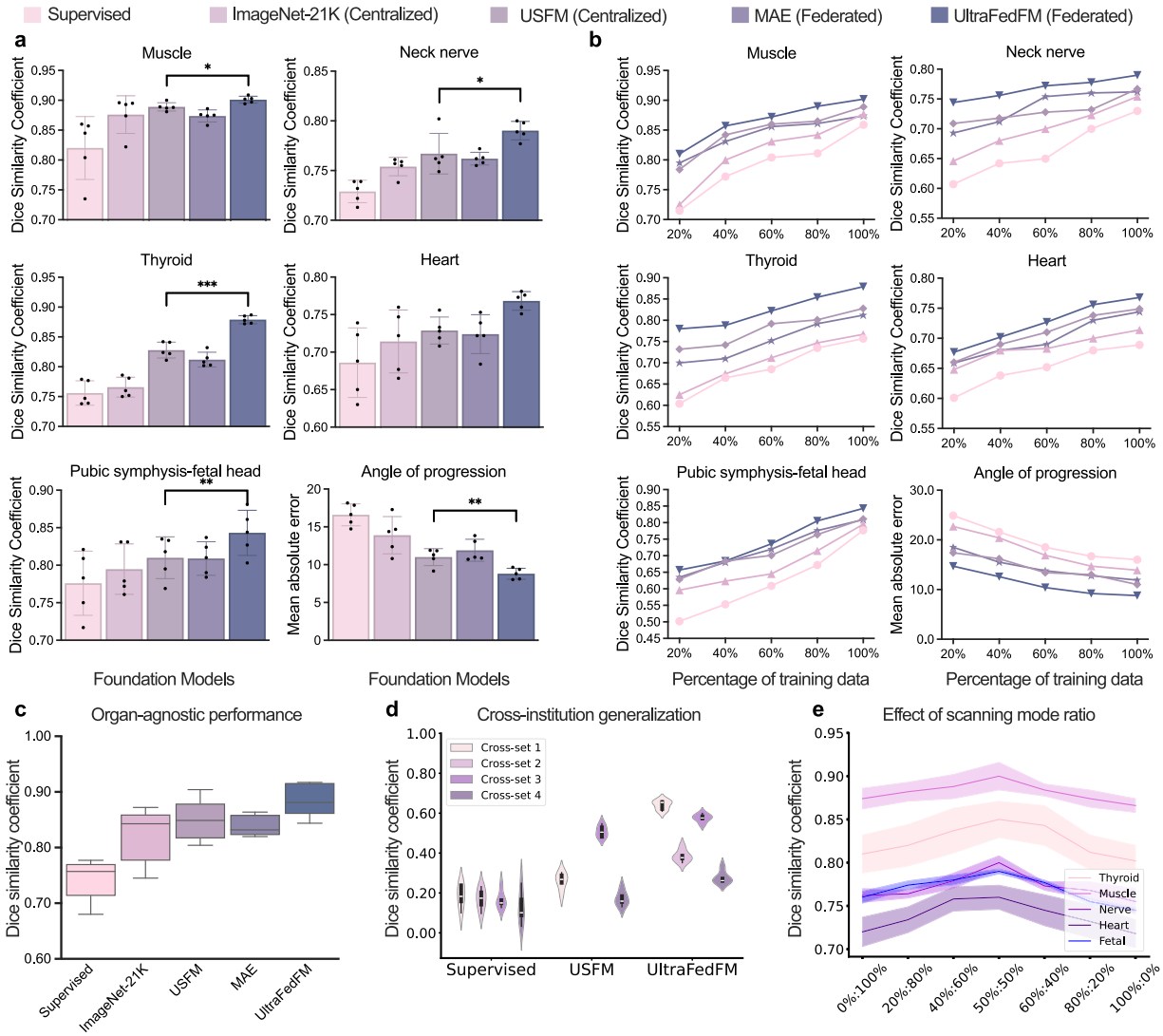

Fig. 4 | Performance for organ and lesion segmentation. a Internal validation of segmentation performance across eight diagnostic tasks. Comparative analysis shows model performance when fine-tuned on complete datasets. We compared the performance using the Dice similarity coefficient (DSC) score. **b** The experimental results of disease classification on limited labeled data subsets. **c** Performance of UltraFedFM on organ-agnostic dataset fine-tuned with a single decoder. **d** Comparison of cross-institution generalization performance. Cross-set 1 denotes fine-tuning on thyroid dataset and test on muscle dataset; Cross-set 2 denotes fine-tuning on thyroid dataset and test on muscle dataset; Cross-set 3 denotes fine-tuning on muscle dataset and test on thyroid dataset; Cross-set 4 denotes fine-tuning on muscle dataset and test on nerve dataset. The width represents the density of the data points at different values. The central line within each violin indicates the median. **e** Organ and lesion segmentation performance of UltraFedFM on different ratios of linear- and convex-array scanning mode ultrasound imaging data. The 95% CI of DSC is plotted in color bands, and the center points of the bands indicate the mean value of DSC.

outperforms most of the comparative models in all evaluation metrics and tasks. In classification tasks, for the fetal plane classification task, UltraFedFM achieves an accuracy of 0.956, significantly higher than cutting-edge models such as HCN[28] and Krishna et al.[29] ($p < 0.001$). On the gallbladder cancer classification dataset, it reaches an accuracy of 0.934, far surpassing classic models like ResNet[30] and Vision Transformer (ViT)[31]. Compared with the latest task-specific methods FocusMAE[32] and GBCHV[33], it shows improvements of 2.3% and 1.9% respectively ($p < 0.001$). In segmentation tasks, on the breast nodule segmentation dataset, UltraFedFM obtains a Dice coefficient of 0.887, greatly outperforming advanced methods such as FABRFnet[34] and EMGANet[35], and achieving performance comparable to the state-of-the-art method nnU-Net[36]. It is worth emphasizing that nnU-Net, as a standardized segmentation framework integrating data-adaptive processing and two-stage segmentation techniques, has achieved top-level performance in multiple segmentation tasks. On the thyroid nodule segmentation dataset, UltraFedFM achieves a Dice coefficient of 0.882, outperforming nnU-Net ($p < 0.01$).

UltraFedFM generalizes to new medical scenarios

Beyond learning ability, a crucial metric to evaluate the practicality of FMs in real-life scenarios is the generalization ability. To assess this, we selected two medical institutions not involved in the pre-training stage (high-frequency skin ultrasound imaging dataset and kidney disease ultrasound imaging dataset). This evaluation aims to determine how well the model performs on unseen ultrasound imaging modalities and organs, both key challenges in ultrasound diagnostics. As shown in Fig. 5b, c. UltraFedFM consistently demonstrates superior generalization across different modalities, achieving an average AUROC of 0.925, significantly outperforming all other baselines ($p < 0.01$). Figure 6c illustrates that UltraFedFM achieves an AUROC of 97.1% and an AP of 0.910, despite the textural and color differences of high-

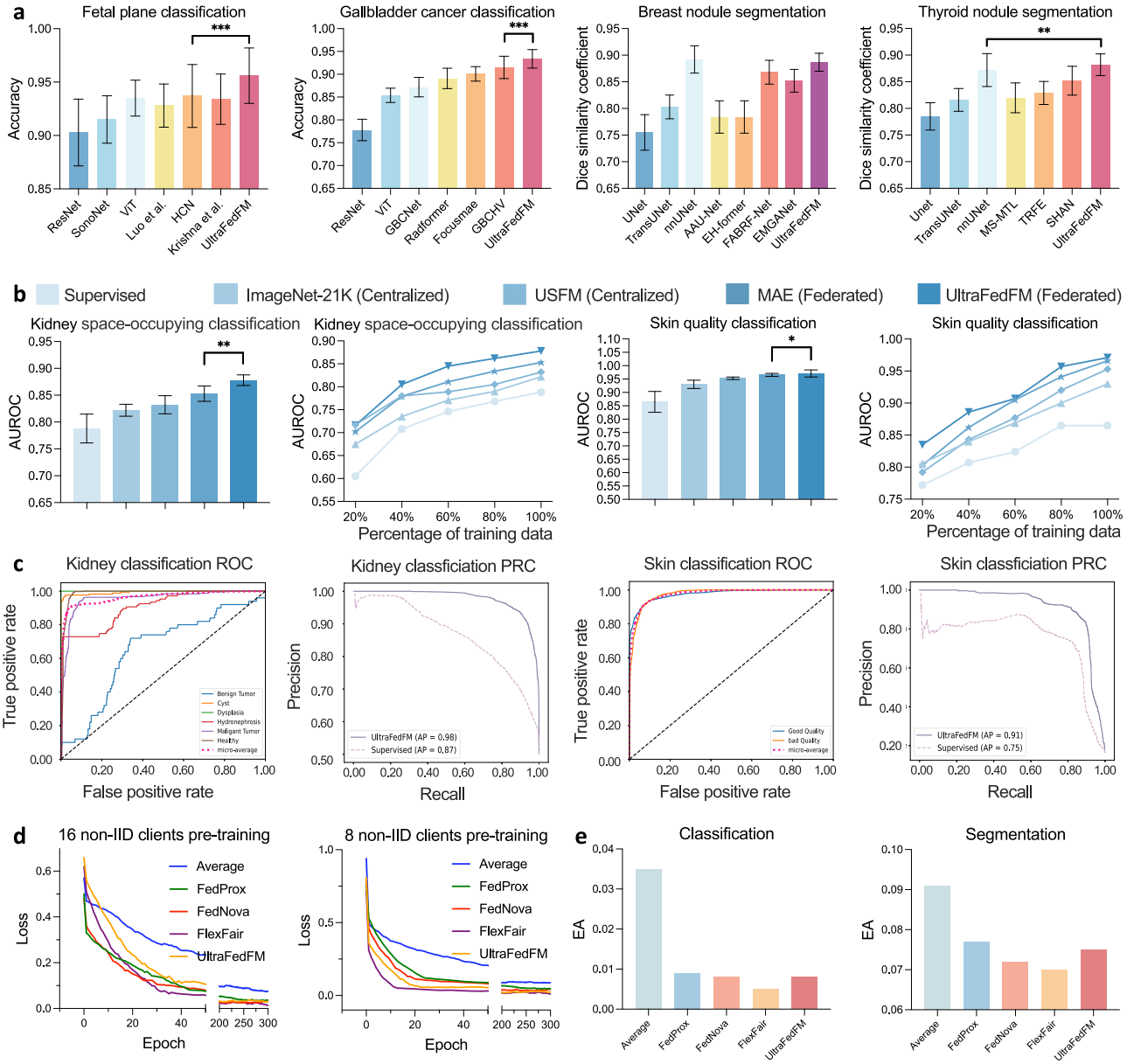

Fig. 5 | Comprehensive evaluation of ultraFedFM demonstrates high performance, strong generalization, and improved fairness. a The quantitative comparison between UltraFedFM and state-of-the-art ultrasound task-specific models across four typical ultrasound tasks. **b** Generalization performance evaluation on out-of-distribution organ (i.e., kidney) and modality (i.e., high-frequency ultrasound). **c** The receiver operating characteristic curve (ROC) and precision-recall curve (PRC) on a new organ and new ultrasound imaging modality. **d**, **e** The quantitative analysis of UltraFedFM with state-of-the-art federated learning methods in terms of pre-training convergence (**d**) and prediction fairness (**e**).

frequency ultrasound imaging from conventional methods. Such generalization is essential for real-world applications where clinicians frequently encounter new organs or modalities that the training data may not easily access.

The stability of UltraFedFM's predictions

The stability of model predictions is essential for ensuring reliable clinical decision-making, particularly in ultrasound-based diagnostics, where inconsistencies can lead to misdiagnosis. To this end, we quantitatively compared the prediction stability of UltraFedFM with the baseline USFM[16] under two settings: organ-specific (Fig. 7a) and organ-agnostic (Fig. 7b). USFM shows a broader distribution (mean $\mu = 0.808$ and standard deviation $\sigma = 0.174$). In contrast, UltraFedFM's predictions concentrate in a high DSC range (mean $\mu = 0.857$ and standard deviation $\sigma = 0.103$). Moreover, stability is paramount when dealing with organs that exhibit significant inter-patient variability, such as the liver or kidneys. Thus, we further introduce random noise to simulate real-world ultrasound imaging perturbations, such as tissue movement, operator variability, or imaging artifacts. We compared the test results under varying levels of noise. Despite these disturbances, UltraFedFM maintained highly correlated test scores (Fig. 7c), demonstrating its robustness and reliability in clinical environments where imaging conditions can be unpredictable.

The scaling efficiency in UltraFedFM

Figure 7e, f presents the scaling efficiency of UltraFedFM during pre-training, including data scaling (Fig. 7e) and model size scaling (Fig. 7f). Data scaling experiments were conducted using different proportions of pre-training data, while model size scaling involved pre-training with encoder architectures of varying parameter sizes (ViT-Base, ViT-Large, and ViT-Huge). In the data scaling experiments, we randomly sampled 10%,

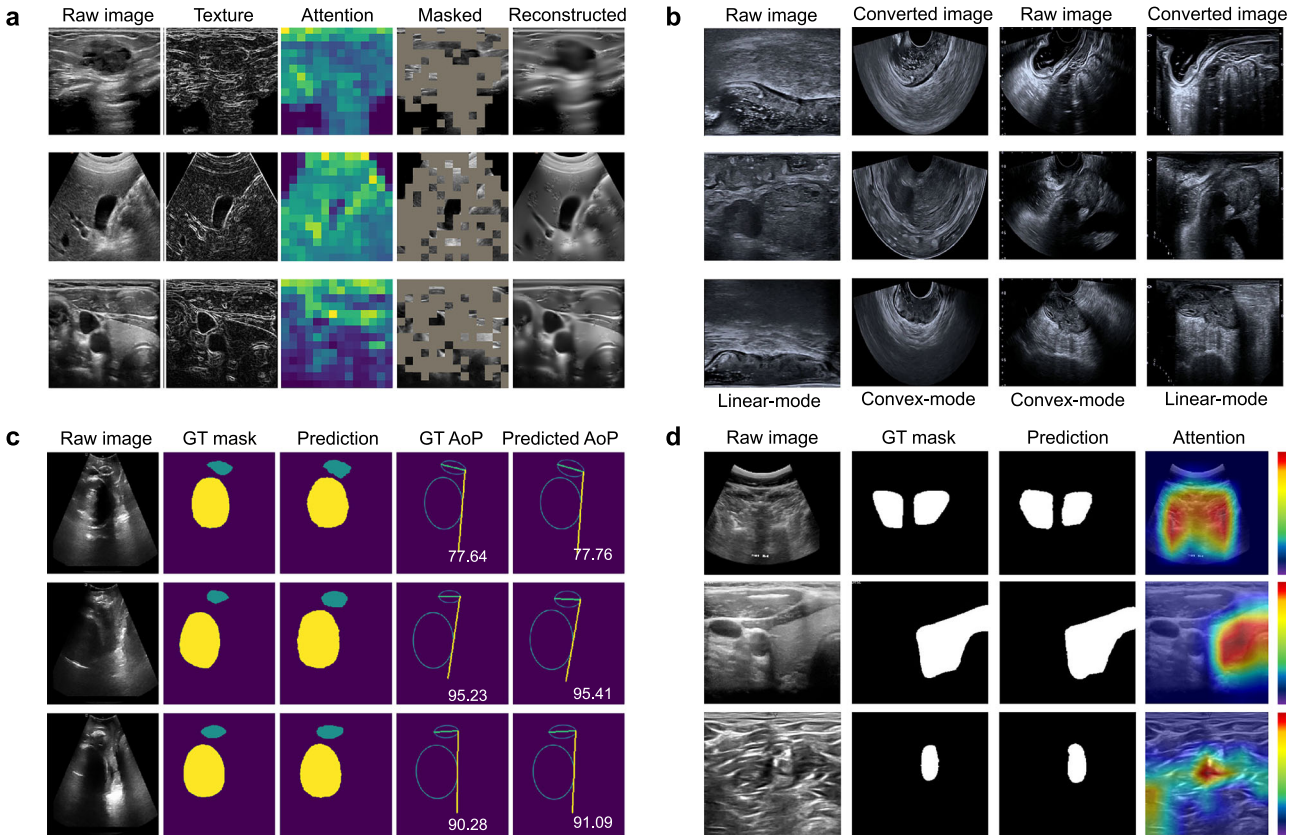

Fig. 6 | Visualization of ultraFedFM's pre-training, data augmentation, and segmentation performance. a The reconstructed ultrasound images from the pre-trained model, where the masked regions are selected based on texture information. **b** To increase the richness and balance of features, images captured in linear-array mode and convex-array mode are transformed into each other. **c** Visualization of multi-class organ segmentation and the prediction of the angle of progression (AoP). **d** Visualization of binary lesion segmentation. Heatmaps highlight the attention areas of the features extracted from the pre-trained encoder. The closer the color is to red, the more the model pays attention to the area.

20%, 50%, and 100% of pre-training data from each client and evaluated performance on eight downstream classification tasks and five downstream segmentation tasks. Overall, increasing the amount of pre-training data improved model performance, consistent with the data scaling principles in self-supervised learning (SSL). However, the growth trends varied slightly across different data modalities. Notably, segmentation tasks exhibited more pronounced performance gains, indicating that high-dimensional pixel-level prediction tasks are more sensitive to pre-trained feature learning. In the model size scaling experiments, we used three ViT variants as encoders to evaluate the impact of increasing the number of trainable parameters during pre-training on classification and segmentation tasks. For classification tasks, larger models generally yielded better performance across most modalities. Specifically, performance gains were more significant for challenging tasks (those with lower AUROC for ViT-Base), while simpler tasks reached a performance plateau, with ViT-Huge potentially introducing noise and overfitting risks. For segmentation tasks, increasing model size consistently improved performance, demonstrating that segmentation tasks demand larger model capacity and that model size scaling is particularly beneficial for addressing more challenging tasks.

Ablation studies validate the effectiveness of proposed strategies

To thoroughly evaluate the contribution of each proposed module, we implemented eight ablated variants of UltraFedFM by replacing individual components and evaluated their performance on both classification and segmentation tasks, as shown in Fig. 7d. Compared to the baseline, incorporating only SMAT (w/SMAT) yielded an increase of 5.5% AUROC and 3.0% DSC, while including only MIC (w/MIC) resulted in a more substantial improvement of 6.9% AUROC and 3.9% DSC. In contrast, removing specific components (i.e., w/o SMAT, w/o MIC, w/o TGM) degrades performance compared to the full UltraFedFM model. Notably, excluding MIC led to the highest decrease of 1.3% AUROC and 3.4% DSC, appearing to exert a particularly notable influence on the model's efficacy as evidenced by the relatively high performance when it is singularly included and the performance retention when other components are removed. To validate the effectiveness of our modified masked autoencoder (MAE) strategy in UltraFedFM, we compare it with several baseline self-supervised learning (SSL) strategies, including vanilla MAE[19], SimCLR[37], SwAV[38], DINO[39], and MoCo[40]. Figure 7g shows that UltraFedFM with the modified MAE significantly outperforms all other baselines ($p < 0.001$) in both disease diagnosis and lesion segmentation tasks. Specifically, UltraFedFM achieves the highest average AUROC of 0.926 across eight classification tasks and an average DSC score of 0.878 across four segmentation tasks. In contrast, the vanilla MAE achieves the second-best performance, with an average AUROC of 0.884 and an average DSC score of 0.839. These results suggest that MAE-based approaches are more effective for ultrasound imaging than contrastive learning-based methods. This success may be attributed to MAE's ability to learn robust feature representations in images where structures can vary significantly across patients or organs. Clinically, this translates to more accurate diagnostic predictions, especially for complex cases involving subtle lesions or challenging anatomical regions.

Qualitative model analysis and visualization

UltraFedFM's performance on downstream tasks depends on the representations learned during training. To investigate how these representations support downstream decisions, we qualitatively analyzed the internal mechanisms of the pre-text task of UltraFedFM during pre-training and how UltraFedFM made task-specific decisions on downstream tasks.

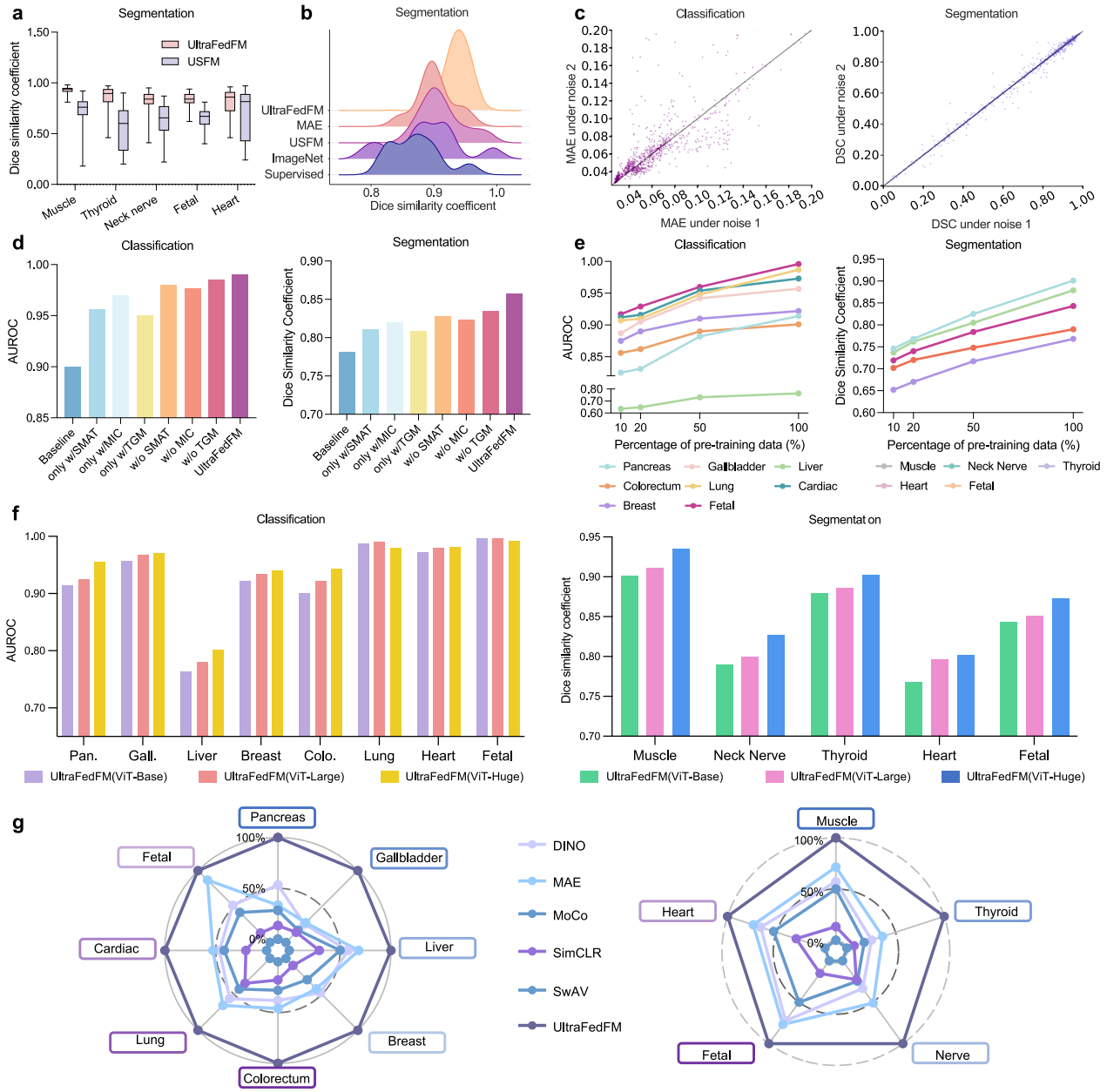

Fig. 7 | Comprehensive analysis of ultraFedFM highlights its performance, scalability, and robustness. a Comparison of the prediction distribution between UltraFedFM and USFM across five independent segmentation tasks. UltraFedFM's predictions are concentrated within a high dice similarity coefficient range (mean $\mu = 0.857$, standard deviation $\sigma = 0.103$), whereas USFM's predictions show greater dispersion (mean $\mu = 0.808$, standard deviation $\sigma = 0.174$) **b** Prediction distribution of all methods on organ-agnostic segmentation tasks. **c** The prediction stability of UltraFedFM under different ratios of input distribution variability. **d** The ablation study of proposed framework components. **e** The scaling effect of pre-training data, evaluated with different proportions of pre-training data. **f** The scaling effect of pre-training model size, evaluated using different ViT architectures (ViT-Base, ViT-Large, and ViT-Huge). **g** Performance impact of different self-supervised learning strategies on classification and segmentation tasks. All results are scaled and normalized relative to UltraFedFM. Specific quantitative results are available in Supplementary Table 6.

During pre-training, the pre-text task enables the model to learn ultrasound-specific context across various ultrasound imaging modalities. As shown in Fig. 6a, UltraFedFM accurately reconstructs images even when large portions are masked while preserving anatomical textures and lesion structures. This qualitative reconstruction behavior suggests the model learns context-aware features that reflect tissue and lesion morphology rather than merely memorizing low-level noise patterns. Figure 6b shows the results of scanning mode-aware transformation (SMAT) used in pre-training. Balancing scan modes reduces the tendency of UltraFedFM to overfit to a single probe configuration and promotes robustness across acquisition settings.

For downstream lesion segmentation tasks, Fig. 6c, d visualize UltraFedFM's precise localization of salient lesion areas and target boundaries. Clinically, the model's ability to focus on salient areas while excluding irrelevant background interference enhances its accuracy in detecting complex lesion structures, which is essential for diagnosing diseases with subtle or overlapping symptoms. Supplementary Fig. 1 illustrates the embedding feature space of different classes in the fine-tuned model. UltraFedFM demonstrates superior class discrimination, with different classes clearly separated in high-dimensional space, resulting in more precise classification boundaries. This ability is crucial in clinical applications where precise differentiation between pathological and non-pathological

tissue can impact treatment decisions. In contrast, the baseline supervised model exhibits a weaker differentiation and less distinct classification results. Supplementary Fig. 2a further shows how UltraFedFM effectively recognizes specific patterns and targets via the attention mechanism. For example, in pancreas, liver, breast, and gallbladder imaging, the model focuses on the center and surrounding areas of tumor lesions. In colorectal and lung imaging, it targets high-density textured regions while ignoring irrelevant hollow regions such as intestines and alveoli. For fetal ultrasound imaging, UltraFedFM focuses on the solid parts of the fetus and excludes irrelevant regions such as the uterus and muscles. In addition, we visualized the evolution of the attention map during pre-training (see Supplementary Fig. 2b). As pre-training progresses, the local model increasingly focuses on meaningful regions, thereby enhancing the effectiveness of the global model.

Discussion

With the growing demand for public health solutions, there is an urgent need to develop AI-based foundation models for wide application to real-world clinical scenarios. In this work, targeting the most widely used ultrasound data, we are the first to propose a comprehensive privacy-preserving ultrasound foundation model (USFM) using federated learning, namely UltraFedFM. By eliminating privacy concerns through decentralized pre-training, UltraFedFM leverages large-scale global datasets, enhancing its generalization capabilities. Regarding the above extensive experimental results, UltraFedFM demonstrates excellent performance, favorable generalization and robustness, and good adaptability to fine-tuning data. Specifically, it can handle different clinical tasks, such as diagnosing diseases, segmenting regions of interest (i.e., pathological tissues or organs), and analyzing spatial relationships of fetal organs, making it versatile for a wide range of medical applications. Moreover, UltraFedFM can be fine-tuned in a modality-agnostic manner to enable a single decoder to diagnose multiple diseases present in different modalities. Even with limited fine-tuning data, it consistently outperforms other baseline methods in both accuracy and stability. Across various ultrasound modalities and clinical tasks, UltraFedFM performs with judgment capabilities comparable to human clinicians.

Ultrasound imaging, a widely used clinical diagnostic tool, is renowned for its convenience and accuracy. Previous research in ultrasound diagnostics primarily focused on deep learning models trained on specific ultrasound modalities, targeting the diagnosis or segmentation of disease types within fixed imaging contexts. For example, Antropova et al.[41] developed a method that utilized pre-trained CNNs to extract and aggregate features, which were then combined with hand-crafted features from CADx for breast cancer diagnosis. Similarly, Basu et al.[7] investigated multi-scale and second-order pooling architectures to address false textures in gallbladder ultrasound, achieving precise localization and detection of malignant gallbladder tumors. Objective assessments and specialized segmentation strategies have been proposed by Yadav et al.[8–11] for thyroid ultrasound, highlighting the importance of robust segmentation and preprocessing in downstream classification. Comparative studies[8,14,15] on despeckling filters and image-quality metrics show that noise suppression and texture preservation strongly affect diagnostic performance. Several recent works also proposed deep-learning-based CAD systems with novel attention mechanisms for systematic disease diagnosis in ultrasound images. Jiang et al.[42] introduced a sparse computation and temporal fusion architecture designed for the accurate and real-time segmentation of colorectal cancer lesions. Yan et al.[13] constructed and validated a deep learning-based radiomics fusion model, enabling accurate identification of bone erosions in rheumatoid arthritis in musculoskeletal ultrasound. These studies have significantly advanced the application of AI in various ultrasound modalities, leading to improvements in automatic ultrasound image analysis, including lesion segmentation, disease diagnosis, and treatment planning. These successes are primarily attributed to the synergistic effects of data, models, and algorithms, specifically the collection and thorough annotation of datasets for specific organs or diseases, as well as specially designed network structures and training methods. However, a critical bottleneck limiting the advancement of medical imaging algorithms is the limited availability of fully annotated medical data. Medical image annotation demands the expertise of trained physicians, and the segmentation tasks in particular require substantial time and effort. Traditional methods were often trained on only hundreds or thousands of samples, significantly impacting the models' stability and generalizability in real-world applications. Meanwhile, with the evolution of ultrasound imaging technology, its application to an increasing number of organs and diseases presents challenges for models trained on single-organ or single-disease data, making it difficult to meet expanding clinical demands. There is a growing interest in developing a label-efficient ultrasound model that can be generalized across various tasks and organs, enabling rapid adaptation and deployment in clinical practice.

This led to the introduction of the foundation model (FM) based on self-supervised learning (SSL), which is capable of learning universal features independent of organs and diseases from unlabeled data. Prior to UltraFedFM, several studies explored FMs in medical imaging[43–47], covering various modalities such as ophthalmic images, endoscopy, and CT scans. In the field of ultrasound imaging, Christensen et al.[17] proposed an FM specifically for cardiac ultrasound, pre-trained on over one million echocardiogram videos for diagnosing various heart diseases. Jiao et al.[16] compiled and organized over two million ultrasound images across 12 different categories to establish a general USFM. However, these foundation models either focus on developing and applying to a single ultrasound modality, which is limited in actual clinical deployment or require centralized collection and processing of multi-center large-scale data. On the one hand, this approach requires expensive servers to store and process data, and on the other hand, the circulation of data inevitably involves the disclosure of patient information, especially rare disease data, which hinders the development of universal models.

UltraFedFM confirms that distributed pre-trained foundation models can match centralized ones by appropriate design. While UltraFedFM is not the first FM developed for ultrasound imaging, it is the first to integrate privacy protection during the model development process. Previous methods highlighted that, in most cases, private data held by different institutions is not shared, and public data must be anonymized to protect patient privacy, making it challenging to utilize a vast amount of available medical data. In this study, UltraFedFM breaks the privacy obstacles by leveraging federated learning that can use large inaccessible private data to train model distributively across sites without sharing sensitive information. UltraFedFM was pre-trained on over 1 million ultrasound images from 16 independent institutions worldwide, covering 19 different organs and 10 ultrasound modalities, which is 58.3% more in organ coverage and 66.7% more in ultrasound modalities compared with centralized baseline 3M-US, and therefore captures a substantially wide range of acquisition devices, operators and clinical scenarios. To address ultrasound-specific pre-training challenges, UltraFedFM developed the ultrasound image masking strategies, which explicitly accounts for probe-dependent texture characteristics and reduces failure modes observed when using generic MAE on ultrasound data. More importantly, the federated architecture also confers practical extensibility, where new private datasets can be incorporated incrementally to update the foundation model, so the utility of the model can increase overtime as more institutions join. Taken together, through careful selection of pre-training image quantity and the design of pre-training algorithms specific to ultrasound imaging, it has been demonstrated for the first time that a distributed pre-training foundation model can achieve comparable performance to centralized foundation models in overall performance across multiple downstream tasks. In ultrasound disease diagnosis tasks, UltraFedFM achieved an average AUROC of 0.927 across eight organs, significantly ($p < 0.001$) surpassing USFM's 0.894 AUROC. In ultrasound lesion segmentation tasks, UltraFedFM achieved an average DSC score of 0.876, significantly ($p < 0.001$) exceeding USFM's 0.858.

The advantages of UltraFedFM go beyond its performance metrics. The federated learning framework it employs provides unique benefits that address long-standing challenges in medical AI. First, it enables the model to

be continuously updatable, allowing further training and refinement on private datasets held by individual institutions or clients without exposing sensitive data. This ensures compliance with privacy regulations, such as GDPR and HIPAA, while maintaining the model's adaptability to evolving clinical needs. Additionally, federated learning facilitates participation from small-scale data contributors with uncommon organs or rare modalities (e.g., uterus, testis, contrast-enhanced ultrasound, high-frequency ultrasound). This fosters a collaborative ecosystem where diverse and distributed medical data can be leveraged to develop a comprehensive and generalizable model. These capabilities establish UltraFedFM as a solution capable of bridging the gap between data privacy and large-scale model training. The security of UltraFedFM during the pre-training process and its accuracy across various ultrasound imaging applications further solidify its potential for clinical translation. Previously, only large institutions with efficient data management workflows could develop foundation models from vast private medical datasets. This study demonstrates that federated learning enables the global medical community to collectively pre-train robust and generalizable foundation models. By addressing the dual imperatives of privacy preservation and model performance, UltraFedFM marks a paradigm shift in medical AI. Its ability to adapt to new data while protecting patient privacy sets a benchmark for the development of privacy-preserving AI in healthcare. Through its innovative use of federated learning, UltraFedFM demonstrates how global collaboration can unlock the potential of distributed medical data, paving the way for advancements in ultrasound AI and broader medical applications.

The core innovation of this study is the application and engineering of a large-scale, ultrasound-specific federated pre-training framework to simulate the distributed data distribution in real clinical environments. During the federated pre-training stage, the ultrasound image data of each participating institution showed significant differences in both sample size and modality type, leading to the typical non-independent and identically distributed (non-IID) characteristics of the data. The sensitive attribute distribution shown in Fig. 2c further confirms the systematic bias existing in the dataset, which poses a substantial challenge for cross-institutional model aggregation. To systematically evaluate the model's adaptability to non-independent and identically distributed (non-IID) data, we designed two sets of experiments focusing on convergence and fairness, respectively. For the convergence validation, we implemented two testing scenarios: the first using original non-IID training data from 16 different institutional clients for pre-training, while the second employed randomly sampled data from 8 institutions. UltraFedFM was compared against four federated learning approaches: Average (the baseline method without volume-based weighting strategy), along with state-of-the-art methods specifically designed for non-IID scenarios, including FedProx[48], FedNova[49], and FlexFair[50]. Experimental results in Fig. 5d demonstrated that UltraFedFM significantly outperformed the Average baseline in both scenarios. Specifically, in the 8-client experiment, FlexFair achieved the fastest convergence speed, followed by UltraFedFM, whereas in the more challenging 16-client experiment, UltraFedFM showed a comparable convergence speed to FedProx, both substantially surpassing the simple averaging method. These findings indicate that the uniform weighting approach of simple averaging causes deviation of the global objective function from its optimal direction, thereby reducing the convergence rate, while simultaneously confirming UltraFedFM's robustness in handling non-IID data. For fairness evaluation, we constructed simulated non-IID experimental environments. The classification task utilized three breast ultrasound datasets (BUS[51], BUS-BRA[52], and BUS-UCLM[53]) as clients, while the segmentation task employed three thyroid ultrasound datasets (DDTI[24], TG3k[25], and TN3k[26,54]). Each client dataset was randomly split into 80% training and 20% testing sets. Equal Accuracy (EA)[50] served as the fairness metric, measuring maximum prediction accuracy disparities across different groups (e.g., hospitals or age cohorts). Results in Fig. 5e showed FlexFair achieved optimal EA fairness performance in both tasks. UltraFedFM ranked second in classification fairness and performed comparably to FedProx in segmentation. Notably, the simple averaging approach performed poorly in both experiments, conclusively demonstrating the necessity of weighted aggregation strategies for non-IID data. Weighting by client data volume effectively prevents information dilution from large-data clients and substantially mitigates negative impacts from extreme data distributions.

While systematically evaluating the advantages of UltraFedFM in ultrasound imaging analysis, this study still has several limitations and unresolved challenges. Firstly, although currently we use federated learning to simulate distributed privacy-preserving training to avoid data information leakage between clients. However, the deployment is not in a fully decentralized clinical environment, and it cannot fully replicate the challenges, such as heterogeneous device differences, network communication delays, and dynamic client participation in a real multi-center scenario. This limitation further highlights the necessity of establishing an inter-institutional joint research network and promoting real distributed training. Secondly, although UltraFedFM has established a comprehensive ultrasound imaging benchmark encompassing 10 ultrasound modalities and 19 human organs, demonstrating outstanding performance in relevant tasks, its data diversity remains limited. Compared to the broader scope of clinical data, the benchmark lacks coverage in certain critical ultrasound imaging domains (e.g., rheumatoid arthritis ultrasound, ocular ultrasound, etc.), primarily due to the scarcity of publicly available datasets. This limitation hinders the full validation of UltraFedFM's generalization capability in these scenarios. Furthermore, the current version of clinical evaluation relies on a relatively small sample size and a limited number of participating physicians, making it difficult to comprehensively reflect the diversity and complexity of real-world clinical practice. Moving forward, we plan to collaborate extensively with clinical institutions to collect multicenter clinical samples covering mainstream ultrasound modalities and rare diseases, while inviting clinical experts from diverse institutions to participate in larger-scale double-blind evaluations. These efforts aim to ensure seamless integration of the model into clinical ultrasound workflows and strict compliance with clinical standards and practical requirements. Thirdly, while our simulated non-IID experiments indicate some robustness of the current volume-weighted federated learning setup, its effectiveness may decline as institutions and heterogeneity grow. Future research efforts should therefore explore targeted extensions with specific fairness designs, such as fairness-aware aggregation to balance overall per-client equity, and scalable federation to improve scalability in real-world deployments. Lastly, although UltraFedFM effectively addresses critical clinical diagnostic and segmentation tasks and has been validated across 12 different types of organs and diseases, it does not utilize the vast amount of textual diagnostic reports available in ultrasound examinations. Integrating multimodal features from both text and images could further improve the foundational model's accuracy in various clinical tasks and enable the development of additional clinically useful tasks like question-answering and diagnostic report generation.

In conclusion, this study provided a robust and reliable framework for developing comprehensive USFMs. It has been demonstrated that high-performance and stable models can be pre-trained without risking privacy leakage. This breakthrough can significantly advance the development of USFMs and potentially lead to the emergence of more powerful general-purpose medical AI. UltraFedFM has already shown excellent performance in various ultrasound clinical tasks and holds promise for further expansion. It has the potential to replace traditional ultrasound AI diagnostic models and play a crucial role in clinical decision support. The theoretical contributions of this research lie in validating the efficacy of federated learning for pre-training medical FMs, inspiring further academic advancements in the field. Practically, the implementation of UltraFedFM can enhance diagnostic accuracy, streamline clinical workflows, and improve patient outcomes. Additionally, these findings offer valuable insights for policy-making, particularly in the areas of data privacy and the integration of AI in healthcare. By ensuring data privacy and leveraging federated learning, the collective power of global medical data can be harnessed to drive innovations in medical diagnostics and treatment planning, ultimately transforming the healthcare delivery landscape.

Methods

Ultrasound dataset curation

UltraFedFM aims to perform universal ultrasound tasks for clinical applications. A crucial aspect of constructing such a model is the adaptability to diverse ultrasound imaging modalities and pathological conditions. To address this issue, we curated a large-scale pre-training dataset consisting of 1,015,754 unlabeled ultrasound images. In this dataset, 782,513 images were publicly available from multiple worldwide hospitals, while the remaining images were our privately owned ultrasound data. Our dataset covers a wide range of clinical ultrasound scenarios and modalities, including common abdominal, heart, fetal, superficial, musculoskeletal, and transvaginal ultrasound, as well as emerging techniques such as lung, endorectal, endoscopic, high-frequency, and contrast-enhanced ultrasound (Fig. 2a). Moreover, to realize federated pre-training, we split and arranged all datasets into 16 clients according to different hospitals. Supplementary Table 1 presents the data distributions of all clients. The multi-organ and multi-modality categories enable UltraFedFM to adapt to a wide range of clinical tasks.

The dataset contains a wide range of real sensitive attributes, and some of these attributes are biased. As illustrated in Fig. 2c. Specifically, at the national level, data from Chinese medical institutions accounts for the largest proportion at 57.30%, primarily because most private data comes from collaborative Chinese institutions. Followed by the United States at 15.94% and Germany at 10.14%, with all other countries comprising less than 20%. This imbalance may lead UltraFedFM to perform better on Chinese patient cases while introducing bias against underrepresented regions. In terms of geographical population distribution, the majority of cases come from Asia, followed by Europe and the Americas, reflecting imbalances in global healthcare resources. At the gender level, although the known gender ratio appears balanced, the large portion of samples with unspecified gender raises concerns about hidden gender bias. In terms of ultrasound modalities, mainstream modalities (such as abdominal, superficial, etc.) are relatively evenly distributed, while other modalities account for a smaller proportion due to their lower usage frequency. The data also primarily uses conventional linear scanning modes. Regarding the image quality of downstream fine-tuning data, nearly 50% of the data has varying degrees of quality issues, which helps us evaluate whether the model exhibits bias with low-quality data. In future work, in addition to expanding the geographical coverage and modality range of the data to minimize information bias introduced by the data, we also plan to explore the effectiveness of fairness-aware federated learning in improving dataset bias.

To further verify the practicality of UltraFedFM on clinically relevant tasks, we curated 15 well-annotated ultrasound datasets for validation. Two common clinical tasks were tested. The first task is ultrasound image diagnosis, which requires the FM to make accurate category judgments based on the organ and lesion information, ranging from two-class cancer recognition to multi-class disease diagnosis (Fig. 2b top). For this task, we utilized 7 publicly available datasets and 3 internal datasets for validation, which included a total of 10 organ categories and 8 ultrasound modalities. The second task is ultrasound image segmentation, which requires the FM to identify key organ/lesion areas and predict boundaries. For this task, we collected 5 public datasets for validation, which included a total of 5 organ categories and 4 ultrasound modalities. Note that small targets are more challenging for the model's prediction ability. In the validation datasets, 64.5% of the images contained targets that are less than 1/10 of the total image area (Fig. 2b bottom).

The overall ultrasound pre-training dataset and validation dataset breakdown are presented in Supplementary Table 1 and Supplementary 2.

Clinician cohort

To truly evaluate the reliability of UltraFedFM as an auxiliary tool in clinical scenarios, we invited multiple clinicians to participate in the evaluation. All participating clinicians hold a physician qualification certificate and a physician's practice certificate issued by the National Health Commission of China, and have completed subspecialty fellowship training in abdominal or musculoskeletal ultrasound. In addition, mid-level clinicians (with 4−8 years of clinical experience) hold a certificate of qualification for intermediate professional and technical positions (CQIPTP), and their professional titles are attending physicians. Expert-level doctors (with 10 years of clinical experience) hold a certificate of qualification for senior professional and technical positions (CQSPTP), and their professional titles are chief physicians or associate chief physicians. Additionally, expert-level clinicians handle an average of 7000 ± 200 ultrasound clinical cases per year (with an average of 240 working days), whereas mid-level clinicians (4–8 years, $n = 2$) average 4, 500 ± 300 clinical cases per year.

Federated pre-training framework

This work aims to collaborate with multiple clients to jointly train a robust FM without sharing their privacy-sensitive ultrasound data. To simulate the real clinical decentralized setting, we partitioned the pre-training dataset in K "virtual clients" to mimic independent institutions. Our simulation ensures raw data never leaves clients and assumes encrypted communication channels. Each client $k \in \{1, \cdots, K\}$ possesses a local dataset $\mathscr{D}_k$ with n_k data samples. The objective is to learn a global model, consisting of a global encoder $\mathtt{E}_g$ and a global decoder $\mathtt{D}_g$ towards minimizing the global loss function, which can be expressed as

$$L(\omega) = \frac{1}{n}\sum_{k=1}^{K} n_k L_k(\omega), \tag{1}$$

where ω is the overall model parameters of the global encoder $\mathtt{E}_g$ and global decoder $\mathtt{D}_g$, $n = \sum_{k=1}^{K} n_k$ is the total number of data of all clients, and $L_k(\omega)$ is the local loss function of client k that measures the local empirical loss on its local dataset $\mathscr{D}_k$. Then, the pre-training stage can be described as follows.

1. In the t-th communication round, the server broadcasts the global model ω^t to all clients.
2. Each client k takes E steps of gradient descent to update the local model based the received global model ω^t, as given by

$$\omega_k^{t+1} = \omega_k^t - \eta \nabla L_k(\omega^t), \tag{2}$$

where ω_k denotes the local model parameter of client k and η is the learning rate.

3. Each client uploads its local model parameter ω_k^{t+1} to the server.
4. The server aggregates the local models from all clients and updates the global model by

$$\omega^{t+1} = \frac{1}{n}\sum_{k=1}^{K} n_k \omega_k^{t+1}. \tag{3}$$

The combination of the four steps is referred to as one communication round. The pre-training process terminates once it reaches a pre-defined number of communication rounds T.

After completing pre-training, the pre-trained global encoder was saved while the global decoder was discarded. In the fine-tuning stage, the global encoder generated high-level features from the ultrasound images. A multi-layer perceptron (MLP) takes these features as input and outputs the probabilities for disease categories. The category with the highest probability was defined as the classification result. The fine-tuning objective was to produce classification results that match the ground-truth labels. After each epoch, the model was evaluated on the validation set. The model weights with the highest accuracy on the validation set were saved as checkpoints for internal and external evaluations.

Local model architecture

As illustrated in Supplementary Fig. 3, we employed a masked autoencoder (MAE) as the local model of each client k, which consists of an encoder $\mathtt{E}_k$ and decoder $\mathtt{D}_k$.

For each image $I \in \mathbb{R}^{H\times W\times C}$ sampled from the local dataset, the corresponding local model undergoes pre-training through an ultrasound masked image modeling (UltraMIM) process, including an ultrasound image masking (UIM) stage and a reconstruction stage, which are respectively defined by

$$\mathscr{P} = \mathtt{UIM}_k(I), \tag{4}$$

$$I^{recon} = \mathtt{D}_k(\mathtt{E}_k(\mathscr{P})), \tag{5}$$

where $\mathscr{P} = \{\mathbf{p}^1, \cdots, \mathbf{p}^L\}$ is the patch set, $\mathbf{p}^\ell \in \mathbb{R}^N$ represents the ℓ-th patch, L denotes the total number of patches, and $\mathtt{UIM}$ denotes the ultrasound image masking operation. In particular, there exist two types of patches in the $\mathscr{P}$, i.e., masked patches $\mathscr{P}_m$ and visible patches $\mathscr{P}_v$.

We utilized Vision Transformer (ViT) as the encoder $\mathtt{E}_k$ and applied it to a sequence of unmasked image patches. Specifically, the encoder compressed the input visible patches $\mathscr{P}_v$ into the latent representation, denoted by $\mathscr{P}'_v$. The latent representation captures the essential features of the input images, allowing the model to learn meaningful patterns regardless of masking. Then, we concatenated $\mathscr{P}'_v$ and $\mathscr{P}_m$ to obtain the overall patch set, defined as $\mathscr{P}^{all} = \mathtt{Concat}(\mathscr{P}'_v, \mathscr{P}_m)$. After that, the overall patch set $\mathscr{P}^{all}$ was passed through the decoder $\mathtt{D}_k$, which is a lightweight ViT that aims to reconstruct the input image by predicting the pixel values of the masked patches. Finally, the output of the decoder was reshaped to obtain a reconstructed image. We note that the reconstruction process enables the model to learn the underlying structure and patterns of the ultrasound images. Without loss of generality, we adopted the mean square error (MSE) as the local loss function to measure the discrepancy between the original image and the reconstructed image, which is defined as

$$L_k(\omega) = \sum_{1\le i\le N_m} \frac{1}{N_m}\left(x^i - \hat{x}^i\right)^2, \tag{6}$$

where N_m is the number of masked patches, while x and $\hat{x}$ are the ground-truth and predicted pixel values of each masked patch.

Ultrasound image masking

In clinical scenarios, different medical institutions may use different probes and equipment to collect ultrasound images. As a result, there exists significant variability in the imaged organs and lesions. To enhance the generalization ability of UltraFedFM across diverse clinical scenarios, we proposed the UIM composed of three modules, including scanning mode-aware transformation (SMAT), mixed image corruption (MIC), and texture-guided masking (TGM), which adaptively adjust the pre-training process based on the characteristics of the local dataset. Specifically, the working mechanism of UIM at client k can be expressed as

$$\mathscr{D}_k^{trans} = \mathtt{SMAT}(\mathscr{D}_k), \tag{7}$$

$$\mathscr{D}_k^{total} = \mathtt{Concat}(\mathscr{D}_k^{trans}, \mathscr{D}_k), \tag{8}$$

$$I^{corru} = \mathtt{MIC}(I, p), \tag{9}$$

$$\mathscr{P} = \mathtt{TGM}(I^{corru}), \tag{10}$$

where $\mathscr{D}_k^{trans}$ is the transformed dataset, $\mathscr{D}_k^{total}$ is the overall dataset for pretraining, and I^{corru} is the corrupted input image.

Scanning mode-aware transformation. To ensure the generalization ability of UltraFedFM to images acquired under different scanning modes, we introduced a data augmentation method called Scanning Mode-Aware Transformation (SMAT). Typically, ultrasound images are collected using convex-array or linear-array probes, each with distinct geometrical properties. Therefore, we leveraged the coordinate mapping relationship between the two modes through Polar-Cartesian transformations[55].

Supplementary Fig. 4a shows the detailed process of SMAT. We first established the Polar coordinates for the convex-array mode image using the origin point (r_o, θ_o) set at the top center of the image and the x-axis at the top edge. Meanwhile, the Cartesian coordinates are set using (x_o, y_o) as the origin point, the top edge as the x-axis, and the vertical central axis as the y-axis. In Cartesian coordinates, we denote (x_1, y_1) as the point in the original image, (x_2, y_2) as the point in the transformed image. In polar system, (r_1, θ_1) is the point in the original image, (r_2, θ_2) is the point in the transformed image. For the convex-array mode, we transform the Polar coordinate into the Cartesian coordinate. Specifically, we obtained the value and position of each pixel (x_2, y_2) according to its corresponding point (r_1, θ_1) on the original image:

$$r_1 = \sqrt{x_1^2 + y_1^2}, \tag{11}$$

$$\theta_1 = \arctan\frac{y_1}{x_1}, \tag{12}$$

$$I^{trans}[x_2, y_2] = f(x_o + r_1\cos\theta_1,\ y_o + r_1\sin\theta_1), \tag{13}$$

where I^{trans} denotes the transformed image, f is a function for getting the information of the corresponding pixel points in the input image.

For the linear-array mode, the transformed position (r, θ) of the convex-array mode image is calculated using the Cartesian-to-Polar transformation, which can be expressed as

$$r_2 = \sqrt{x_1^2 + y_1^2}, \tag{14}$$

$$\theta_2 = \arctan\frac{y_1}{x_1}, \tag{15}$$

$$I^{trans}[x_2, y_2] = f(x_o + r_2\cos\theta_2,\ y_o + r_2\sin\theta_2). \tag{16}$$

After image transformation, we then concatenated the transformed dataset and the original dataset to obtain an enhanced pre-training dataset, i.e., $\mathscr{D}_k^{total}$. Note that the number of images in $\mathscr{D}_k^{total}$ is balanced across different scanning modes. Therefore, the risk of model over-fitting to any particular mode is significantly reduced.

Mixed image corruption. To ensure that UltraFedFM is robust to both low-quality and high-quality images, we introduced an additional de-corruption branch in the pre-training process, as shown in Supplementary Fig. 4b. The image corruption operations were inspired by three common cases encountered in clinical practice, i.e., motion blur, low resolution, and random noise.

1. Motion blur arises from the swift movement of the ultrasound probe, leading to artifacts and image distortion. We simulated this effect by convolving the original image with a motion blur kernel $\mathbf{K}(d, \phi)$ that accounts for both the degree of blur d and the angle of motion ϕ. Specifically, we first constructed an identity matrix $\mathbf{U} \in \mathbb{R}^{d\times d}$. Moreover, we defined a rotation matrix $\mathbf{R}(\phi)$ with respect to the motion angle ϕ, as given by

$$\mathbf{R}(\phi) = \begin{bmatrix} cos(\phi) & -sin(\phi) \\ sin(\phi) & cos(\phi) \end{bmatrix}. \tag{17}$$

Then, by applying the affine transformation, the motion blur kernel $\mathbf{K}(d, \phi)$ is derived as

$$\mathbf{K}(d, \phi) = \frac{1}{d} \cdot \mathbf{U} \cdot \mathbf{R}(\phi), \tag{18}$$

where each element in $\mathbf{K}(d, \phi)$ was normalized by the degree of blur d to ensure that the kernel maintains the same intensity as the original image. Finally, the motion-blurred image can be obtained by convolving the original image with the motion kernel, which is expressed as

$$I^{corru} = I * \mathbf{K}(d, \phi), \tag{19}$$

where * denotes the convolution operation.

2. Low resolution makes it hard to distinguish the critical parts of images. It can be achieved by a simple Gaussian blur operation. Gaussian blur involves convolving an image with a Gaussian kernel $G(\sigma)$, which is a two-dimensional Gaussian function with a standard deviation σ. Mathematically, the Gaussian blur operation can be expressed as

$$\begin{aligned} I^{corru} &= \texttt{Gaussian}(I, \sigma) \\ &= I * G(u, v; \sigma) \\ &= I * \tfrac{1}{2\pi\sigma^2} e^{-(u^2+v^2)/2\sigma^2}, \end{aligned} \tag{20}$$

where u is the distance to the origin in the horizontal axis, v is the distance to the origin in the vertical axis.

3. Random noise is the disturbance of pixel values, which is caused by the aging of components in old equipment. Random noise often destroys the appearance features of the target and affects the doctor's judgment. In this work, we simulated salt-and-pepper noise by the random occurrence of black and white pixels in an image. Given a grayscale image $I(x, y)$ where (x, y) denotes the pixel coordinates. The corrupted image is given by

$$I^{corru}(x, y) = \begin{cases} 0, & \text{with probability } p_s, \\ 255, & \text{with probability } p_p, \\ I(x, y), & \text{with probability } 1 - p_s - p_p, \end{cases} \tag{21}$$

where p_s is the probability of a pixel being set to the minimum intensity value ("0" corresponds to "salt"), and p_p is the probability of a pixel being set to the maximum intensity value ("255" corresponds to "pepper").

The above three image corruption transformations can be combined to constitute a variety of composite transformations. We randomly selected one, two, or three operations from the triplet of [motion blur, gaussian blur, random noise] according to probability p to form a composite operation.

Texture-guided masking. In ultrasound images, the edge prior reveals the sharpness of local regions and contains high anatomical information. Therefore, we quantified the edge information of each image patch to measure the texture complexity. The process is illustrated in Supplementary Fig. 4c. Specifically, given an ultrasound image I with the spatial size of $H \times W$, we computed the texture map $I^{texture} \in \mathbb{R}^{H \times W \times 1}$ based on the edge information of I using a second-order Laplacian differential operator[56], as defined by

$$I^{texture} = \frac{\partial^2 I}{\partial x^2} + \frac{\partial^2 I}{\partial y^2}, \tag{22}$$

where x and y are the indices of the image I. Then we split $I^{texture}$ into a series of texture patches, denoted by a set $\mathscr{P}_t = \{\mathbf{p}_t^1, \cdots, \mathbf{p}_t^L\}$. Here, $\mathbf{p}_t^\ell \in \mathbb{R}^{h \times w \times 1}$ is the ℓ-th texture patch with spatial size of $h \times w$. For each texture patch, the texture complexity score was calculated by summing the absolute values of all elements, which is given by

$$\text{Score}_t^l = \sum_{i=1}^{h} \sum_{j=1}^{w} \| \mathbf{p}_t^l(i, j) \|, \tag{23}$$

where Score_t^l is the score of the l-th texture patch and (i, j) denotes the position of the texture patch. By doing so, we can obtain the scores for all texture patches, denoted by $[\text{Score}_t^1, \text{Score}_t^2, \ldots, \text{Score}_t^L]$. Then, we generated the texture attention mask by concatenating the score of all texture patches, as given by

$$\mathbf{A} = \texttt{Concat}(\text{Score}_t^1, \text{Score}_t^2, \ldots, \text{Score}_t^L). \tag{24}$$

Given the attention mask $\mathbf{A} \in \mathbb{R}^L$, the texture patches with higher weights are more likely to be the foreground critical objects and may contain more information than those with lower weights. Therefore, we raised the masking probabilities of texture patches with high weights and used the texture patches with low weights as the visible hints in the masked image modeling process. More precisely, we first sorted the values in $\mathbf{A}$ from largest to smallest. Then, we guided the token selection process to generate the masked patch set $\mathscr{P}$. The tokens with the top M highest probabilities were discarded, while the remaining tokens were preserved as visible hints for masked image modeling. Here, we set $M = 75\%$ to keep consistent with the original MAE settings.

Self-supervised learning implementation

For comparison purposes, we replaced the MAE[19] in UltraFedFM, with other SSL methods, including SimCLR[37], SwAV[38], DINO[39], and MoCo-v3[40] to generate different pre-trained models. For each SSL method, we followed the network architectures and hyperparameter settings recommended in the literature to achieve the optimal performance. First, we loaded the pre-trained weights on ImageNet-1k into the models. Subsequently, we trained the models using the ultrasound pre-training dataset with each SSL method to obtain the pre-trained models. Following the same process for UltraFedFM, we transferred the MAEs to downstream disease detection tasks and fine-tuned these pre-trained models.

Baseline methods implementation

We compared the performance of UltraFedFM with 4 pre-trained comparison models: Supervised, ImageNet-21K (centralized), USFM (centralized), and MAE (federated). "Supervise" uses the supervised learning strategy with a randomly initialized ViT encoder. "ImageNet-21K" uses the transfer learning strategy, where the model is centralized pre-trained on ImageNet-21K (about 14 million natural images with classification labels) through self-supervised learning. "USFM" uses the Universal Ultrasound Foundation Model[16] to perform centralized pre-training on the 3M-US dataset (about 2 million ultrasound images of 12 organs). MAE and UltraFedFM use the same ultrasound dataset for federated pre-training, but MAE uses the original masked image modeling algorithm as a control to observe the advantages of our newly designed modules. All methods are fine-tuned on the same downstream datasets using the same experimental settings until convergence.

Performance metrics

To evaluate the performance of all disease diagnosis tasks, we utilized four widely employed metrics, including accuracy, F1-score, AUROC, and recall. Accuracy is a fundamental metric in classification tasks, which is defined as the ratio of correctly classified instances to the total number of instances. Mathematically, it is expressed as

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{25}$$

where TP, TN, FP, and FN represent the true positives, true negatives, false positives, and false negatives, respectively.

F1-score is often used to evaluate the performance of classification models, particularly in scenarios where the data distribution is imbalanced. It is defined as the harmonic mean of precision and recall, offering a balance between the two metrics. F1-score is mathematically expressed as

$$\mathrm{F1-score} = 2 \cdot \frac{\mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}, \tag{26}$$

$$\mathrm{Precision} = \frac{TP}{TP + FP}, \tag{27}$$

$$\mathrm{Recall} = \frac{TP}{TP + FN}. \tag{28}$$

AUROC is utilized to assess the performance of binary classification models, particularly in distinguishing between positive and negative classes across various threshold settings. Typically, AUROC is defined as the area under the receiver operating characteristic curve, which plots the true positive rate (TPR) against the false positive rate (FPR) as the decision threshold is varied. Mathematically, AUROC is expressed as

$$\mathrm{AUROC} = \int_0^1 \mathrm{TPR}(t) d\mathrm{FPR}(t), \tag{29}$$

$$\mathrm{TPR} = \frac{TP}{TP + FN}, \tag{30}$$

$$\mathrm{FPR} = \frac{FP}{FP + TN}. \tag{31}$$

For binary and multi-class lesion segmentation tasks, we utilized the dice similarity coefficient (DSC) for evaluation. DSC is useful for assessing the accuracy of a model in scenarios where spatial overlap between predicted and true masks is of primary interest. It is defined as

$$\mathrm{DSC} = \frac{2 \cdot |P \cap T|}{|P| + |T|}, \tag{32}$$

where P represents the predicted segmentation and T represents the ground truth.

In addition to DSC, we used the Hausdorff distance (HD) to evaluate the binary segmentation models. HD is a critical metric for image segmentation tasks, which measures the maximum discrepancy between the boundaries of the predicted segmentation and the ground truth. It evaluates the worst-case scenario by identifying the greatest distance from any point on one boundary to the closest point on the other boundary. Mathematically, the HD between two prediction segmentation P and ground truth T is defined as

$$\mathrm{HD} = \max\left\{ \max_{p \in P}\{\min_{t \in T}\}||p - t||, \max_{t \in T}\{\min_{p \in P}\}||t - p|| \right\}. \tag{33}$$

For pubic symphysis-fetal head tasks, the measurement of the angle of progression was conducted by constructing two lines from three specific landmarks. Firstly, we identified the two furthest points on the pubic symphysis contour based on the segmented image. Then, we drew a tangent line through the rightmost point of the pubic symphysis to define the fetal head region. The tangent line on the right side of the image, intersecting with the fetal head region, determines the third point for calculating the angle of progression (AoP). Finally, the angle formed by these three points constituted AoP. The performance of AoP prediction was evaluated using the mean absolute error (MAE). MAE provides a straightforward interpretation of how far, on average, the predicted values deviate from the actual observed values, which is given by

$$\mathrm{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|, \tag{34}$$

where n is the number of samples, y_i is the actual value for the i-th sample, $\hat{y}_i$ is the predicted value for i-th sample.

To evaluate the prediction fairness of the method, we adopted equal accuracy (EA) as the fairness evaluation metric, which measures the maximum gap in prediction accuracy between different groups (e.g., different hospitals, age groups):

$$EA = \max_k |\mathrm{Score}\left(\mathscr{A}_k\right) - \overline{\mathrm{Score}}|. \tag{35}$$

Here, A_k represents the k-th client, Score (A_k) denotes the AUROC score of the test set for client k, and $\overline{\mathrm{Score}}$ is the average score across all clients. Minimizing EA (i.e., narrowing the performance gap between groups) indicates achieving maximum fairness.

The models for all tasks were trained using five different random seeds to determine the shuffling of the training data. We calculated the mean and standard deviation of the performance over the three iterations and computed the standard error, i.e., the standard deviation divided by the square root of 5. We obtained the 95% confidence interval (CI) by multiplying the standard error by 1.96. Moreover, to determine whether there were significant differences, we performed two-sided t-tests between the significance of UltraFedFM compared to other methods.

Implementation details

In the pre-training stage, the image encoder of UltraFedFM is implemented by a basic vision Transformer33 (ViT-base) with 12 Transformer blocks and an embedding vector size of 768, whereas the decoder is a small vision Transformer (ViT-small) with 8 Transformer blocks and an embedding vector size of 512. The masking ratio is configured to 0.75, with an input size of 224×224. The model was pre-trained for 600 communication rounds (epochs) with a batch size of 512, and the warm-up period is 60 epochs. The local model was trained with 1 epoch in each communication round. We employed the AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.95$, an initial learning rate of $1.5e-4$, and a weight decay of 0.05. For the fine-tuning stage, the input images are resized to 224×224, and Random rotations, flips, and crops were used as data augmentation. The batch size was set as 16. All models were trained with 100 epochs using the AdamW optimizer with a cosine learning rate scheduler. The first 10 epochs were used for learning rate warm-up. The drop-path probability was set to 0.1. The detailed configurations are listed in the Supplementary Table 4 and Supplementary Table 5.

Computing hardware and software

We used Python (version 3.7.4) for all experiments and analyses in the study, which can be replicated using open-source libraries as outlined below. For pre-training, we used 8 32-GB NVIDIA GeForce Tesla V100 GPUs configured for multi-GPU training using DistributedDataParallel (DDP) as implemented by the framework PyTorch (version 1.11.0, CUDA 11.3). For fine-tuning, we used 1 32-GB NVIDIA GeForce Tesla V100 GPU. Pillow library (version 9.5.0) and opencv-python (version 4.7.0) libraries were used to read images, which were then converted to the base64 string format using Python. Timm library (version 0.9.2), torchvision (version 0.12.0) and opencv-python were applied for image processing and loading during training. Einops library (version 0.6.1) was applied for tensor operations in modeling. For model evaluation, we use the torchmetrics library (version 1.3.2) and pycm library (4.0) for classification task evaluations, and the segmentation-models-pytorch library (version 0.3.3) for segmentation task evaluations. Numpy (version 1.23.2) and Pandas (version 2.2.2), were used in data collection, preprocessing and data analysis.

Data availability

The publicly available dataset for pre-training can be accessed from: BUV (\href{https://github.com/jhl-Det/CVA-Net/tree/main}{https://github.com/jhl-Det/CVA-Net/tree/main}), CLUST (\href{https://clust.ethz.ch/data.html}{https://clust.ethz.ch/data.html}), EchoNet-Dynamic (\href{https://echonet.github.io/dynamic/}{https://echonet.github.io/dynamic/}), FETAL-PLANES (\href{https://zenodo.org/records/3904280}{https://zenodo.org/records/3904280}), TDSC-ABUS (\href{https://tdsc-abus2023.grand-challenge.org/Dataset/}{https://tdsc-abus2023.grand-challenge.org/Dataset/}), Leg-3D-US (\href{https://www.cs.cit.tum.de/camp/publications/leg-3d-us-dataset/}{https://www.cs.cit.tum.de/camp/publications/leg-3d-us-dataset/}), Thyroid Ultrasound Cine-clip (\href{https://stanfordaimi.azurewebsites.net/datasets/a72f2b02-7b53-4c5d-963c-d7253220bfd5}{https://stanfordaimi.azurewebsites.net/datasets/a72f2b02-7b53-4c5d-963c-d7253220bfd5}), SYSU-FLL-CEUS (\href{https://github.com/lemondan/Focal-liver-lesions-dataset-in-CEUS}{https://github.com/lemondan/Focal-liver-lesions-dataset-in-CEUS}), CAMUS (\href{https://www.creatis.insa-lyon.fr/Challenge/camus/index.html}{https://www.creatis.insa-lyon.fr/Challenge/camus/index.html}), COVID-BLUES (\href{https://github.com/NinaWie/COVID-BLUES}{https://github.com/NinaWie/COVID-BLUES}), NerveUS (\href{https://www.kaggle.com/competitions/ultrasound-nerve-segmentation}{https://www.kaggle.com/competitions/ultrasound-nerve-segmentation}), LEPset (\href{https://zenodo.org/records/8041285}{https://zenodo.org/records/8041285}), FPUS (\href{https://github.com/bharathprabakaran/FPUS23?tab=readme-ov-file}{https://github.com/bharathprabakaran/FPUS23?tab=readme-ov-file}), GBUSV (\href{https://github.com/sbasu276/FocusMAE}{https://github.com/sbasu276/FocusMAE}). The publicly available datasets for downstream tasks can be accessed from: LEPset-labeled (\href{https://zenodo.org/records/8041285}{https://zenodo.org/records/8041285}), SYSU-FLL-CEUS-labeled (\href{https://github.com/lemondan/Focal-liver-lesions-dataset-in-CEUS}{https://github.com/lemondan/Focal-liver-lesions-dataset-in-CEUS}), GBCU (\href{https://gbc-iitd.github.io/data/gbcu}{https://gbc-iitd.github.io/data/gbcu}), BUSI (\href{https://www.kaggle.com/datasets/aryashah2k/breast-ultrasound-images-dataset}{https://www.kaggle.com/datasets/aryashah2k/breast-ultrasound-images-dataset}), BUV-labeled (\href{https://github.com/jhl-Det/CVA-Net/tree/main}{https://github.com/jhl-Det/CVA-Net/tree/main}), BUS-BRA (\href{https://zenodo.org/records/8231412}{https://zenodo.org/records/8231412}), BUS-UCLM (\href{https://data.mendeley.com/datasets/7fvgj4jsp7/3}{https://data.mendeley.com/datasets/7fvgj4jsp7/3}), POCUS (\href{https://github.com/jannisborn/covid19_ultrasound}{https://github.com/jannisborn/covid19ultrasound}), FETAL-PLANES-labeled (\href{https://zenodo.org/records/3904280}{https://zenodo.org/records/3904280}), HFUS (\href{https://data.mendeley.com/datasets/td8r3ty79b/1}{https://data.mendeley.com/datasets/td8r3ty79b/1}), NerveUS-labeled (\href{https://www.kaggle.com/competitions/ultrasound-nerve-segmentation}{https://www.kaggle.com/competitions/ultrasound-nerve-segmentation}), DDTI (\href{https://www.kaggle.com/datasets/dasmehdixtr/ddti-thyroid-ultrasound-images}{https://www.kaggle.com/datasets/dasmehdixtr/ddti-thyroid-ultrasound-images}), Thyroid Ultrasound Cine-clip labeled (\href{https://stanfordaimi.azurewebsites.net/datasets/a72f2b02-7b53-4c5d-963c-d7253220bfd5}{https://stanfordaimi.azurewebsites.net/datasets/a72f2b02-7b53-4c5d-963c-d7253220bfd5}), TG3k (\href{https://github.com/haifangong/TRFE-Net-for-thyroid-nodule-segmentation}{https://github.com/haifangong/TRFE-Net-for-thyroid-nodule-segmentation}), TN3k (\href{https://github.com/haifangong/TRFE-Net-for-thyroid-nodule-segmentation}{https://github.com/haifangong/TRFE-Net-for-thyroid-nodule-segmentation}), LUMINOUS (\href{https://users.encs.concordia.ca/~impact/luminous-database/}{https://users.encs.concordia.ca/~impact/luminous-database/}), CardiacUDA (\href{https://www.kaggle.com/datasets/xiaoweixumedicalai/cardiacudc-dataset}{https://www.kaggle.com/datasets/xiaoweixumedicalai/cardiacudc-dataset}), JNU-IFM (\href{https://figshare.com/articles/dataset/JNU-IFM/14371652}{https://figshare.com/articles/dataset/JNU-IFM/14371652}). The UltraFedFM private dataset consists of routinely collected healthcare data. Owing to its sensitive nature and the risk of reidentification, the dataset is subject to controlled access by means of a structured application process. Data access enquiries may be made by \href{https://forms.gle/sdS5uX5FJjFRcr74A}{Google form}. We will review and aim to respond in a few weeks. The pre-trained and fine-tined models, as well as source code for pre-training, fine-tuning, inference, and data preprocessing, can be accessed at \href{https://github.com/yuncheng97/UltraFedFM}{https://github.com/yuncheng97/UltraFedFM}

Received: 9 May 2025; Accepted: 11 October 2025;
Published online: 21 November 2025

References

1. Whitworth, M., Bricker, L. & Mullan, C. Ultrasound for fetal assessment in early pregnancy. *Cochrane Database Syst. Rev.* **14**, CD007058 (2015).
2. Akkus, Z. et al. A survey of deep-learning applications in ultrasound: artificial intelligence–powered ultrasound for improving clinical workflow. *J. Am. College Radiol.* **16**, 1318–1328 (2019).
3. Donofrio, M. T. et al. Diagnosis and treatment of fetal cardiac disease: a scientific statement from the american heart association. *Circulation* **129**, 2183–2242 (2014).
4. Feldman, M. K., Katyal, S. & Blackwood, M. S. Us artifacts. *Radiographics* **29**, 1179–1189 (2009).
5. Lin, Z. et al. A new dataset and a baseline model for breast lesion detection in ultrasound videos. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 614–623 (Springer, 2022).
6. Burgos-Artizzu, X. P. et al. Evaluation of deep convolutional neural networks for automatic classification of common maternal fetal ultrasound planes. *Sci. Rep.* **10**, 10200 (2020).
7. Basu, S., Gupta, M., Rana, P., Gupta, P. & Arora, C. Surpassing the human accuracy: Gallbladder cancer detection from usg with curriculum learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 20854–20864 (IEEE, 2022).
8. Yadav, N., Dass, R. & Virmani, J. Despeckling filters applied to thyroid ultrasound images: a comparative analysis. *Multimedia Tools Appl.* **81**, 8905–8937 (2022).
9. Yadav, N., Dass, R. & Virmani, J. Objective assessment of segmentation models for thyroid ultrasound images. *J. Ultrasound* **26**, 673–685 (2023).
10. Yadav, N., Dass, R. & Virmani, J. Deep learning-based cad system design for thyroid tumor characterization using ultrasound images. *Multimedia Tools Appl.* **83**, 43071–43113 (2024).
11. Yadav, N., Dass, R. & Virmani, J. Machine learning-based cad system for thyroid tumour characterisation using ultrasound images. *Int. J. Med. Eng. Inform.* **16**, 547–559 (2024).
12. Yadav, N., Dass, R. & Virmani, J. A systematic review of machine learning based thyroid tumor characterisation using ultrasonographic images. *J. Ultrasound* **27**, 209–224 (2024).
13. Yan, L. et al. Development and validation of ultrasound-based radiomics deep learning model to identify bone erosion in rheumatoid arthritis. *Clin. Rheumatol.* **44**, 2635−2645 (2025).
14. Virmani, J. & Agarwal, R. et al. Assessment of despeckle filtering algorithms for segmentation of breast tumours from ultrasound images. *Biocybernetics Biomed. Eng.* **39**, 100–121 (2019).
15. Dass, R. & Yadav, N. Image quality assessment parameters for despeckling filters. *Procedia Comput. Science* **167**, 2382–2392 (2020).
16. Jiao, J. et al. Usfm: A universal ultrasound foundation model generalized to tasks and organs towards label efficient image analysis. *Med.Image Anal.* **96**, 103202 (2024).
17. Christensen, M., Vukadinovic, M., Yuan, N. & Ouyang, D. Vision–language foundation model for echocardiogram interpretation. *Nat. Med.* **30**, 1481–1488 (2024).

18. Voigt, P. & Von dem Bussche, A. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing* **10**, 10–5555 (2017).
19. He, K. et al. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16000–16009 (IEEE, 2022).
20. Anna, M. et al. *Ultrasound Nerve Segmentation.* https://kaggle.com/competitions/ultrasound-nerve-segmentation (2016).
21. Belasso, CJ, Behboodi, B, Benali, H, Boily, M, Rivaz, H, Fortin, M. Lumbar multifidus muscle segmentation from ultrasound. *BMC Musculoskelet Disord*. **21**, 703 (2020).
22. Yang, J., Ding, X., Zheng, Z., Xu, X. & Li, X. Graphecho: Graph-driven unsupervised domain adaptation for echocardiogram video segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11878–11887 (2023).
23. Stanford AIMI. *Stanford Aimi Shared Datasets*. https://stanfordaimi.azurewebsites.net/ (2020).
24. Pedraza, L. et al. An open access thyroid ultrasound image database. In *10th International Symposium on Medical Information Processing and Analysis*, 188–193 (SPIE, 2015).
25. Wunderling, T. et al. Comparison of thyroid segmentation techniques for 3d ultrasound. In *Medical Imaging 2017: Image Processing*, vol. 10133, 346–352 (SPIE, 2017).
26. Gong, H. et al. Multi-task learning for thyroid nodule segmentation with thyroid region prior. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 257–261 (IEEE, 2021).
27. Lu, Y. et al. The jnu-ifm dataset for segmenting pubic symphysis-fetal head. *Data in brief* **41**, 107904 (2022).
28. Tanwar, S. K., Choudhary, P., Priyanka & Agrawal, T. Hcn: Hybrid capsule network for fetal plane classification in ultrasound images. *Int. J. Imag. Syst. Technol.* **34**, e23149 (2024).
29. Krishna, T. B. & Kokil, P. Standard fetal ultrasound plane classification based on stacked ensemble of deep learning models. *Expert Syst. Appl.* **238**, 122153 (2024).
30. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778 (IEEE, 2016).
31. Dosovitskiy, A. et al. An image is worth 16 x 16 words: transformers for image recognition at scale. *arXiv* https://doi.org/10.48550/arXiv.2010.11929 (2020).
32. Basu, S., Gupta, M., Madan, C., Gupta, P. & Arora, C. Focusmae: Gallbladder cancer detection from ultrasound videos with focused masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11715–11725 (2024).
33. Hasan, M. Z., Rony, M. A. H., Chowa, S. S., Bhuiyan, M. R. I. & Moustafa, A. A. Gbchv an advanced deep learning anatomy aware model for accurate classification of gallbladder cancer utilizing ultrasound images. *Sci. Rep.* **15**, 7120 (2025).
34. Liu, Y., Yang, Y., Jiang, Y., Zhao, X. & Xie, Z. Fabrf-net: A frequency-aware boundary and region fusion network for breast ultrasound image segmentation. *Information Fusion* **123**, 103299 (2025).
35. Huang, J. et al. Emganet: Edge-aware multi-scale group-mix attention network for breast cancer ultrasound image segmentation. In *IEEE Journal of Biomedical and Health Informatics*, 5631−5641(IEEE, 2025).
36. Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J. & Maier-Hein, K. H. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **18**, 203–211 (2021).
37. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 1597–1607 (PMLR, 2020).
38. Caron, M. et al. Unsupervised learning of visual features by contrasting cluster assignments. *Adv. Neural Inform. Process. Syst.* **33**, 9912–9924 (2020).
39. Caron, M. et al. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Cnference on Computer Vision*, 9650–9660 (IEEE, 2021).
40. Chen, X., Xie, S. & He, K. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Cnference on Computer Vision*, 9640–9649 (IEEE, 2021).
41. Antropova, N., Huynh, B. Q. & Giger, M. L. A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets. *Med. Phys.* **44**, 5162–5171 (2017).
42. Jiang, Y. et al. Towards a benchmark for colorectal cancer segmentation in endorectal ultrasound videos: Dataset and model development. *arXiv* https://arxiv.org/abs/2408.10067 (2024).
43. Zhou, Y. et al. A foundation model for generalizable disease detection from retinal images. *Nature* **622**, 156–163 (2023).
44. Qiu, J. et al. Development and validation of a multimodal multitask vision foundation model for generalist ophthalmic artificial intelligence. *NEJM AI* **1**, AIoa2300221 (2024).
45. Wang, Z., Liu, C., Zhang, S. & Dou, Q. Foundation model for endoscopy video analysis via large-scale self-supervised pre-train. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 101–111 (Springer, 2023).
46. Hamamci, I. E. et al. A foundation model utilizing chest ct volumes and radiology reports for supervised-level zero-shot detection of abnormalities. *CoRR* (2024).
47. Pai, S. et al. Foundation model for cancer imaging biomarkers. *Nat. Mach. Intell.* 6, 354−367 (2024).
48. Li, T. et al. Federated optimization in heterogeneous networks. *Proc. Mach. Learn. Syst.* **2**, 429–450 (2020).
49. Wang, J., Liu, Q., Liang, H., Joshi, G. & Poor, H. V. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Adv. Neural Inform. Process syst.* **33**, 7611–7623 (2020).
50. Xing, H. et al. Achieving flexible fairness metrics in federated medical imaging. *Nat. Commun.* **16**, 3342 (2025).
51. Al-Dhabyani, W., Gomaa, M., Khaled, H. & Fahmy, A. Dataset of breast ultrasound images. *Data Brief* **28**, 104863 (2020).
52. Gómez-Flores, W., Gregorio-Calas, M. J. & Coelho de Albuquerque Pereira, W. Bus-bra: a breast ultrasound dataset for assessing computer-aided diagnosis systems. *Med. Phys.* **51**, 3110–3123 (2024).
53. Vallez, N., Bueno, G., Deniz, O., Rienda, M. A. & Pastor, C. Bus-uclm: Breast ultrasound lesion segmentation dataset. *Sci. Data* **12**, 242 (2025).
54. Gong, H. et al. Thyroid region prior guided attention for ultrasound segmentation of thyroid nodules. *Comput. Biol. Med.* **155**, 106389 (2023).
55. Park, W. & Chirikjian, G. S. Interconversion between truncated cartesian and polar expansions of images. *IEEE Trans. Image Process. Publication IEEE Signal Process. Soc.* **16**, 1946 (2007).
56. Li, A. et al. You can mask more for extremely low-bitrate image compression. *arXiv* https://doi.org/10.48550/arXiv.2306.15561 (2023).

Acknowledgements

This work was supported by NSFC with Grant No. 62293482 (awarded to S.C), by the Basic Research Project No. HZQB-KCZYZ-2021067 of Hetao Shenzhen-HK S&T Cooperation Zone (awarded to S.C.), by NSFC with Grant No. 62573371 (awarded to Z.L.), by the Shenzhen-Hong Kong Joint Funding No. SGDX20211123112401002 (awarded to Z.L.), by the Shenzhen General Program No. JCYJ20220530143600001 (awarded to Z.L.), by the Shenzhen Outstanding Talents Training Fund 202002 (awarded to S.C.), by the Guangdong Research Project No.2017ZT07X152 and No. 2019CX01X104 (awarded to S.C.), by the Guangdong Provincial Key Laboratory of Future Networks of Intelligence (Grant No. 2022B1212010001) (awarded to J.R., S.C.), by the Guangdong Provincial Key Laboratory of BigData Computing CHUK-Shenzhen

(awarded to Z.L.), by the NSFC 61931024&12326610 (awarded to Z.L.), by the Key Area R&D Program of Guangdong Province with grant No. 2018B030338001 (awarded to Z.L., S.C.), by the Shenzhen Key Laboratory of Big Data and Artificial Intelligence (Grant No. SYSPG20241211173853027) (awarded to Z.L., S.C.), by China Association for Science and Technology Youth Care Program (awarded to Z.L.), by the NSFC 62225113 (awarded to B.D.), by National Key Research and Development Program of China under Grants 2023YFC2705702 (awarded to B.D.) and by Tencent & Huawei Open Fund (awarded to Z.L.). Y.J. worked as a postdoctoral fellow at West China Hospital of Sichuan University after graduation.

Author contributions

Y.J., C.-M.F., J.R., J.W., Z. L. and Z.Z. designed the method.Y.J. conducted the experiments. Y.J., C.-M.F., J.R., and J.W. wrote the main manuscript text. Z.Z., Y.L., and R.S. revised the manuscript text. Y.J. and Y.H. prepared the main figures. All authors reviewed the manuscript.

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/s41746-025-02085-0.

Correspondence and requests for materials should be addressed to Zhen Li.

[1]FNii-Shenzhen, The Chinese University of Hong Kong, Shenzhen, Shenzhen 518172, China. [2]School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, Shenzhen 518172, China. [3]Department of General Surgery & Laboratory of Gastric Cancer, State Key Laboratory of Biotherapy, Collaborative Innovation Center of Biotherapy and Cancer Center, West China Hospital, Sichuan University, Chengdu, China. [4]Gastric Cancer Center, West China Hospital, Sichuan University, Chengdu, China. [5]School of Computer Science, University College Dublin, Dublin, Ireland. [6]College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China. [7]South China Hospital, Health Science Center, Shenzhen University, Shenzhen 518111, China. [8]School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China. [9]Affiliated Hospital of North Sichuan Medical College, Sichuan 637000, China. [10]North Sichuan Medical College, Sichuan 637000, China. [11]Shenzhen Research Institute of Big Data, Shenzhen 518172, China. [12]Bio-Computing Research Center, Harbin Institute of Technology, Shenzhen, Shenzhen 518055, China. [13]School of Computer Science, Wuhan University, Wuhan 430072, China. [14]Computer Science Program, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Kingdom of Saudi Arabia. [15]Center of Excellence for Smart Health (KCSH), King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Kingdom of Saudi Arabia. [16]Beijing University of Posts and Telecommunications, Beijing 100876, China. [17]School of Biomedical Engineering, Suzhou Institute for Advanced Research, University of Science and Technology of China, Suzhou 215123, China. [18]Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China. [19]These authors contributed equally: Yuncheng Jiang, Chun-Mei Feng. ✉e-mail: lizhen@cuhk.edu.cn