nature communications

Article

https://doi.org/10.1038/s41467-025-59478-8

Improving AI models for rare thyroid cancer subtype by text guided diffusion models

Received: 6 July 2024

Accepted: 24 April 2025

Published online: 13 May 2025

Check for updates

Fang Dai[1,2,3,17], Siqiong Yao[2,17,18], Min Wang[4], Yicheng Zhu[5], Xiangjun Qiu[6], Peng Sun[1], Cheng Qiu[7], Jisheng Yin[8], Guangtai Shen[9], Jingjing Sun[10], Maofeng Wang[11], Yun Wang[12], Zheyu Yang[13], Jianfeng Sang[14], Xiaolei Wang[1], Fenyong Sun[3,18], Wei Cai[13,18], Xingcai Zhang[15,18] & Hui Lu[1,2,16,18]

Artificial intelligence applications in oncology imaging often struggle with diagnosing rare tumors. We identify significant gaps in detecting uncommon thyroid cancer types with ultrasound, where scarce data leads to frequent misdiagnosis. Traditional augmentation strategies do not capture the unique disease variations, hindering model training and performance. To overcome this, we propose a text-driven generative method that fuses clinical insights with image generation, producing synthetic samples that realistically reflect rare subtypes. In rigorous evaluations, our approach achieves substantial gains in diagnostic metrics, surpasses existing methods in authenticity and diversity measures, and generalizes effectively to other private and public datasets with various rare cancers. In this work, we demonstrate that text-guided image augmentation substantially enhances model accuracy and robustness for rare tumor detection, offering a promising avenue for more reliable and widespread clinical adoption.

Recent advancements in medical artificial intelligence (AI) have significantly aided clinical decision-making. However, the issue of insufficient sample collection has long been a concern for researchers[1,2]. This is particularly true for the diagnosis of rare diseases or rare subtypes, where low incidence rates result in limited data and sample diversity, posing significant challenges in training AI models[3–5]. More specifically, their low incidence rates lead to scarce data, often resulting in collected data that cannot cover all features of the disease, manifesting in a phenomenon of certain feature deficiencies[6]. Given these limitations, contemporary research primarily focuses on

[1]Department of Bioinformatics and Biostatistics, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China. [2]SJTU-Yale Joint Center of Biostatistics and Data Science, Technical Center for Digital Medicine, National Center for Translational Medicine, Shanghai Jiao Tong University, Shanghai, China. [3]Department of Clinical Laboratory Medicine, Shanghai Tenth People's Hospital, School of Medicine, Tongji University, Shanghai, PR China. [4]Department of Critical Care Medicine, Jiuquan Hospital of Shanghai General Hospital, Shanghai Jiaotong University School of Medicine, Jiuquan, Gansu, PR China. [5]Department of Ultrasound, Pudong New Area People's Hospital Affiliated to Shanghai University of Medicine and Health Sciences, Shanghai, PR China. [6]Department of Automation, Tsinghua University, Beijing, PR China. [7]Medical college, Nantong University, Nantong, Jiangsu, PR China. [8]Shcool of Artificial Intelligence, University of Chinese Academy of sciences, Beijing, PR China. [9]Xin'an League People's Hospital, Xing'an League, Inner Mongolia, PR China. [10]Department of Ultrasound, Shanghai Fourth People's Hospital Affiliated to Tongji University, Shanghai, PR China. [11]Department of Biomedical Sciences Laboratory, Affiliated Dongyang Hospital of Wenzhou Medical University, Dongyang, Zhejiang, PR China. [12]Department of Hepatobiliary pancreatic center, Xuzhou City Central Hospital, Xuzhou, Jiangsu, China. [13]Department of General Surgery, Ruijin Hospital, Shanghai Jiaotong University School of medicine, Shanghai, PR China. [14]Nanjing Drum Tower Hospital, The Affiliated Hospital of Nanjing University Medical School, Nanjing, Jiangsu, PR China. [15]World Tea Organization, Cambridge, MA, USA. [16]Institute of Bioinformatics, Shanghai Academy of Experimental Medicine, Shanghai, China. [17]These authors contributed equally: Fang Dai, Siqiong Yao. [18]These authors jointly supervised this work: Siqiong Yao, Fenyong Sun, Wei Cai, Xingcai Zhang, Hui Lu.
e-mail: yaosiqiong@sjtu.edu.cn; sunfenyong@263.net; caiwei@shsmu.edu.cn; drzhangxingcai@outlook.com; huilu@sjtu.edu.cn

achieving high predictive performance within the general or common disease cohorts, making it challenging to estimate progress on rare subpopulations. This leads to lower detection rates, higher misdiagnosis and underdiagnosis rates, and potentially increased mortality rates among rare subpopulation groups[7,8] and can lead to potential medical AI unfairness[9,10].

Although individual rare diseases or subtypes have low incidence rates, the overall population affected by such conditions is substantial. This disparity not only raises ethical concerns but also constitutes a significant barrier to the widespread adoption and practical application of medical AI models[11,12]. In the case of thyroid cancer, for instance, current clinical guidelines generally follow standard procedures for papillary thyroid carcinoma (PTC)[13], such as fine-needle aspiration biopsy and surgical treatment. However, more accurate pre-diagnosis of subtypes, particularly rare ones, could enable the adoption of more targeted approaches, such as conservative treatment strategies or further diagnostic investigations in different directions. For example, while PTC typically warrants total thyroidectomy with central compartment neck dissection, the surgical approach for follicular thyroid carcinoma (FTC) varies depending on tumor size and vascular invasion, sometimes requiring only a hemithyroidectomy[14]. Conversely, medullary thyroid carcinoma (MTC), which is associated with genetic factors, often involves regional lymph node metastasis and distant metastases to the lungs and bones, necessitating the observation of metastatic sites before proceeding with treatment[15]. Some subtypes are prone to metastasis and may not be suitable for surgical intervention[16,17]. Hence, the precise diagnosis of rare subtypes is essential to refine the existing ultrasound-based risk stratification systems.

Building on previous work[10], in our AI model for the diagnosis of thyroid cancer subtypes through ultrasound imaging, we identified significant predictive imbalances. The scarcity of certain subtypes, characterized by low incidence rates and features that are difficult to distinguish from those of more common subtypes, means that imbalances in sample categories and the uneven distribution of distinguishing features are the primary reasons for these predictive discrepancies. This poses challenges to the training of deep learning AI algorithms. Data augmentation is an approach to address sample size limitations in AI[18–21]. Traditional augmentation methods such as basic image manipulations (rotations, stretching, translation, scaling, arbitrary cropping, etc.), image mixing (overlaying two images), and random erasing (obscuring certain parts of an image)[19], are more suitable for handling natural images. When applied to medical images, these techniques can alter the morphological information of lesions, making them unsuitable for medical AI training sets due to the changes in clinical significance and potentially leading to inaccuracies in diagnosis or treatment recommendations. An increasing number of researches are turning their focus to generative models, which automatically learn and replicate the patterns in input images, enabling them to produce examples that resemble those from the original dataset[20,22]. Such generative methods, based on existing datasets for feature augmentation, often face issues of insufficient sample diversity when applied to rare diseases with limited sample sizes[23]. These problems directly impact the generalization performance of model predictions.

Recently, generative models have demonstrated a performance advantage in feature fusion between image and text[24]. This method uses text-guided approaches to avoid mode collapse, producing high-quality generated images in natural settings, resulting in significant performance improvements for downstream tasks. For instance, Mohamed et al.[25] investigated the impact of using diffusion-based text-to-image data augmentation methods on the classification of skin diseases, especially focusing on the comparison between original medical datasets and entirely synthetic images. Through this approach, the study aimed to enhance the accuracy and generalization capability of skin disease classification models. Chambon et al.[26] used the stable diffusion method to guide the generation of CT images based on the names of various diseases in chest radiographs, which improved the prediction accuracy for different types of diseases. In fact, generative data augmentation methods guided by disease knowledge are one of the more effective means to enhance the diversity and authenticity of generated images[27]. However, the majority of existing studies primarily employ category-level classifications, such as disease types or subtype names, to direct the generation of new images. This approach is suitable for augmenting datasets with images of prevalent diseases or common subtypes, yet may be faced with challenges when utilized for generating diverse samples of insufficiently represented diseases or variations (rare subtypes)[28]. To address this limitation, it is crucial to incorporate detailed textual descriptions that highlight the unique features of these rare subtypes at an attribute level. This addition will facilitate the generation of truly diverse and representative images of rare conditions.

In this work, we introduce the Tiger Model, a text-guided medical image generation deep learning model to specifically address diversity and generalization in the prediction of rare medical diseases. It enables controlled generation of features at a fine granularity by introducing extensive clinical knowledge of imaging. The model comes with text and image control modules, aiming to enhance the diversity and authenticity of rare subtype feature generation. The text-guided data generation framework of the Tiger Model shows promise in diagnosing rare medical diseases, offering images that are more diverse and realistic with enhanced user comprehensibility that could accelerate the translation of medical AI advancements.

Results

Overview of the Tiger Model

To achieve models that realize detailed feature control in order to generate new images with realistic diversity, we designed the Tiger Model as outlined below. The overall usage process of the Tiger Model is illustrated in Fig. 1. Firstly, we constructed and summarized a disease knowledge base detailing the imaging feature differences between common and rare subtypes, based on literature search and clinical reports including Composition, Echogenicity, Echotexture, Calcification, Aspect, Shape, Margin, Halo, Diameter, Nodule Location, Extension and Sonogram Direction. We categorized similar descriptions and summarized the commonalities and distinctive features between common and rare subtypes, forming a disease knowledge base. We refer to Supplementary Table 1 for details. Next, utilizing the clinical knowledge guidance mechanism of the Tiger Model, we trained the model to generate features of benign and common tumors, and then trained it to generate unique features of rare tumors through common feature transfer. Finally, by combining the permutations of benign nodules and common tumor features with the unique features of rare tumors, we achieved the realistic and diverse synthesis of rare subtype features.

The architecture of the Tiger Model is shown in Fig. 2, consisting of two main phases: Coarse-Training and Fine-Training. Coarse-Training utilizes ultrasound images and corresponding textual reports as inputs (Fig. 2). Clinically, features of rare subtypes are usually derived from combinations and permutations of benign and common tumor features, with a few unique features[29]. Therefore, transferring features from benign and common tumors can foster the reconstruction of features for rare subtype tumors. During Fine-Training, we conducted rule-based generative training according to the characteristic combinations of disease subtypes. The ultrasonic diagnosis of thyroid cancer subtypes often relies on comprehensive judgment based on the co-occurrence and synergy features of ultrasonography. Considering the issue of intra-class imbalance between the tumor area (foreground, abbr. FG) and background (abbr. BG) information in ultrasound images, we designed detailed feature control methods for extracting features from the foreground (FG) and the background (BG) separately, using FG-Encoder and BG-Encoder (Fig. 2). The computation process of the attention mechanism in the

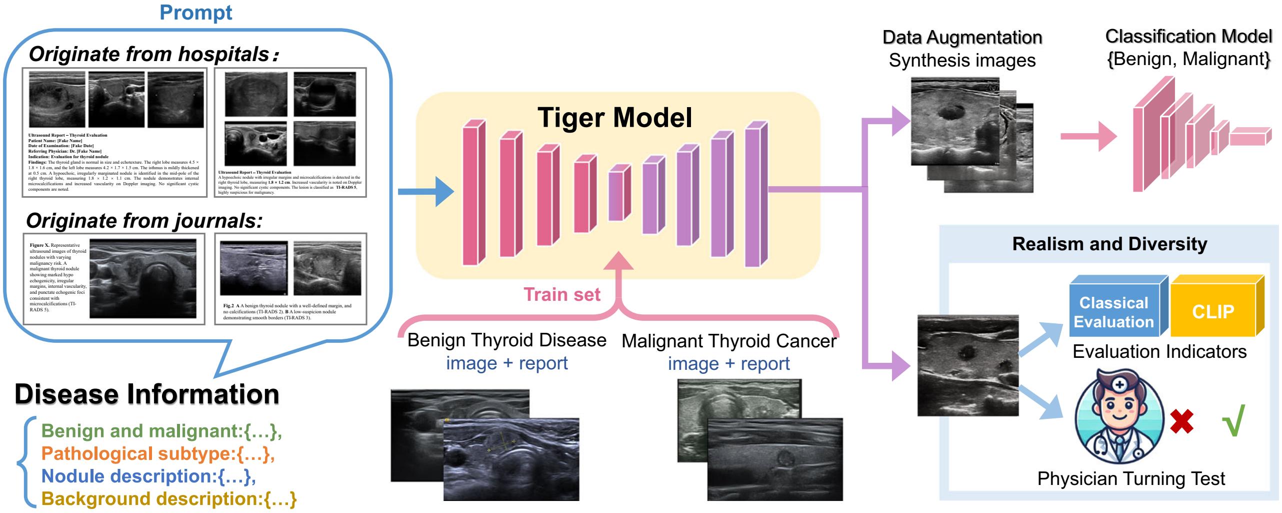


Fig. 1 | Training and Evaluation Process of Tiger Model. The Tiger Model is trained and constructed based on the differences in disease subtype features, utilizing disease knowledge(prompt). Through the model's extraction of commonality and differences in the feature domain, it achieves the generation of true diversity of features for rare subtypes. The training data of the model are ultrasonic images and ultrasonic reports. Tiger generates augmented datasets and merges them with real data for training classification models. The realism and diversity of the images are then tested by professionals and evaluated using computer vision indices.

j-th layer of each encoder can be expressed by the following formula:

$$\mathbb{C}^j_m = \text{Attention}^j_m(Q, K, V) = \text{softmax}\left(W^j_m\left(Q^j, K^j\right)\right) \cdot V^j \qquad (1)$$

Where $m \in \{\text{FG}, \text{BG}\}$. The output of FG-Encoder and BG-Encoder (denoted as F_{FG} and F_{BG}) was combined with a weighted module for foreground-background fusion represented in the formula (2) and (3). Detailed implementations are elaborated in the Method section 2: Tiger Model architecture and Method section 3: Training details. The fusion design enhances the realism and rigor of the generated images, including the following aspects: color and resolution consistency, texture rationality, scanning plane authenticity, and the correctness of the synthesized tumor subtype.

$$W_{\text{FG+BG}} = \text{sigmoid}(\text{Conv}(F_{\text{FG}} \oplus F_{\text{BG}})) \qquad (2)$$

$$\mathcal{F} = e_{\text{FG+BG}} F_{\text{FG}} + (1 - W_{\text{FG+BG}}) F_{\text{BG}} \qquad (3)$$

Data collection and experimental design

The data collection process strictly followed the inclusion and exclusion criteria detailed in the Method section. We collected thyroid ultrasound reports and pathological reports from 68,386 patients across 10 hospitals. After exclusion, we enrolled a final cohort of 40,571 patients. Among these, there were 21,920 benign cases, 13,552 cases of Papillary Thyroid Carcinoma (PTC, with an incidence rate of 90–95%)[13], 4102 cases of the rare subtype Follicular Thyroid Carcinoma (FTC, with an incidence rate of only 3–5%)[30], 997 cases of the rare subtype Medullary Thyroid Carcinoma (MTC, with an incidence rate of only 1–3%)[31], and 274 cases of the rare subtypes Anaplastic Thyroid Carcinoma (ATC, with an incidence rate of only 0.2%)[32]. For gender, the male-to-female ratio was 0.46; for nodule sizes, they ranged from 0.1 cm to 1.9 cm, with the < 1 cm group accounting for 44%. We collected a total of 186,812 thyroid cancer ultrasound images corresponding to the 40,845 reports. Specifically, there were 93,220 images for benign cases, 48,585 for PTC, 31,725 for FTC, 12,910 for MTC, and 372 for ATC (Table 1).

To validate the enhancement in model generalization through data augmentation, we observed the impact on the benign-malignant diagnosis task across four subtypes of thyroid cancer (one common subtype, PTC, and three rare subtypes FTC, MTC, and ATC, as outlined in Table 1a), focusing primarily on the malignancy degree of thyroid nodules. To improve nodule identification, prior to constructing the Tiger Model, we trained a thyroid and nodule segmentation network using independent 4000 images annotated by medical professionals (2000 benign and 2000 malignant). The dataset splits for Classification, Generation, Segmentation, and CLIP tasks, including thyroid, pediatric chest, and breast datasets, are detailed in Supplementary Table 2.

Quantitative analysis of image generation quality

In Table 2, we utilize a comprehensive set of metrics to evaluate the performance of our generative network, including the Structural Similarity Index (SSIM)[33], CLIP-MMD (CMMD)[34], Gradient Similarity (GS)[35], and Density and Coverage (D&C)[36]. The SSIM scores for the Tiger-F model are significantly higher than those of the stable diffusion model (SD-S) (by 13.16, 39.34, and 31.75%), indicating superior preservation of structural information in the Tiger Model. In addition, Tiger-F outperforms the stable diffusion model in both CMMD and GS metrics, showing reductions of 37.14, 38.46, and 26.79% in CMMD and 18.65, 82.02, and 73.43% in GS. These results suggest that the Tiger Model generates more realistic images. Furthermore, the D&C score of the Tiger Model surpasses that of the stable diffusion model, reflecting a more compact and comprehensive clustering solution. Compared to the stable diffusion model, images generated by the Tiger Model exhibit improved authenticity, with a 16.69% reduction in FID and a 19.64% reduction in p-Hash, and greater diversity, as indicated by a 39.17% increase in IS. Detailed experimental results are provided in Supplementary Table 3.

Doctor assessment of image generation quality

We conducted three Turing test experiments with medical professionals (Figs. 3 and 4) to qualitatively evaluate the diversity and quality control of the images generated by the Tiger Model. We recruited 50 ultrasound physicians, evenly divided between senior clinicians (with over 5 years of clinical experience) and junior clinicians (with 3–5 years of clinical experience).

In the first test, doctors were shown images and asked to assess whether they were real or generated. The images included real images,

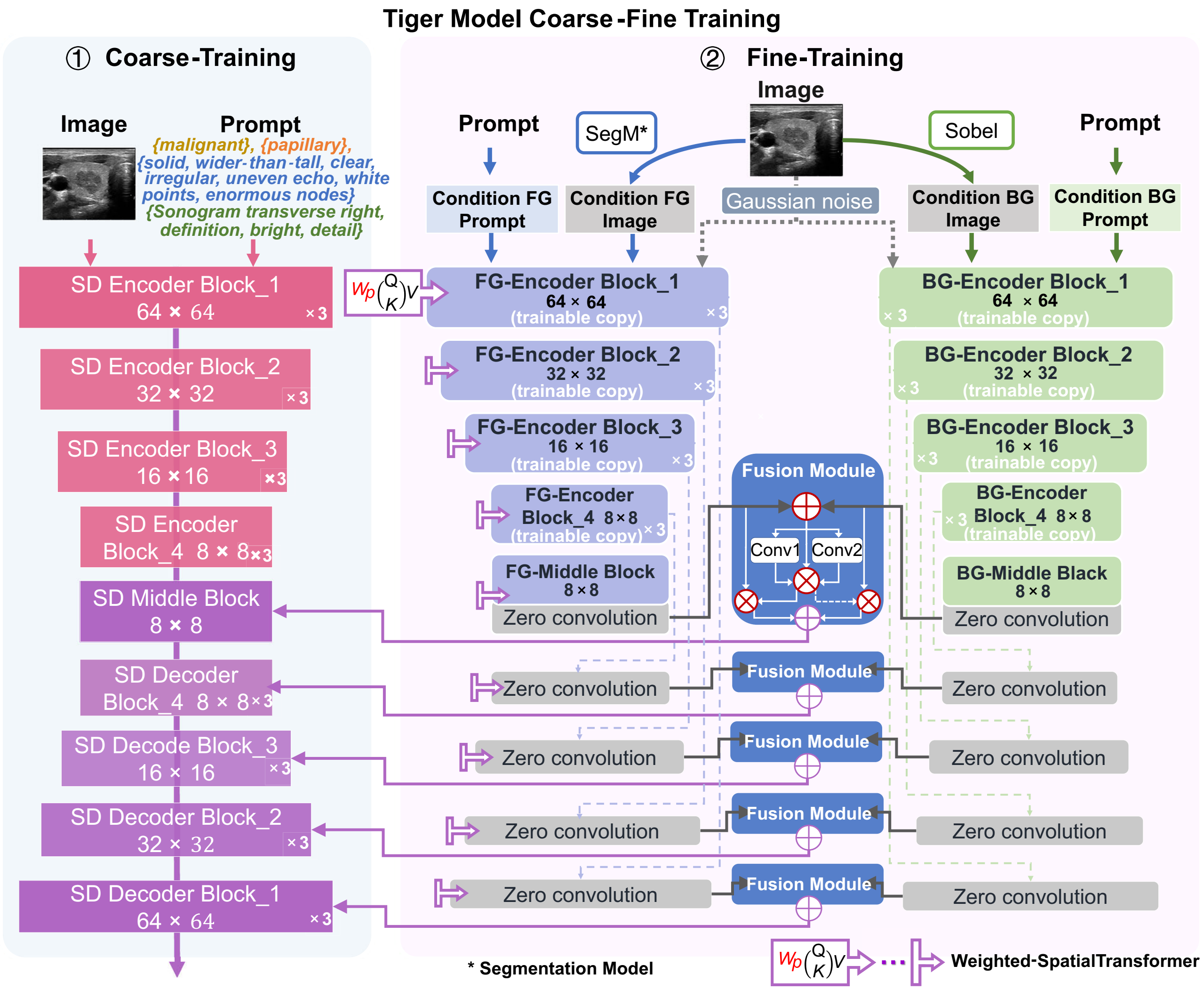

Tiger Model Coarse-Fine Training
① Coarse-Training
Image
Prompt
{malignant}, {papillary},
{solid, wider-than-tall, clear, irregular, uneven echo, white points, enormous nodes}
{Sonogram transverse right, definition, bright, detail}
SD Encoder Block_1 64 × 64 ×3
SD Encoder Block_2 32 × 32 ×3
SD Encoder Block_3 16 × 16 ×3
SD Encoder Block_4 8 × 8 ×3
SD Middle Block 8 × 8
SD Decoder Block_4 8 × 8 ×3
SD Decode Block_3 16 × 16 ×3
SD Decoder Block_2 32 × 32 ×3
SD Decoder Block_1 64 × 64 ×3
② Fine-Training
Image
Prompt
SegM*
Sobel
Prompt
Condition FG Prompt
Condition FG Image
Gaussian noise
Condition BG Image
Condition BG Prompt
$W_p\binom{Q}{K}V$
FG-Encoder Block_1 64 × 64 (trainable copy) ×3
FG-Encoder Block_2 32 × 32 (trainable copy) ×3
FG-Encoder Block_3 16 × 16 (trainable copy) ×3
FG-Encoder Block_4 8×8 (trainable copy) ×3
FG-Middle Block 8×8
Zero convolution
BG-Encoder Block_1 64 × 64 (trainable copy) ×3
BG-Encoder Block_2 32 × 32 (trainable copy) ×3
BG-Encoder Block_3 16 × 16 (trainable copy) ×3
BG-Encoder Block_4 8×8 (trainable copy) ×3
BG-Middle Black 8×8
Zero convolution
Fusion Module
Conv1
Conv2
Zero convolution
Fusion Module
Zero convolution
Zero convolution
Fusion Module
Zero convolution
Zero convolution
Fusion Module
Zero convolution
Zero convolution
Fusion Module
Zero convolution
* Segmentation Model
$W_p\binom{Q}{K}V$ ··· Weighted-SpatialTransformer

Fig. 2 | Tiger Model Architecture Design and Applications. The Tiger Model is trained and constructed based on the differences in disease subtype features, utilizing disease knowledge(prompt). The model identifies commonality and differences within the feature domain, allowing it to generate diverse feature combinations reflecting true rare subtypes. The training data of the model encompass ultrasonic images and ultrasonic report text. Tiger generates augmented datasets and merges them with real data for training classification models. The realism and diversity of the images are then tested by professionals and evaluated using computer vision indices.

Tiger-F generated images, and images generated by other methods (PG-GAN, Diffusion Transformers, Imagen, and Stable Diffusion), with 500 images per category. As shown in Fig. 3a, doctors correctly identified real images as authentic in 98.0% of cases, Tiger-F generated images in 92.2%, and images generated by other methods in only 33.7% of cases (Fig. 3b). This demonstrates the Tiger Model's strong potential in enhancing the realism of radiological data augmentation.

In the second test, doctors were asked to select the correct image from four candidate images based on a provided description (Fig. 3c). The correct images were randomly chosen from a dataset of both real and Tiger-F generated images, covering benign, PTC, FTC, and MTC types. For each correct image, three similar images were selected using pHash values but paired with different textual descriptions. This test aimed to evaluate whether physicians could accurately identify rare subtypes, which are often prone to misdiagnosis. Each doctor participated in 20 rounds, earning one point for each correct choice. The results (Fig. 3d) revealed that, among the 25 junior clinicians, 12 answered all questions correctly, 8 achieved an accuracy rate above 90%, while 5 scored below 90%. Among the 25 senior clinicians, 20 answered all rounds correctly, while the remaining 5 achieved over 90% accuracy. Our method was compared with Stable Diffusion and Imagen, demonstrating a superior average accuracy: a 31.09% improvement over Stable Diffusion and 36.95% over Imagen for junior clinicians, and a 37.01% improvement over Stable Diffusion and 44.58% over Imagen for senior clinicians.

In the third test, doctors were asked to identify the presence of 10 specific radiological features in Tiger-generated images and to annotate the images accordingly. Images containing all 10 features received the highest scores. Each doctor assessed 30 images randomly—a mix of real and generated images—across PTC, FTC, and MTC subtypes, with 500 images per subtype (Fig. 4a). Figure 4b shows the frequency of correct feature identification within the 500 images. The results indicated high accuracy in feature estimation for the generated samples: 84.1% for PTC, 81.5% for FTC, and 82.8% for MTC, which closely aligned with the accuracy for real samples (PTC 87.9%, FTC 85.3%, MTC 84.2%). For comparison, we fine-tuned the CLIP model with medical text and image features for the same task[37]. The CLIP model achieved estimations of 81.5%, 81.2%, and 78.8% for PTC, FTC, and MTC in generated images, respectively, which did not significantly differ from the results for real images (85.3%, 83.8%, 83.9%). Examples comparing CLIP with physician assessments are provided in Supplementary Table 4. In addition, we tested the performance of Stable

Table 1 | Description of Dataset

Thyroid Ultrasound	ALL	Benign	Malignancy			
			PTC	FTC	MTC	ATC*
Patient	40,845	21920	13552	4102	997	274
Image	186,812	93220	48585	31725	12910	372
Report	40,845	21920	13552	4102	997	274
Breast Ultrasound	**ALL**	**Benign**	**Malignancy**			
			Sum	**IDC***	**ILC***	**PBC***
Patient	4581	3058	1523	1042	58	423
Image	6275	3383	2892	2274	89	529
Report	4581	3058	1523	1042	58	423
Chest X-ray	**ALL**	**Benign**	**Malignancy**			
			Bronchitis	**Pneumonia**	**Broncho-pneumonia**	**Bronchiolitis**
VinDr-PCXR$^{\Delta}$	2032	\	842	148	545	497

The dataset includes 186,812 ultrasound images from 40,845 thyroid patients, along with their corresponding ultrasound and pathological reports. In addition, breast ultrasound images and chest X-ray (VinDr-PCXR) images are incorporated for external evaluation.

In the Breast Ultrasound and Chest X-ray dataset, each patient has only one image and one report about the image.

PTC Papillary Thyroid Carcinoma, *FTC* Follicular Thyroid Carcinoma, *MTC* Medullary Thyroid Carcinoma, *ATC* Anaplastic Thyroid Carcinoma, *IDC* Invasive Ductal Carcinoma, *ILC* Infiltrating Lobular Carcinoma, *PCB* Papillary Carcinoma of the Breast.

All the classification experiments adopted random 5-fold cross-validation.

*External validation set for ATC(Thyroid), IDC(Breast), ILC(Breast), PCB(Breast).

$^{\Delta}$Public dataset for chest.

Table 2 | Comparative analysis of generated Image quality

Method	PTC Synthesis Image (30k)				FTC Synthesis Image (30k)				MTC Synthesis Image (30k)			
	SSIM	CMMD	GS	D&C	SSIM	CMMD	GS	D&C	SSIM	CMMD	GS	D&C
PG-GAN	0.72	0.54	3.25×10^{-3}	0.08	0.52	0.78	5.82×10^{-3}	0.04	0.55	0.84	4.82×10^{-3}	0.03
DiT	0.77	0.35	2.22×10^{-3}	0.07	0.48	0.66	4.94×10^{-3}	0.06	0.63	0.72	4.15×10^{-3}	0.05
Imagen	0.79	0.43	1.93×10^{-3}	0.10	0.63	0.56	3.94×10^{-3}	0.08	0.72	0.63	3.03×10^{-3}	0.03
SD-S	0.76	0.35	8.41×10^{-4}	0.8	0.61	0.52	5.46×10^{-3}	0.05	0.63	0.56	4.14×10^{-3}	0.03
SD-C	0.73	0.39	7.28×10^{-4}	0.06	0.68	0.52	3.42×10^{-3}	0.06	0.69	0.74	4.20×10^{-3}	0.04
Tiger-N	0.82	0.27	6.39×10^{-4}	0.14	0.75	0.35	2.41×10^{-3}	0.13	0.69	0.48	2.16×10^{-3}	0.07
Tiger-F	0.86	0.22	6.84×10^{-4}	0.22	0.85	0.32	9.82×10^{-4}	0.17	0.83	0.41	1.10×10^{-3}	0.08

Tiger-N refers to Tiger Name-Guided, Tiger-F refers to Tiger Feature-Guided; SD-S refers to Stable Diffusion + Segmentation Model; SD-C refers to Stable Diffusion + ControlNet.

SSIM scores range from 0 to 1, and 0.8 means the generated images retain most of the structural and perceptual qualities of the reference images. If CMMD value is close to 0, real images and fake images are very similar; if the two sets differ significantly, the CMMD value will be relatively high. A lower GS indicates that the generated data is more topologically similar to the real data. D&C refers to Density/Coverage, the high Density and high Coverage indicate that the generative model can produce high-quality images that also cover the diversity of real images.

The real dataset involved in the calculation of evaluation indicators (SSIM, CMMD, GS, D&C) are from test sets of PTC, FTC and MTC in Table 1.

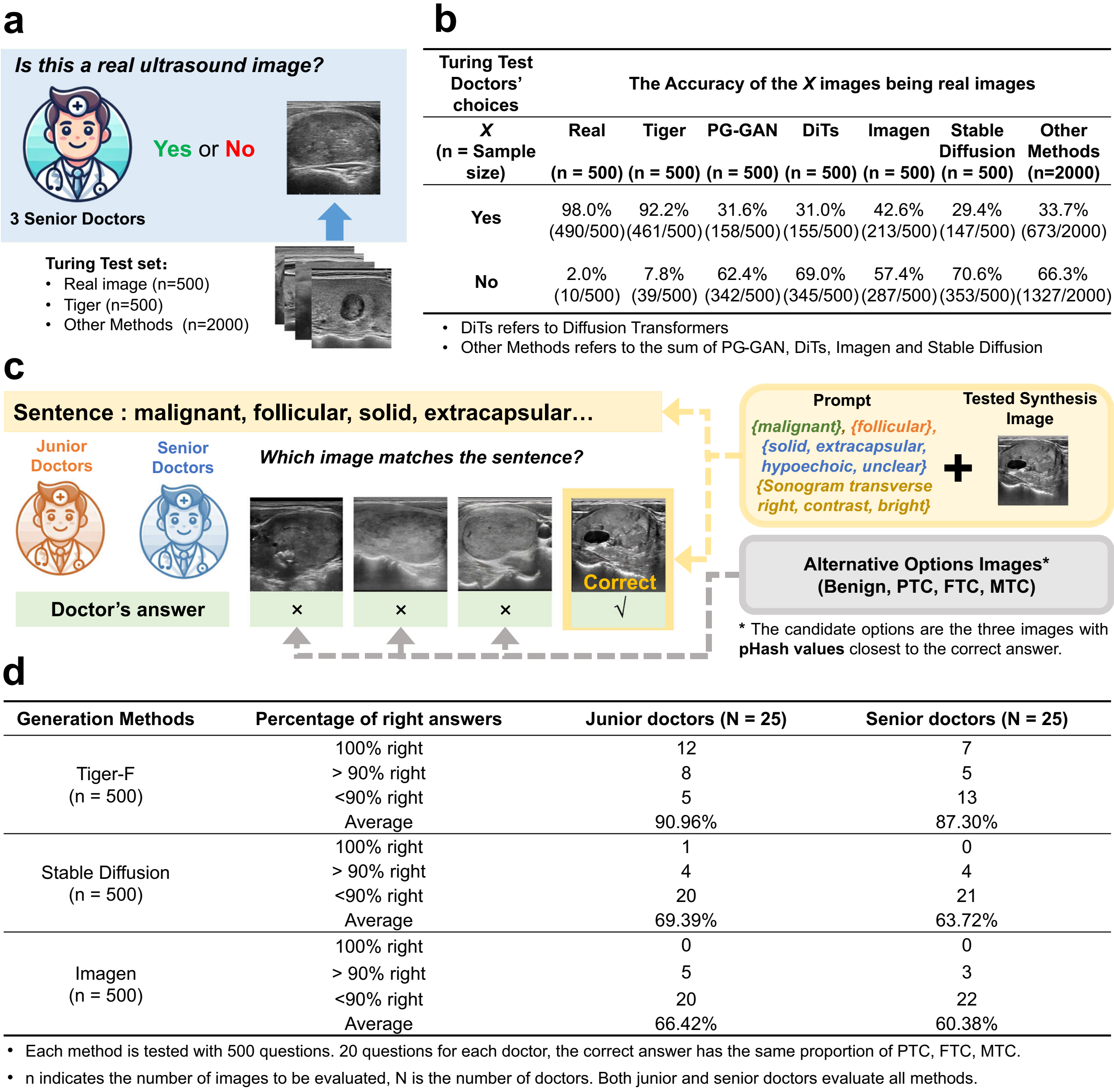
Diffusion and Imagen. The physician evaluation indicated that our method outperformed Stable Diffusion by over 13% (PTC 13.3%, FTC 17.7%, MTC 15.4%) and Imagen by over 10% (PTC 10.9%, FTC 10.3%, MTC 11.6%). The CLIP model test also showed that our method outperformed Stable Diffusion by over 6% (PTC 6.7%, FTC 17.2%, MTC 13.3%) and Imagen by over 9% (PTC 9.0%, FTC 13.6%, MTC 9.6%) (detailed results are presented in Fig. 4c).

Through multi-level validation using both human and machine assessments, it is evident that the Tiger Model generates highly realistic images, capable of standing up to scrutiny. A model that is both interpretable and trustworthy forms the cornerstone for widespread clinical diagnostic applications. Moreover, our model's training generation is based on learning the features of disease subtypes, allowing it to interpret and predict the characteristics of each subtype. As such, the model can assist clinical experts in gaining deeper insights into the disease.

Diagnoses for rare thyroid cancer subtype - downstream task

In the binary classification of benign and malignant cases in rare subtype thyroid cancer, we present the comparative efficacy of the Tiger Model in Table 3, benchmarking it against baselines that include four traditional visual augmentation methods (Basic image manipulation[38], Mixing image[39], Random erasing[40], Feature space augmentation[41]), and four deep learning augmentation methods (PG-GAN[42], Diffusion Transformers[43], Imagen[44], Stable Diffusion[24]). Details of the training processes are provided in Supplementary Method 1. To assess the effectiveness of feature descriptions in our prompt design, we evaluated the performance of the Tiger Model guided by two types of prompts: Tiger Name-Guided (abbr. Tiger-N; prompts including only benign-malignant labels and pathological subtypes) and Tiger Feature-Guided (abbr. Tiger-F; prompts including benign-malignant labels, pathological subtypes, and feature descriptions of nodules and background; default settings for the Tiger Model). For each augmentation method, we applied it to the original training set to create an augmented set and then trained a binary classification model based on ResNet50[45] on the augmented data. We compared the predictive performance of classification models derived from each augmentation method.

It is worth noting that the Tiger Model is also adaptable for multi-class classification tasks, as demonstrated in external evaluations outlined in subsequent sections. Each method augmented the dataset to 30,000 images. The results in Table 3 demonstrate that, compared to the baseline, Tiger-F showed the most significant improvement in


a
Is this a real ultrasound image?
Yes or No
3 Senior Doctors
Turing Test set:
Real image (n=500)
Tiger (n=500)
Other Methods (n=2000)
b
c
Sentence : malignant, follicular, solid, extracapsular…
Junior Doctors
Senior Doctors
Which image matches the sentence?
Doctor's answer
×
×
×
Correct
√
Prompt
{malignant}, {follicular}, {solid, extracapsular, hypoechoic, unclear} {Sonogram transverse right, contrast, bright}
+
Tested Synthesis Image
Alternative Options Images* (Benign, PTC, FTC, MTC)
* The candidate options are the three images with pHash values closest to the correct answer.
d

Turing Test Doctors' choices	The Accuracy of the *X* images being real images						
X (n = Sample size)	Real (n = 500)	Tiger (n = 500)	PG-GAN (n = 500)	DiTs (n = 500)	Imagen (n = 500)	Stable Diffusion (n = 500)	Other Methods (n=2000)
Yes	98.0% (490/500)	92.2% (461/500)	31.6% (158/500)	31.0% (155/500)	42.6% (213/500)	29.4% (147/500)	33.7% (673/2000)
No	2.0% (10/500)	7.8% (39/500)	62.4% (342/500)	69.0% (345/500)	57.4% (287/500)	70.6% (353/500)	66.3% (1327/2000)

- DiTs refers to Diffusion Transformers
- Other Methods refers to the sum of PG-GAN, DiTs, Imagen and Stable Diffusion

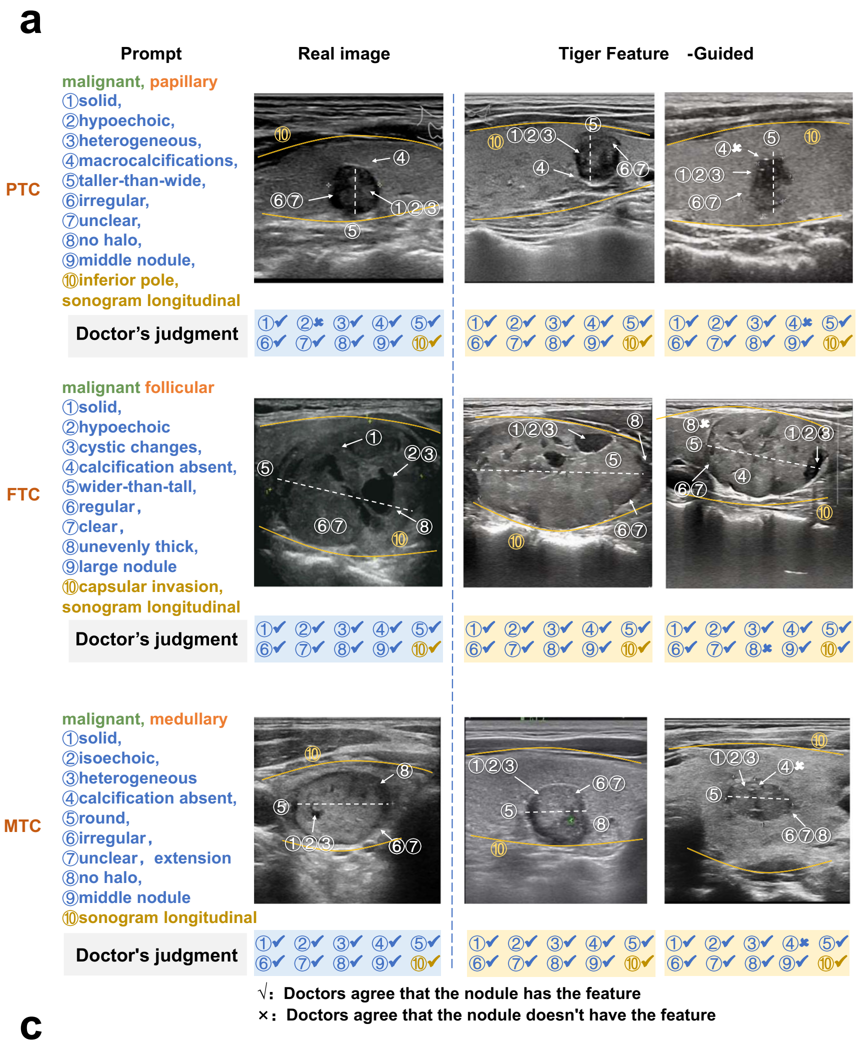
Generation Methods	Percentage of right answers	Junior doctors (N = 25)	Senior doctors (N = 25)
Tiger-F (n = 500)	100% right	12	7
	> 90% right	8	5
	<90% right	5	13
	Average	90.96%	87.30%
Stable Diffusion (n = 500)	100% right	1	0
	> 90% right	4	4
	<90% right	20	21
	Average	69.39%	63.72%
Imagen (n = 500)	100% right	0	0
	> 90% right	5	3
	<90% right	20	22
	Average	66.42%	60.38%

- Each method is tested with 500 questions. 20 questions for each doctor, the correct answer has the same proportion of PTC, FTC, MTC.
- n indicates the number of images to be evaluated, N is the number of doctors. Both junior and senior doctors evaluate all methods.

Fig. 3 | Design and results of the first two Turing tests. a Turing test 1: Three doctors judged each picture to be real or fake. The images are randomly selected from the Turing Test Set. **b** In Turing test 1, compared to other generative methods, the Tiger Model (Tiger-F) received the closest scores to real images. **c** Turing test 2: The expert chooses the one that corresponds to the text from four pictures. The four choices include one correct choice that corresponds to the sentence, and three different images randomly chosen from the Turing Test Set. **d** In Turing test 2, compared to other generative methods, images generated by the Tiger Model received the highest scores from medical experts in selecting the correct ones. Source data are provided as a Source Data file.

predictive performance, with AUC increases of 14.64%, from 0.7364(95% CI:0.69–0.75) to 0.8442(95% CI:0.75–0.85) and 9.45%, from 0.7523(95% CI:0.67–0.79) to 0.8234(95% CI:0.78–0.85) for the rare subtypes FTC and MTC, respectively. In addition, we observed an increase in sensitivity for FTC classification, from 0.5833(95% CI:0.54–0.61) to 0.7983(95% CI:0.77–0.82), and in specificity, from 0.6250(95% CI:0.60-0.65) to 0.8404(95% CI:0.82–0.86). For MTC, sensitivity improved from 0.6529(95% CI:0.63–0.68) to 0.7891(95% CI:0.76–0.80), and specificity rose from 0.6976(95% CI:0.68–0.82) to 0.8325(95% CI:0.80–0.85). While Tiger-N also exceeded baseline performance (FTC and MTC) with AUC improvements of 5.01% and 6.71%, respectively, it fell short of Tiger-F by 3.81% and 2.58%. Our approach demonstrates robustness even when limited to disease names alone, with substantial performance gains when labels incorporate comprehensive feature descriptions. Additional comparative analysis demonstrated statistical significance of the findings (Supplementary Table 5). Further details of the Tiger model compared to baseline models are shown in Supplementary Fig. 1.

We then quantitatively assessed the predictive performance of the Tiger Model across varying sample size proportions in classification tasks (Fig. 5a). The results show that the Tiger Model consistently achieved higher predictive efficiency across all three subtype tasks compared to other methods. As the volume of generated data increased, the Tiger Model exhibited a steeper rise in AUROC than



b

Attributes	Synthesis PTC Accuracy, n = 500	
	Doctors	CLIP
Composition (①)	80.4% (402/500)	82.4% (412/500)
Echogenicity (②)	84.8% (424/500)	80.8% (404/500)
Echotexture (③)	85.8% (429/500)	74.2% (371/500)
Calcification (④)	81.0% (405/500)	77.4% (387/500)
Aspect (⑤)	79.4% (397/500)	79.8% (399/500)
Shape (⑥)	77.2% (386/500)	75.0% (375/500)
Margin (⑦)	83.2% (416/500)	84.0% (420/500)
Halo (⑧)	71.4% (357/500)	72.2% (361/500)
Diameter (⑨)	99.2% (496/500)	96.8% (484/500)
Direction (⑩)	98.2% (491/500)	89.4% (467/500)
Average	84.1%	81.5%
Real image (100)	87.9%	85.3%

Attributes	Synthesis FTC Accuracy, n = 500	
	Doctors	CLIP
Composition (①)	76.2% (381/500)	75.8% (379/500)
Echogenicity (②)	84.2% (421/500)	82.2% (411/500)
Echotexture (③)	64.4% (322/500)	74.2% (371/500)
Calcification (④)	83.0% (415/500)	83.4% (417/500)
Aspect (⑤)	76.2% (381/500)	79.2% (396/500)
Shape (⑥)	81.2% (406/500)	75.6% (378/500)
Margin (⑦)	83.2% (416/500)	83.4% (417/500)
Halo (⑧)	67.4% (337/500)	85.4% (427/500)
Diameter (⑨)	99.8% (499/500)	84.2% (421/500)
Direction (⑩)	99.2% (496/500)	89.6% (448/500)
Average	81.5%	81.2%
Real image (100)	85.3%	83.8%

Attributes	Synthesis MTC Accuracy, n = 500	
	Doctors	CLIP
Composition (①)	73.2% (366/500)	76.4% (382/500)
Echogenicity (②)	76.2% (381/500)	74.6% (373/500)
Echotexture (③)	84.2% (421/500)	91.2% (456/500)
Calcification (④)	74.2% (371/500)	79.0% (395/500)
Aspect (⑤)	84.2% (421/500)	83.8% (419/500)
Shape (⑥)	84.0% (420/500)	72.4% (362/500)
Margin (⑦)	80.2% (401/500)	71.4% (357/500)
Halo (⑧)	74.4% (372/500)	63.2% (316/500)
Diameter (⑨)	99.2% (496/500)	84.8% (424/500)
Direction (⑩)	98.4% (492/500)	91.2% (456/500)
Average	82.8%	78.8%
Real image (100)	84.2%	83.9%

c

Generation Methods*	Evaluation	Synthesis PTC Accuracy (n = 500)	Synthesis FTC Accuracy (n = 500)	Synthesis MTC Accuracy (n=500)
Imagen	Doctors	70.8% (354/500)	60.4% (302/500)	68.2% (341/500)
Stable Diffusion		73.2% (366/500)	62.6% (313/500)	69.2% (346/500)
Tiger Name -Guided		81.2% (406/500)	68.4% (342/500)	76.4% (382/500)
Imagen	CLIP	75.0% (375/500)	63.6% (318/500)	65.6% (328/500)
Stable Diffusion		72.6% (363/500)	68.2% (341/500)	67.2% (336/500)
Tiger Name -Guided		77.4% (387/500)	73.6% (368/500)	74.6% (373/500)

* Stable Diffusion and Imagen are the only two models in the comparison that use text generation.

Fig. 4 | Design and results of the Turing test 3. a Presents the assessment of 50 doctors based on 10 prompts regarding the correspondence of features between real images and Tiger-F generated images. Annotations illustrate doctors' associations of feature descriptions with image features in ultrasound images. **b** Compiles the results of correct feature judgments by doctors and the CLIP model. Both the doctors and the CLIP model exhibit similar proficiency in feature judgment between generated images and real images. **c** The results of three alternative models (Stable Diffusion, Imagen, and Tiger-N) evaluated using physician assessments and the CLIP evaluation method. The performance of these models is inferior compared to that of Tiger-F in Fig. 5b. Source data are provided as a Source Data file.

other methods. In contrast, other generative augmentation methods showed an initial rise in performance, followed by a decline when data augmentation reached a certain threshold. Tiger Model, however, demonstrated a more stable and sustained increase in predictive efficiency, with Tiger-F delivering the best results. A lack of diversity in the generated samples appeared to limit the sustained performance improvements of other models, leading to potential mode collapse. Tiger Model addressed this issue by enhancing its ability to accommodate a broader range of features, effectively elevating the upper limit of predictive performance.

Next, we evaluated the impact of sample size on the predictive performance of the Tiger Model. As illustrated in Fig. 5b, we compared the performance of binary classification models trained on different proportions of basic real data x (Task 1), Tiger Model-generated datasets y added to real data (Task 2,3), and models trained with an equal number of real samples Y added to the training set (Task 4). The results indicate significant improvements in the predictive ability of models utilizing Tiger Model sample augmentation, particularly in the benign-malignant classification of rare subtypes (FTC and MTC). The Tiger Model achieved a 7.61% and 10.18% improvement for FTC and

Table 3 | Comparative analysis of Tiger Model against other data augmentation methods in the classification of benign and malignant rare subtypes of thyroid cancer

Augmentation Method (Augmentation data number = 30 k)	AUROC (95% CI)					
	Benign vs PTC (Resnet50)		Benign vs FTC (Resnet50)		Benign vs MTC (Resnet50)	
	Valid	Test	Valid	Test	Valid	Test
① No Data Augmentation	0.8532 (0.83–0.93)	0.8677 (0.83–0.92)	0.7312 (0.68–0.75)	0.7364 (0.69–0.75)	0.7438 (0.63–0.78)	0.7523 (0.67–0.79)
Basic Data Augmentation						
② Basic image manipulation[38]	0.8763 (0.85–0.88)	0.8614 (0.84–0.87)	0.6999 (0.68–0.72)	0.6872 (0.65–0.69)	0.6988 (0.63–0.70)	0.6511 (0.63–0.66)
③ Mixing image[39]	0.8321 (0.81–0.84)	0.8563 (0.82–0.87)	0.6761 (0.65–0.69)	0.6321 (0.60–0.64)	0.6947 (0.64–0.72)	0.6827 (0.63–0.72)
④ Random erasing[40]	0.8426 (0.83–0.86)	0.8637 (0.82–0.89)	0.6590 (0.56–0.72)	0.6472 (0.62–0.73)	0.6743 (0.63–0.69)	0.7182 (0.65–0.72)
⑤ Feature space augmentation[41]	0.8532 (0.83–86)	0.8452 (0.81–0.87)	0.6842 (0.65–0.73)	0.6828 (0.60–0.72)	0.7083 (0.67–0.73)	0.7427 (0.72–0.75)
Deep Learning Generation						
⑥ PG-GAN[42]	0.8427 (0.82–0.86)	0. 8453 (0.82–0.86)	0.7022 (0.68–0.75)	0.6732 (0.63–0.72)	0.7529 (0.68–0.81)	0.7468 (0.72–0.79)
⑦ Diffusion Transformers[43]	0.8743 (0.84–0.92)	0.8703 (0.85–0.92)	0.7380 (0.71–0.75)	0.7328 (0.69–0.75)	0.7121 (0.65–0.73)	0.7252 (0.69–0.76)
⑧ Imagen[44]	0.8538 (0.82–0.86)	0.8412 (0.84–0.86)	0.7543 (0.72–0.80)	0.7577 (0.68–0.82)	0.7574 (0.74–0.82)	0.7581 (0.63–0.79)
⑨ Stable Diffusion[24]	0.8694 (0.84–0.89)	0.8503 (0.85–0.89)	0.7471 (0.73–0.79)	0.7737 (0.76–0.82)	0.7627 (0.74–0.82)	0.7582 (0.73–0.76)
⑩ Stable Diffusion + ControlNet[63]	0.8473 (0.84–0.89)	0.8527 (0.83–0.87)	0.7033 (0.65–0.75)	0.7172 (0.69–0.74)	0.7261 (0.69–0.76)	0.7425 (0.71–0.76)
⑪ Tiger-N	0.8882 (0.84–0.89)	0.8969 (0.84–0.92)	0.8067 (0.79–0.82)	0.8132 (0.78–0.83)	0.7733 (0.76–0.83)	0.8027 (0.79–0.82)
⑫ Tiger-F	0.9127 (0.90–0.93)	0.9263 (0.85–0.94)	0.8338 (0.76–0.83)	0.8442 (0.75–0.85)	0.8043 (0.75–0.83)	0.8234 (0.78–0.85)

① refers to the baseline without any data augmentation. ② refers to Flipping, Cropping, Rotation, Translation, and Noise injection operations. ⑥⑦ refers to the type of image-to-image generation models. ⑧⑨⑩⑪⑫ refers to the type of Text-to-image generation models. ⑧⑨⑩⑪⑫ are fine-tuned using our thyroid image-level features. ⑪ is the Tiger Model trained and generated solely using disease subtype names as prompts. ⑫ is a Tiger Model trained and generated using the nodule image's detailed description text as prompts.

MTC, respectively, compared to models without data augmentation, and performed only 1.8% worse on average in FTC and MTC when compared to models augmented with an equal number of real images. This suggests that Tiger Model-generated data closely aligns with the realism and diversity of real data distributions, making it a viable method for augmenting real data samples.

With 30,000 images each for benign and PTC cases, adding 10, 50, and 100 images of FTC and MTC led to improvements in malignancy diagnosis. Notably, adding 50 images increased the average AUC for classification by 0.07 or more (Supplementary Table 6). In addition, we varied the volume of benign and PTC data in the Tiger Model's training set to observe changes in model performance (Supplementary Table 7). Generated images obtained from training the Tiger Model exclusively on rare datasets are shown in Supplementary Fig. 2b.

External evaluation of the Tiger Model on public datasets

We carried out external evaluation experiments using both private and public ultrasound datasets to validate the generalizability of the Tiger Model (Tiger-F).

First, we assessed the model's performance on a benign-malignant binary classification task. The private data utilized includes 372 images from 274 patients diagnosed with Anaplastic Thyroid Carcinoma[46] (ATC, a rare thyroid cancer subtype), as well as images from patients with specific types of breast cancer, including Fibroadenoma of the Breast (benign; 2894 patients, 3987 images), Invasive Ductal Carcinoma (IDC; 2083 patients, 3468 images), Infiltrating Lobular Carcinoma (ILC; 77 patients, 100 images), and Papillary Carcinoma of the Breast (PCB; 389 patients, 639 images). Experiments were conducted separately on ATC data (trained on benign, PTC, FTC, MTC, and ATC; tested on ATC) and breast cancer data (trained on benign, IDC, ILC, and PCB; tested on IDC, ILC, and PCB, respectively), with the results summarized in Table 4. More details of the results can be found in the Supplementary Table 8

The results in Table 4 demonstrate that Tiger-F achieved significantly higher test AUC and improved calibration compared to other models. Notably, ATC served as out-of-distribution data relative to the thyroid data previously used, and the results indicated a 12.6% improvement in AUC over the baseline, highlighting the Tiger Model's capability to adapt effectively to out-of-distribution rare cancer types. The results on the three breast cancer subtypes, particularly the rare subtypes ILC and PCB, further validate the effectiveness of the Tiger Model in capturing relevant features specific to these conditions. Additional results from the external validation tests (Breast Ultrasound BrEaST and BUSI datasets) are provided in Supplementary Table 9.

Then, we expanded the task categories to a multi-class classification task. We used the VinDr-PCXR dataset, which comprises 9125 pediatric chest radiography studies retrospectively collected from a major pediatric hospital in Vietnam between 2020 and 2021[47]. Each scan has been manually annotated by experienced radiologists, identifying 36 critical findings and 15 diseases, with each abnormal finding marked by a rectangle bounding box on the image. To the best of our knowledge, this is a pediatric CXR dataset containing lesion-level and image-level labels for multiple findings and diseases. The dataset is segmented into 7728 training samples and 1397 test samples, making it suitable for algorithm development. The images were labeled for a total of 36 findings and 15 diagnoses. To test the capabilities of the Tiger Model, we used the 36 findings from the training set as prompts and paired them with images from the

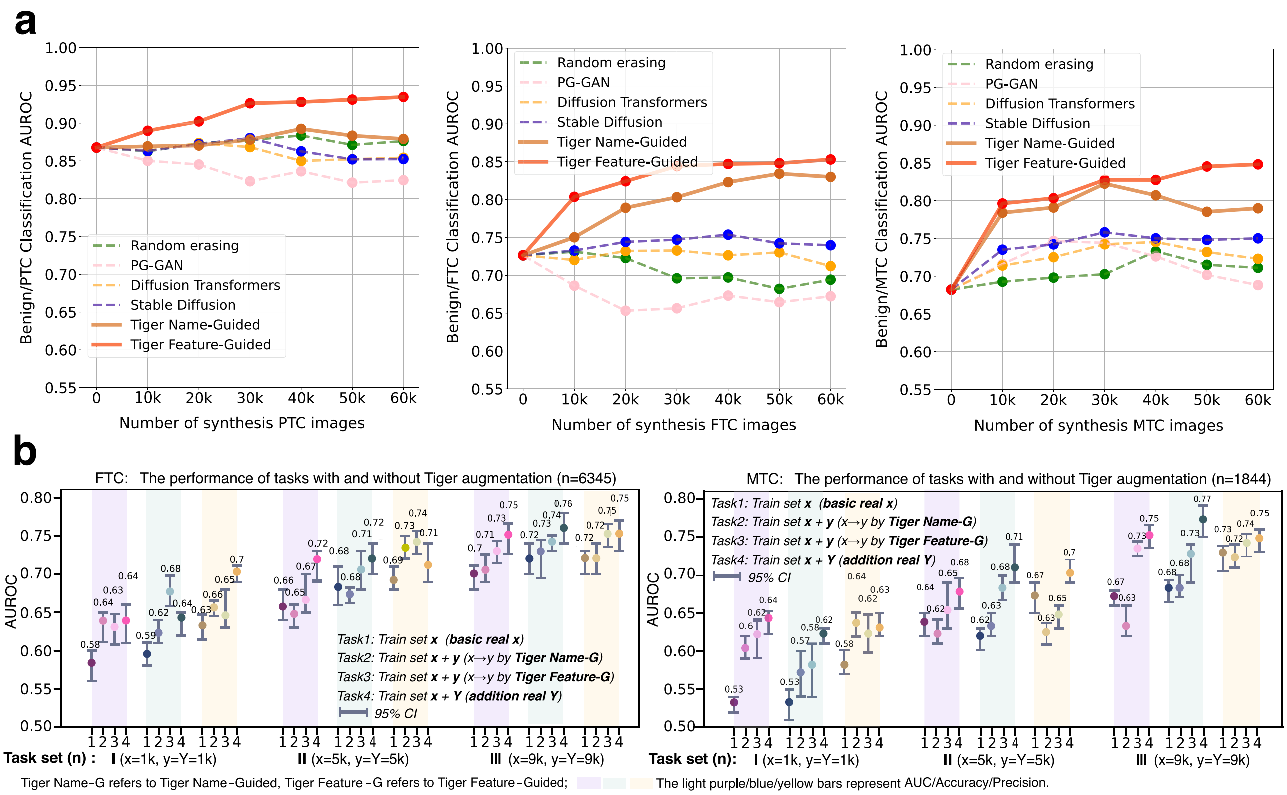


Fig. 5 | Effects of the number of generated images. a Comparative evaluation of the effect of downstream tasks based on different amplification ratios of different amplification methods in benign-malignant binary thyroid cancer prediction tasks. The results reveal performance limitations of other methods, possibly due to insufficient diversity in the detailed features of generated samples, while the Tiger Model shows substantial advantages by continuously improving performance. **b** Tiger data amplification and the addition of an equivalent amount of real image amplification were compared with the results of the downstream classification of rare subtypes. The FTC and MTC malignant tumor classification tasks showed that the model trained on the Tiger Model enhancement data was significantly better than the unamplified model, and the results were similar to the real image amplification. The AUC results are presented as mean values, with error bars representing 95% confidence intervals derived from $n = 50$ experimental replicates for each task setting. In each replicate trial, the basic real images (x) were selected through bootstrap sampling from the real image set. Source data are provided as a Source Data file.

Table 4 | External Evaluation of Tiger Model on private datasets with rare cancer subtypes

Augmentation Method	Benign vs ATC*		Benign vs IDC**		Benign vs ILC**		Benign vs PCB**	
	Test (ACC, 95%CI)	Test (AUC, 95%CI)	Test (ACC, 95%CI)	Test (AUC, 95%CI)	Test (ACC, 95%CI)	Test (AUC, 95%CI)	Test (ACC, 95%CI)	Test (AUC, 95%CI)
None Augmentation	0.7661 (0.69, 0.83)	0.7686 (0.72, 0.83)	0.8905 (0.89, 0.93)	0.8326 (0.80,0.86)	0.8158 (0.68, 0.92)	0.8011 (0.71, 0.98)	0.5000 (0.33, 0.67)	0.6889 (0.59, 0.93)
Stable Diffusion	0.8682 (0.83, 0.90)	0.8552 (0.82, 0.86)	0.9089 (0.90, 0.94)	0.8340 (0.81, 0.86)	0.8684 (0.76, 0.97)	0.8466 (0.72, 0.98)	0.6382 (0.60, 0.65)	0.7467 (0.70, 0.81)
Tiger-F	0.8468 (0.78, 0.90)	0.8658 (0.83, 0.88)	0.9344 (0.92, 0.95)	0.8670 (0.84, 0.89)	0.8421 (0.78, 0.88)	0.8665 (0.79, 0.88)	0.7857 (0.69, 0.87)	0.8533 (0.80, 0.89)

* the train set of augmentation models are {Benign (93220) + PTC (48585) + FTC (31725) + MTC (12910) + ATC (310)}.

** the train set of augmentation models are {Benign (3383) + IDC (2774) + ILC (80) + PCB (529)}.

training dataset. We selected four diseases (Bronchitis, Pneumonia, Bronchopneumonia, Bronchiolitis) and used their image-text pairs to train the Tiger Model. To validate the effectiveness of our method on rare cancer types, we under-sampled the Pneumonia class to form a 'rare' class, making its representation in the training set 10% of the most common class, Bronchitis. The results in Table 5 compare the four-class classification models with and without various types of data augmentation. After augmentation, the four-class classification task showed an improvement in ACC by 20.3%, and the AUC for category Bronchopneumonia improved by 13.4%. The lowest AUC improvement was 11.6% for category Bronchitis. In addition, the Brier Score decreased by 44.7%. It is evident that the Tiger-F method achieved the best performance, thereby validating the effectiveness of the proposed model architecture on the dataset.

Discussion

This study conducted training and validation of a text-guided image generation model, Tiger Model, marking an instance of its application in improving the diagnosis of rare subtypes. Indeed, although the incidence rate of individual rare tumors might be low, the total number of individuals affected by rare diseases is substantial around the world[5]. Diagnostic prediction differences in specific populations can lead to misdiagnosis, missed diagnosis, and poor prognosis, raising issues of unfairness[10]. In clinical practice, the diagnosis of rare subtypes often

Table 5 | External Evaluation of Tiger Model on VinDr-PCXR dataset

Augmentation Methods	VinDr-PCXR: Four-Classification {Bronchitis, Pneumonia, Broncho-pneumonia, Bronchiolitis}(ResNet-50)						
	Test (ACC, 95%CI)	Test (AUC, 95%CI)					Brier Score
		Mean	Bronchitis	Pneumonia	Broncho-pneumonia	Bronchiolitis	
None Aug	0.6842 (0.66–0.71)	0.7031 (0.66–0.72)	0.6982 (0.67–0.73)	0.6742 (0.64–0.69)	0.7173 (0.70–0.73)	0.6491 (0.63–0.67)	0.3321
Stable Diffusion	0.7301 (0.72–0.74)	0.7423 (0.72–0.75)	0.7214 (0.71–0.73)	0.7341 (0.71–0.74)	0.7323 (0.71–0.75)	0.6930 (0.68–0.71)	0.2472
Tiger-F	0.8229 (0.81–0.84)	0.8303 (0.82–0.85)	0.7791 (0.76–0.79)	0.8302 (0.83–0.84)	0.8131 (0.80–0.83)	0.7291 (0.70–0.73)	0.1837

The number of augmentation image are Bronchitis (0) Pneumonia (606) Bronchopneumonia (237) Bronchiolitis (275). Generate data to reduce the imbalance between classes.
All generation models are trained by VinDr-PCXR train set including images and texts. 'Stable Diffusion' refers to Stable Diffusion methods with image-level feature prompts. 'Tiger-F' refers to Tiger Model with image-level feature as prompts. Prompts are 36 findings of chest radiography, and the foreground-background segmentation model for chest X-rays uses TorchXRayVision[71].

requires detailed observation of disease characteristics. Healthcare professionals frequently misjudge rare subtype cases as common subtypes due to their similar features. For example, Follicular Thyroid Carcinoma (FTC) often resembles benign nodules, and Medullary Thyroid Carcinoma (MTC) cases can be similar to Papillary Thyroid Carcinoma (PTC) cases[48,49]. Misdiagnosis or missed diagnosis can severely impact prognosis. For example, MTC can easily metastasize to regional lymph nodes and distant organs such as the lungs and bones. In such scenarios, accurate preoperative diagnosis is highly beneficial. On the other hand, Ultrasonography is one of the preferred non-invasive methods for determining the nature and type of thyroid tumors in clinical practice[13]. Ultrasound is more cost-effective compared to CT and MRI, and can clearly display key features for benign– malignant diagnosis including size, location, boundaries, and internal structure (such as echogenicity, calcification, and vascular distribution) of thyroid nodules[50,51]. Meanwhile, it can guide fine-needle aspiration biopsy (FNAB), increasing the accuracy and safety of sampling, which is particularly important for diagnosing indeterminate nodules[52]. Ultrasound can also assess the condition of cervical lymph nodes, helping to evaluate the presence of metastatic lesions, which is crucial for staging and treatment planning of thyroid cancer[53,54]. Therefore, improving the diagnostic performance of thyroid ultrasound imaging is of significant clinical importance, especially for the diagnosis of rare subtypes.

The scarcity of rare disease samples can lead to issues of insufficient sample diversity and representativeness, which is addressed by this study. Literature examining the ultrasonic diagnosis of thyroid cancer subtypes often relies on a comprehensive judgment based on the co-occurrence and synergistic features of ultrasonography[55]. The training of the generative model is a process of systematically arranging and combining ultrasonographic features according to disease subtypes. This approach enabled the generation of images corresponding to rare subtypes through clinical knowledge-guided image generation. In diagnostic tasks, the Tiger Model exhibited a notable improvement in detecting rare subtypes, with an increase of over 10%. It is pertinent to note that the features of rare subtypes are mostly composed of combinations and permutations of benign and common features, along with their own unique characteristics[56,57]. The incorporation of text-guided feature training with extensive sample sizes significantly aids the model in migrative learning of features in rare diseases, which effectively addresses the challenges posed by sample scarcity in training. Consequently, the Tiger Model has enhanced the generalization ability of diagnostic models across different subtype data, demonstrating effectiveness across multiple datasets and ensuring equitable benefits for specific populations. This study primarily focuses on the application of the Tiger model in thyroid cancer. However, external evaluations also demonstrate that the Tiger model exhibits excellent generalization performance and can be effective in other cancer types.

Compared to conventional image generation methods, text-guided image augmentation yields more diverse and clinically relevant data, significantly elevating the predictive performance of models. Our experimental results reveal that traditional training methods predominantly rely on image characteristics, which poses a challenge in generating varied images from the limited samples available for rare diseases (more high-quality and defective generated ultrasound images of thyroid in the Supplementary Fig. 3). In diagnostic tasks, the accuracy of conventional methods plateaus after 30k images. Conversely, the Tiger Model leverages text-guided image generation to produce a broader range of images with richer features, leading to continued enhancements in predictive accuracy. Rare diseases, characterized by their low prevalence and insufficient sample sizes, often lack cases demonstrating specific feature combinations. This is a key obstacle for prevalent research that employs category-level text descriptions to guide the generation of data for prevalent diseases or common subtypes in this area, because the efficiency of feature extraction from existing images is limited, which can be remedied through supplemental guidance using extensive, literature-based descriptions of disease characteristics. This approach aims to align the sample distribution more closely with actual datasets. In addition, images generated from literary sources carry significant medical interpretability and clinical relevance, offering potential for application in clinical studies where actual samples are scarce but theoretical support exists.

This study designed three types of Turing Tests to comprehensively evaluate the authenticity and diversity of generated images, as well as the model's control over feature details and user comprehensibility. Compared to other state-of-the-art generative sample amplification methods, the Tiger Model shows superior performance of more than 10% higher in terms of scores for realism and diversity, lending to its level of being clinically comprehensible. In addition, the ability of the Tiger Model to control fine details in image generation enhances user interaction, improving the system's operability and editability. Such features make the Tiger Model particularly suitable for practical clinical applications in the medical field. The Tiger Model holds significant potential for assisting doctors in accurately identifying detailed disease characteristics, particularly when dealing with rare diseases and subtypes. It can aid in confirming diagnoses and reducing the risk of misdiagnosis and missed diagnoses. This is crucial for enhancing the overall quality of healthcare services and ensuring patient safety. In addition to assisting with diagnosis, doctors can utilize clinically guided disease feature generation to visualize and explain the decision-making process. This helps doctors understand the complex manifestations and detailed differences of diseases, particularly in cases that are less frequently encountered in daily practice. Through interaction with the advanced model, clinicians may also continuously update their knowledge base and experience level (Supplementary Fig. 4 for more details).

The study validated the few-shot generation capabilities of the Tiger Model using a small dataset. The results demonstrated that even with a limited set of 20 images, the model could generate diverse, realistic, detailed, and comprehensible images, contributing to improved diagnostic performance. Currently, there are numerous advanced artificial intelligence algorithms designed for diseases with abundant standardized data. These algorithms rely heavily on high-quality medical sample annotations for initial training, especially in the burgeoning era of large-scale models[58]. However, this ideal often deviates from reality. Statistics indicate the existence of hundreds of rare diseases and subtypes[59], where data collection is hampered by limited samples, patient privacy concerns, and data security challenges. Moreover, the annotation of medical images incurs substantial costs, requiring specialized tools and the expertise of medical professionals[60]. A representative existing method utilizes semantic knowledge from diffusion models to identify semantic correspondences in natural images, which is locating positions with the same semantic meaning across multiple images[61]. This approach optimizes the prompt embeddings of generative models to maximize attention to regions of interest. These optimized embeddings capture semantic information about locations, which can then be used to prompt semantic positions on another image. For datasets with little segmentation labels, this type of approach can be combined with the proposed Tiger Model to achieve text-guided segmentation of nodule regions. Despite the scarcity of data in rare subtypes, the feature distribution for specific attributes remains relevant and in-domain, consistent with clinical diagnosis. During the training of the Tiger Model, textual guidance assists in the disentanglement of features and enables the reutilization of attributes learned from common subtypes. This method shows potential for advancing the understanding of rare diseases and influencing the development of related health policies in the future.

This study, however, is not without limitations and areas for further exploration. Primarily, due to constraints in sample collection, the study's focus was restricted to thyroid cancer subtypes. Future research should extend this approach to a wider range of rare diseases and patient subgroups, including different age, gender, nation or ethnic groups, to evaluate its broader medical significance. Additionally, while this study concentrated on the improvement in diagnosing rare diseases through text-guided image generation, it did not provide a detailed quantitative analysis of the relationship between the number of generated samples and predictive accuracy. Future studies should investigate these dynamics more thoroughly. Furthermore, during the training process, a small fraction of generated samples exhibited quality issues, likely attributable to the unique challenges of medical imaging. Future efforts will aim to identify factors influencing medical image quality and develop strategies to address these challenges.

Tiger Model, as a text-guided image generation methodology, is adept at synthesizing clinically relevant and trustworthy high-fidelity medical data based on detailed disease feature descriptions. Its ability to control the generation of lesion features at a granular level is crucial for disease understanding and facilitating user-friendly human-machine interactions. Looking ahead, it is anticipated that the Tiger Model will find applications in the analysis of rare diseases and subgroups, improving equitable access for minority populations and enhancing the system's generalizability and expedite the translation of medical AI innovations into practical applications.

Methods

Ethical issues

This study was approved by the institutional review board (IRB) of Shanghai Tong Ren Hospital and undertaken according to the Declaration of Helsinki. Informed consent from patients with thyroid cancer and controls was exempted by the IRB because of the retrospective nature of this study.

Data collection and literature collection process

We retrospectively collected preoperative thyroid ultrasound images of patients undergoing thyroidectomy. Inclusion criteria include: age between 18 and 75 years old, with a median age of 38, and female patients accounted for 63%, thyroid nodules size ranging from 0.1 cm to 1.9 cm, with an average size of 0.44 cm, postoperative pathologically confirmed thyroid cancer; 68,386 patients met one of the following conditions: (1) postoperative pathologically confirmed thyroid cancer or benign diagnosis; (2) After more than one year of follow-up by experienced radiologists, it was diagnosed as benign. After inclusion, patients who met one of the following criteria were excluded: (1) other life-threatening disease comorbidities; (2) Preoperative ultrasound reports were lacking or incomplete; (3) Poor ultrasonic image quality; (4) Disputed pathological diagnosis; (5) History of thyroidectomy or other head and neck tumors. The patient's demographic information, Thyroid Imaging Reporting and Data Systems (TIRADS) grade[62], and postoperative pathology were collected at the same time. Specifically, for benign patients, there are two groups: one group consists of patients who underwent surgery and were confirmed to be benign (initially suspected to have a high likelihood of malignancy), and the other group includes patients who were confirmed to be benign through 5 years of follow-up. Sex information was based on self-report and ultrasound report, but not included as variables in the analytical framework, as the study focused on rare thyroid images generations. We included all eligible patients from four top-tier hospitals from July 2013 to October 2023 (List of hospitals see Supplementary Method 2) and divided the patients into training set and a validation set, and included patients from six independent institutions as the external test set. The acquisition and quality control of thyroid ultrasound images were conducted using various ultrasound machines, with specific model details and quality control measures outlined in Supplementary Method 3. Our study population came from different regions of China, including multi-ethnic groups. 50 senior physicians supervised the entire data collection process and data Quality control, all of which passed Quality validation by doctors. Their average clinical duration was more than 5 years, and the correct diagnosis rate of common thyroid cancer subtypes (PTC) was more than 87.9%. TIRADS scores were evaluated and recorded by experienced sonographers. The correspondence between text and features is derived from the descriptions in the ultrasound reports. If descriptions are lacking, they are added by two ultrasound specialists with over five years of experience, based on the ultrasound images and corresponding pathological diagnosis reports. Patient information is used to link the ultrasound images, reports, and pathological information. The subtyping diagnoses from the ultrasounds are confirmed by pathological results.

The literature search conducted in this study involved keywords such as "thyroid ultrasonic image", "PTC," "FTC," "MTC," and similar terms. Articles were selected based on the presence of statistical tables illustrating ultrasonic features or the inclusion of ultrasonic images with specific feature descriptions in the text. The results of the literature collection are presented in the Supplementary Table 10, encompassing summaries of Benign characteristics, PTC features, FTC features, MTC features, and a collection table for the four ultrasound image prompts (Supplementary Data 1).

The utilized textual reports are divided into those from the hospital and those from the literature. The reports from the hospital were generated by ultrasound specialists and clinical surgeons with over 5 years of experience. All the reports were carefully verified by 20 ultrasound specialists and clinical surgeons using standardized terminology. The data processing details of the Condition FG image and Condition BG image are in Supplementary Method 4.

Tiger Model architecture

The devised Tiger Model primarily consists of three functional modules, namely the Coarse-Training module, Fine-Training module, and

Attentional Fusion module. The image generated by Coarse-Training is displayed in Supplementary Fig. 2a. The Fine-Training module consists of an FG-Encoder and a BG-Encoder, each of which incorporates a ControlNet structure formulated by replicating and finetuning the encoder and middle blocks from the pretrained Stable Diffusion model, following[63]. The outputs of both the FG-Encoder and BG-Encoder are channeled into an Attentional Fusion module for integration before being incorporated into the middle block and the decoder of the Stable Diffusion model. Within both FG-Encoder and BG-Encoder, we implement a condition encoder f_e for text conditions (prompt) and f_i for image conditions (mask image or Sobel map).

Weighted-spatialtransformer

We implement the basic Stable Diffusion (SD) model as a Coarse-Training module[24]. We utilized a standard U-Net model with encoder blocks (12 blocks), a middle block (1 block), and decoder blocks (12 blocks) to form the backbone of the SD model. Denote the U-Net model as ϵ_θ. During inference, given a text prompt y, a text encoder (CLIP encoder) first extracts its text representation $f_e(y)$, With classifier-free guidance (CFG)[64] and Denoising Diffusion Implicit Models (DDIM) sampling[65], the U-Net predicts a latent representation $\epsilon_\theta(f_e(y))$, which is then used by an image decoder to construct the output image via the latent representation.

The latent representation is obtained through the Stable Diffusion process[24]. Through time $t \in \text{Uniform}\{1, \ldots, T\}$, the model ϵ_θ predicts a denoised variant of its input image x_t, which is a noisy version of the original image x with noise ϵ. It is learned through the objective:

$$\mathcal{L} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,I),t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\right] \tag{4}$$

Where, $\mathcal{N}(0, I)$ refers to the standard normal distribution. However, in our thyroid ultrasound data, the images directly generated by the coarse trained Tiger Model are plagued by issues pertaining to details. For example, the edges of nodules are often deformed; The nodules are usually cavernous and do not contain the details of the lesion; The thyroid is often simple in shape and lacks diversity; The authenticity of surrounding tissues, such as trachea and blood vessels are insufficient (see Supplementary Fig. 1a).

To overcome the issues pertaining to detail in Stable Diffusion model-generated images, we borrowed from ControlNet and implemented two symmetric ControlNet structures, namely an FG-Encoder and a BG-Encoder. The two Encoders are formulated to implement different functions in the Fine-Training module, each containing 12 Encoder Blocks (three blocks per size, encompassing sizes 64 × 64, 32 × 32, 16 × 16, and 8 × 8). FG-Encoder is used for detailed feature control of nodule generation, while BG-Encoder is used for background detail generation.

During model fine-training, we used segmentation model to segment the tumor (foreground, abbreviated as FG) region, generating the Condition FG image (here we employ YoloV8[66]), while the Sobel edge detection algorithm[67] is employed to generate feature edge maps for obtaining the region outside the tumor (background, abbreviated as BG), resulting in the Condition BG image. The segmentation model serves to localize the relative positions of the thyroid and nodules within the training dataset, supplementing textual descriptions with detailed descriptions of nodule locations. Particularly, by analyzing the location statistics of nodules, they can be categorized into the inferior pole, middle pole, or upper pole in the "Nodule Location" section. During downstream image generation, models trained with detailed location descriptions can customize the shape and position of the nodules relative to the thyroid according to user requirements. The prompt includes detailed descriptions of foreground-background information, such as morphological features, positional details, and multiple characteristics guiding clinical assessments of malignancy levels (Fig. 1 and Supplementary Table 1). These conditions prompts aid in optimizing the model to generate images with more precise detailed features that better align with the actual data distribution.

The Condition FG image and Condition BG image are separately fed into the FG-Encoder and BG-Encoder. In addition, images with added noise are simultaneously directed into both encoders. Furthermore, the Condition FG prompt and Condition BG prompt undergo embedding before entering the model. Adjustments are made to the Cross Attention weights during forward propagation of the FG-Encoder to enhance focus on rare samples. Moreover, a Fusion Module is integrated between the FG-Encoder and BG-Encoder, enabling feature fusion of corresponding outputs from the FG-Block and BG-Block at each layer. These fused features, combined with outputs from the SD-Encoder, enter the SD-Decoder. Injected noise latent vectors are inputted into the U-Net to estimate noise, and the KL divergence loss is computed by comparing these estimated noise outputs with the true noise information labels, subsequently updating the model parameters. Zero convolution layers, initialized with zero weights to protect the pretrained backbone model from initial training noise, are gradually optimized during ControlNet training through backpropagation[63].

An image x fixed with a Markov chain that gradually adds Gaussian noise and x_t (t is uniform$\{1 \ldots T\}$) is considered a noisy version. For a conditional image, we first obtain the condition FG image x_{FG} and the BG image $x_{\mathrm{BG}}\mathtt{from}x$. x_{FG} used pretrained yolov8 to mask the nodule regions of x, the mask map input contains the position and shape information of the nodule. x_{BG} is the background map input of x generated by the Sobel edge detection algorithm. Then, $x_t, x_{\mathrm{FG}}, x_{\mathrm{BG}}$ are encoded through distinct encoders f_i into z_t, z_{FG} and z_{BG}. Meanwhile, for the given prompt y, we use keyword detection based on keywords shown in Table 2. We divide y into two prompts, describing the edge and internal morphology of the nodule and background information, as y_{FG} and y_{BG} respectively. Through the text encoder to get their representation $f_e(y_{\mathrm{FG}})$ and $f_e(y_{\mathrm{BG}})$, f_e here is a type of the public training CLIP model of text encoder[37].

FG-Encoder takes $z_t^{\mathrm{FG}} = z_t + z_{\mathrm{FG}}$ as input and $f_e(y_{\mathrm{FG}})$ as a condition to infer the characteristics of the detailed content of the nodular region, denoted as F_{FG}; BG-Encoder takes $z_t^{\mathrm{BG}} = z_t + z_{\mathrm{BG}}$ as input and $f_e(y_{\mathrm{BG}})$ as a condition to infer the detailed content of the background region, denoted as F_{BG}. Within both FG-Encoder and BG-Encoder, we mapped the condition to the intermediate layers of the model through a cross-attention layer, based on the following formula:

$$Q = \varphi_i\left(z_t^{\mathrm{FG}}\right) W_Q^i \tag{5}$$

$$K = f_e\left(y_{\mathrm{FG}}\right) W_K^i \tag{6}$$

$$V = f_e\left(y_{\mathrm{FG}}\right) W_V^i \tag{7}$$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{W(QK^T)}{\sqrt{d}}\right) V \tag{8}$$

Where, $\varphi_i\left(z_t^{\mathrm{FG}}\right)$ is the (flattened) intermediate representation at the i^{th} layer, while W_Q^i, W_K^i and W_V^i are the learnable parameters of the attention layer; d is the representation dimension; $W \in \{W_{\mathrm{FG}}, W_{\mathrm{BG}}\}$ is an auxiliary weight matrix used to scale the impact of different keywords in y. In BG-Encoder, $W_{BG} = I$ is an identity matrix; In FG-Encoder, $W_{\mathrm{FG}} = \{\omega^{ij}\}$ is determined based on the following formula:

$$\omega^{ij} = a + \frac{b}{1 + e^{b\delta_j}} \tag{9}$$

Where, ω^{ij} can be viewed as the scaling factor of the attention weight between token$_i$ and token$_j$; a and b are hyperparameters determined through experiment, we set a = 0.6 and b = 5, b is approximately obtained from rare subtypes samples/total samples; δ_j is determined based on the following formula:

$$\delta_j = \begin{cases} 0\,, \mathrm{token}_j \in \left\{ f_{\mathrm{txt}}\left(y_{\mathrm{FG}}^{\mathrm{pr}}\right) \right\} \\ \tau_j\,, \mathrm{token}_j \in \left\{ f_{\mathrm{txt}}\left(y_{\mathrm{FG}}^{\mathrm{jn}}\right) \right\} \\ 1, \qquad\qquad \text{otherwise}. \end{cases} \tag{10}$$

Here, τ_j is the term frequency of token$_j$ in the one training batch[68], which is the number of times a word appears divided by the total amount in one batch. $y_{\mathrm{FG}}^{\mathrm{pr}}$ and $y_{\mathrm{FG}}^{\mathrm{jn}}$ refers to the particular terms and joint terms (according to Supplementary Table 1) within y_{FG} respectively. 'Otherwise' means common terms and others. In the experiment, during training, we calculate τ_j within every batch.

Attentional fusion

The Attentional Fusion module integrates the outputs of FG-Encoder and BG-Encoder in ControlNet module and incorporate them into the Stable Diffusion module.

Due to the differing focus and scope of information encoding between the FG-Encoder and BG-Encoder, the features extracted by them cannot simply be concatenated together. Instead, they require interdependent training of foreground and background features. Specifically, the foreground is determined based on background details to define the edge morphology, while the background generates textures of reasonable scales based on the foreground position. To facilitate this interaction, we employed an attentional fusion module, composed of two convolutional modules.

The Fusion Module[69] is responsible for integrating the foreground feature F_{FG} and the background feature F_{BG} during the training process of the two encoders. This module strengthens the connection between foreground and background, avoiding later separate generation processes from producing unrelated errors that deviate from natural data patterns. The blue blocks represent the Fusion Module, which consists of two convolutions (Fig. 2). Conv1 comprises {Point-wise Conv + ReLU + Point-wise Conv}, while Conv2 comprises {GlobalAvgPooling + Point-wise Conv + ReLU + Point-wise Conv}.

$$F_{\mathrm{fusion}} = C(F_{\mathrm{FG}} \circledcirc F_{\mathrm{BG}}) \otimes F_{\mathrm{FG}} \oplus (1 - C(F_{\mathrm{FG}} \circledcirc F_{\mathrm{BG}})) \otimes F_{\mathrm{BG}} \tag{11}$$

$$F_{\mathrm{FG}} \circledcirc F_{\mathrm{BG}} = F_{\mathrm{FG+BG}} \otimes \mathrm{sigmoid}\left(\mathrm{Conv1}\left(F_{\mathrm{FG+BG}}\right) \oplus \mathrm{Conv2}\left(F_{\mathrm{FG+BG}}\right)\right) \tag{12}$$

Here, ⊕ denotes the broadcasting addition; ⊗ denotes the element-wise multiplication; $F_{FG+BG} = F_{FG}F_{BG}$. When feeding F_{fusion} back into Stable Diffusion, we employ CFG Resolution Weighting. The final output of the model is:

$$F_{\mathrm{PRED}} = F_{\mathrm{SD}} + \beta_{\mathrm{CFG}}(F_{\mathrm{fusion}} - F_{\mathrm{SD}}) \tag{13}$$

Where, F_{SD} is the direct output of the Stable Diffusion module, F_{fusion} is the output after passing through the Fusion Module, and β_{CFG} represents the scaling hyperparameter introduced by CFG metric.

Training details

During model training, we initially train the basic Stable Diffusion model with the <prompt-image> pairs to enable the model to learn the general characteristics of thyroid cancer ultrasound images. Importantly, during Coarse-Training, the Fine-Training Module is not included in the training process. After completing the pre-training, to achieve controllability in the generated images, we further train the Coarse-Training module within the Tiger Model using both the prompt and the generated condition image as condition information. More Training parameter details about the Tiger Model are in Supplementary Method 4.

The FG-Encoder needs to focus on generating the nodular part. Therefore, we use a masked image to highlight the nodular area, which is then passed into the model for targeted training. During the training process, only the parameters of the FG-Encoder and the FG-Middle sections are updated. The loss function guiding the updates to the FG-Encoder model during training is as follows:

$$\mathcal{L}_{\mathrm{FG}} = \mathrm{MSE}(\hat{x}_{\mathrm{FG}}, x_{\mathrm{FG}}) \tag{14}$$

During training, the model only updates parameters related to the BG-Encoder and BG-Middle sections. The loss function guiding the model's updates based on the BG-Encoder model during training is as follows:

$$\mathcal{L}_{\mathrm{BG}} = MSE(\hat{x}_{\mathrm{BG}}, x_{\mathrm{BG}}) \tag{15}$$

In the experiment, due to a limited variety of corresponding prompts for background organization, achieving controlled generation was challenging. Therefore, Sobel edge detection operators were utilized to retain essential structural attributes of the images. These edge detection operators were incorporated as crucial distribution information for the background during joint training with prompts (illustrative effects of Sobel edge detection operators can be observed in Supplementary Fig. 5). Sobel edge detection operators derive essential structures from real thyroid images, such as trachea, thyroid lobe, carotid artery, resembling sketches or outlines.

Notably, both FG-Encoder and BG-Encoder are equipped with individual optimizers (both utilizing AdamW) and undergo parameter updates by backpropagating their respective losses. The training is conducted using 32 A100 Graphics Processing Units (GPUs) and continuous for 82 h. More Training parameter details about YoloV8, CLIP, Resnet50 are in Supplementary Method 5.

Inference details

Tiger Model's application scenarios (inference) can be divided into two categories (Supplementary Fig. 6). The first type is Diversify Inference, which involves generating thyroid feature textual prompts based on prompt input combinations. Tiger Model generates synthetic images based on the prompt content, controlling the synthesis of corresponding fine-grained foreground-background features within the model. The second type is Refinement Inference, where the input comprises real images. Tiger Model generates images consistent with the subtype of the input image. Both generation scenarios allow for the control of corresponding foreground-background features as needed during the generation process.

During the Inference process, the FG-Encoder and BG-Encoder trained in this context can be used separately or in combination. To meet the data augmentation requirements, the choice can be made to solely utilize the FG-Encoder to diversify nodules within images. However, it's crucial to consider that when the augmentation quantity is substantial, leaving the background unchanged may lead to model overfitting and local optima. Therefore, we recommend a methodology that involves using the FG-Encoder to generate the foreground and subsequently using the BG-Encoder to generate the background. All the results in the results section of this work follow this approach. Generate foreground and background pictures respectively, and display them in Supplementary Fig. 5b.

Source of the text used by the FG-Encoder in the Inference process: During the augmentation process, the diversity and accuracy of the text directly impact the final model's performance. In this task, for the substantially larger datasets of Benign and PTC, the text used during inference is randomly selected from the entirety of the Benign and PTC datasets. However, for the smaller datasets of FTC and MTC, leveraging prior knowledge from medical literature summarizing the characteristics of these rare subtypes and instances appearing in article illustrations serves as descriptions for FTC and MTC. This ensures that the guiding text for the generated images is derived from real-world existent descriptions. (Supplementary Method 6).

Source of the Sobel maps used by the BG-Encoder in the Inference process: In background generation, the Sobel maps determine the specific positions and shapes of various areas within the background. It is recommended to derive Sobel maps from authentic data sources, as manually created Sobel maps may introduce significant distortions. Sobel maps can be derived from a substantial volume of existing data, such as benign, PTC, and normal thyroid images. Sobel maps extract a large number of common data backgrounds as the background of rare data, which greatly improves the diversity of rare data backgrounds and ensures the effectiveness. See Supplementary Fig. 5a for more details. We opt to generate Sobel maps from real images (i.e., Benign and PTC), ensuring that surrounding organs appear in their correct positions during background generation. The Tiger Model then complements different subtypes nodule details based on this generated structure. During the experiment, we guarantee that the Sobel maps are not come from any test set and verification set, and test set and verification set are completely independent. In background generation, we distinguish between two kinds of ultrasonic positions, one is transverse and the other is longitudinal. This setting follows standard practices in the clinician's diagnostic process. See Supplementary Fig. 5b for more details. Ablation study of the Tiger model is presented in Supplementary Table 13.

To effectively generate the attributes of Nodule Location and Extension (Supplementary Table 1), we rely on a well-trained thyroid shape segmentation model to obtain the thyroid's segmented region(Supplementary Fig. 7). The results of different segmentation methods are compared in Supplementary Tables 11 and 12. During FG-Inference for Condition FG image preparation, we alter the mask's region (white area) positions. For the Nodule Location definition, altering the mask's position relative to the thyroid's edge is necessary based on textual guidance. The thyroid region is vertically divided into three equal parts—the inferior pole, middle pole, and upper pole. For instance, when the text specifies "middle pole," the model randomly changes the mask region's position while ensuring its centroid remains within the middle pole area. Implementing Extension involves setting the mask region's edge to intersect with the upper edge of the thyroid, defining the maximum distance of intersection within the range of [0-10] indices, as depicted in Supplementary Fig. 5c.

The evaluation criteria used in this article, including AUROC, Accuracy, Sensitivity, Specificity, CLIP Score, SSIM, CMMD, GS, D&C, FID, IS, Brier Score, among others, are detailed in Supplementary Method 7.

Reporting summary

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

Data availability

All data supporting the findings of this study are available within the article and its Supplementary Information files. The minimum thyroid dataset required to interpret, verify, and extend the results of this study – including model predictions and performance metrics – has been deposited in Hugging Face under accession code: https://huggingface.co/datasets/FangDai/Thyroid_Ultrasound_Images. The Source Data file containing the detailed model outputs and key evaluation metrics is also available at https://github.com/fangdai-dear/Tiger-Model/blob/master/dataset/SourceData.xlsx. This includes: - Pre-processed imaging data (ultrasound images with anonymized metadata). - Clinical feature tables (age, gender, tumor size) with all direct identifiers removed. -Due to ethical restrictions and patient confidentiality agreements, the full dataset (e.g., raw imaging data, detailed clinical records) cannot be made publicly available. This pertains to detailed clinical records and high-resolution imaging data that, even after de-identification, may pose a risk of re-identification given the unique characteristics of thyroid cancer cases. Researchers who wish to access additional data for non-commercial academic purposes may submit a formal request to the corresponding author. Requests will be reviewed by the institutional ethics committee and data custodians. The following conditions apply: -Purpose: Data will only be shared for research purposes that align with the original study objectives. -Access Restrictions: Requesters must sign a data use agreement prohibiting re-identification or redistribution. -Data Retention: Approved data will be available for 2 years from the date of publication. -The chest X-ray dataset is sourced from https://physionet.org/content/vindr-pcxr/1.0.0/, and the breast cancer ultrasound dataset used in the Supplementary is available from https://github.com/best-ippt-pan-pl/BrEaST/ and https://scholar.cu.edu.eg/?q=afahmy/pages/dataset. The data for each figure/table in this study are included in the Source Data section, with the file named Source Data.xlsx. This file can also be downloaded from the following link: https://github.com/fangdai-dear/Tiger-Model/blob/master/dataset/Source
Data.xlsx. Source data are provided in this paper.

Code availability

The code and models are available on both GitHub (https://github.com/fangdai-dear/Tiger-Model.git) and Hugging Face (https://huggingface.co/FangDai/Tiger-Model). Installation instructions are provided in both repositories. We have provided the permanent reference for the version of the code used in this study[70]. The code is open-source and released under the Apache License 2.0, which allows free use, modification, and redistribution under its terms. The implementation is based on multiple publicly available open-source projects. We have retained all original license information and copyright notices in the corresponding source files. Specifically, we acknowledge the following contributions: Diffusers by HuggingFace (Apache 2.0): https://github.com/huggingface/diffusersStable Stable Diffusion by Rombach et al. (Apache 2.0): https://github.com/CompVis/stable-diffusion. ControlNet (MIT): https://huggingface.co/papers/2302.05543. Fusion Module by Yimian Dai (MIT): https://github.com/YimianDai/open-aff.

References

1. Ricci Lara, M. A., Echeveste, R. & Ferrante, E. Addressing fairness in artificial intelligence for medical imaging. *Nat. Commun.* **13**, 4581 (2022).
2. Rajpurkar, P., Chen, E., Banerjee, O. & Topol, E. J. AI in health and medicine. *Nat. Med.* **28**, 31–38 (2022).
3. Willemink, M. J. et al. Preparing medical imaging data for machine learning. *Radiology* **295**, 4–15 (2020).
4. Bria, A., Marrocco, C. & Tortorella, F. Addressing class imbalance in deep learning for small lesion detection on medical images. *Comput. Biol. Med.* **120**, 103735 (2020).
5. Stark, Z. & Scott, R. H. Genomic newborn screening for rare diseases. *Nat. Rev. Genet.* **24**, 755–766 (2023).
6. Jiang, X., Hu, Z., Wang, S. & Zhang, Y. Deep learning for medical image-based cancer diagnosis. *Cancers* **15**, 3608 (2023).
7. Chen, R. J. et al. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nat. Biomed. Eng.* **7**, 719–742 (2023).
8. Seyyed-Kalantari, L., Zhang, H., McDermott, M. B., Chen, I. Y. & Ghassemi, M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat. Med.* **27**, 2176–2182 (2021).
9. Ktena, I. et al. Generative models improve fairness of medical classifiers under distribution shifts. *Nat. Med.* **30**, 1166–1173 (2024).

10. Yao, S. et al. Enhancing the fairness of AI prediction models by Quasi-Pareto improvement among heterogeneous thyroid nodule population. *Nat. Commun.* **15**, 1958 (2024).
11. Varoquaux, G. & Cheplygina, V. Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ Digit. Med.* **5**, 48 (2022).
12. Liang, W. et al. Advances, challenges and opportunities in creating data for trustworthy AI. *Nat. Machi. Intell.* **4**, 669–677 (2022).
13. Chen, D. W., Lang, B. H., McLeod, D. S., Newbold, K. & Haymart, M. R. Thyroid cancer. *Lancet* **401**, 1531–1544 (2023).
14. Grani, G., Lamartina, L., Durante, C., Filetti, S. & Cooper, D. S. Follicular thyroid cancer and Hürthle cell carcinoma: challenges in diagnosis, treatment, and clinical management. *Lancet Diabetes Endocrinol.* **6**, 500–514 (2018).
15. Pelizzo, M. R., Mazza, E. I., Mian, C. & Merante Boschin, I. Medullary thyroid carcinoma. *Exp. Rev. Anticancer Ther.* **23**, 943–957 (2023).
16. Vuong, H. G., Le, M. K., Hassell, L., Kondo, T. & Kakudo, K. The differences in distant metastatic patterns and their corresponding survival between thyroid cancer subtypes. *Head Neck* **44**, 926–932 (2022).
17. Liu, W., Yan, X. & Cheng, R. Continuing controversy regarding individualized surgical decision-making for patients with 1–4 cm low-risk differentiated thyroid carcinoma: a systematic review. *Eur. J. Surg. Oncol.* **46**, 2174–2184 (2020).
18. Chlap, P. et al. A review of medical image data augmentation techniques for deep learning applications. *J. Med. Imaging Radiat. Oncol.* **65**, 545–563 (2021).
19. Garcea, F., Serra, A., Lamberti, F. & Morra, L. Data augmentation for medical imaging: A systematic literature review. *Comput. Biol. Med.* **152**, 106391 (2023).
20. Wang, H. et al. Scientific discovery in the age of artificial intelligence. *Nature* **620**, 47–60 (2023).
21. Gao, C. et al. Synthetic data accelerates the development of generalizable learning-based algorithms for X-ray image analysis. *Nat. Mach. Intell.* **5**, 294–308 (2023).
22. Özbey, M. et al. Unsupervised medical image translation with adversarial diffusion models. *IEEE Trans. Med. Imaging* **42**, 3524–3539 (2023).
23. Xu, Z. et al. Cross-domain attention-guided generative data augmentation for medical image analysis with limited data. *Comput. Biol. Med.* **168**, 107744 (2024).
24. Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* 10684–10695 (2022).
25. Akrout, M. et al. Diffusion-based data augmentation for skin disease classification: Impact across original medical datasets to fully synthetic images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 99–109 (Springer, 2023).
26. Chambon, P. J. M., Bluethgen, C., Langlotz, C. & Chaudhari, A. Adapting pretrained vision-language foundational models to medical imaging domains. In *NeurIPS 2022 Foundation Models for Decision Making Workshop*. (2022).
27. Kather, J. N., Ghaffari Laleh, N., Foersch, S. & Truhn, D. Medical domain knowledge in domain-agnostic generative AI. *NPJ Digit. Med.***5**, 90 (2022).
28. DuMont Schütte, A. et al. Overcoming barriers to data sharing with medical image generation: a comprehensive evaluation. *NPJ Digi. Med,* **4**, 141 (2021).
29. Li, W. et al. The value of sonography in distinguishing follicular thyroid carcinoma from adenoma. *Cancer Manag. Res.* **13**, 3991–4002 (2021).
30. Sobrinho-Simoes, M., Eloy, C., Magalhaes, J., Lobo, C. & Amaro, T. Follicular thyroid carcinoma. *Modern Pathol.* **24**, S10–S18 (2011).
31. Wells, S. A. Jr et al. Revised American Thyroid Association guidelines for the management of medullary thyroid carcinoma: the American Thyroid Association Guidelines Task Force on medullary thyroid carcinoma. *Thyroid* **25**, 567–610 (2015).
32. Zhou, S.-Y. & Luo, L.-X. An overview of the contemporary diagnosis and management approaches for anaplastic thyroid carcinoma. *World J. Clin. Oncol.* **15**, 674 (2024).
33. Bakurov, I., Buzzelli, M., Schettini, R., Castelli, M. & Vanneschi, L. Structural similarity index (SSIM) revisited: A data-driven approach. *Exp. Syst. Appl.* **189**, 116087 (2022).
34. Jayasumana, S. et al. Rethinking fid: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 9307–9315 (2024).
35. Liu, A., Lin, W. & Narwaria, M. Image quality assessment based on gradient similarity. *IEEE Trans. on Image Process.* **21**, 1500–1512 (2011).
36. Naeem, M. F., Oh, S. J., Uh, Y., Choi, Y. & Yoo, J. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning* 7176–7185 (PMLR, 2020).
37. Radford, A. et al. Learning transferable visual models from natural language supervision. in *International conference on machine learning* 8748–8763 (PMLR, 2021).
38. Shorten, C. & Khoshgoftaar, T. M. A survey on image data augmentation for deep learning. *J. Big Data* **6**, 1–48 (2019).
39. Carratino, L., Cissé, M., Jenatton, R. & Vert, J.-P. On mixup regularization. *J. Mach. Learn. Res.* **23**, 1–31 (2022).
40. Zhong, Z., Zheng, L., Kang, G., Li, S. & Yang, Y. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence* **34**, 13001–13008 (2020).
41. Kuo, C.-W., Ma, C.-Y., Huang, J.-B. & Kira, Z. in *Computer Vision–ECCV 2020* (Springer, 2020).
42. Karras, T., Aila, T., Laine, S. & Lehtinen, J. Progressive growing of GANs for improved quality, stability, and variation. In *6th International Conference on Learning Representations (ICLR)* (2018).
43. Peebles, W. & Xie, S. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* 4195–4205 (2023).
44. Saharia, C. et al. Photorealistic text-to-image diffusion models with deep language understanding. *Adv. Neural Inf. Process. Syst.* **35**, 36479–36494 (2022).
45. 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 (2016).
46. Jannin, A. et al. Anaplastic thyroid carcinoma: an update. *Cancers* **14**, 1061 (2022).
47. Pham, H. H., Tran, T. T. & Nguyen, H. Q. VinDr-PCXR: An open, large-scale pediatric chest X-ray dataset for interpretation of common thoracic diseases. *(version 1.0. 0) PhysioNet* https://doi.org/10.13026/k8qc-na36 (2022).
48. Lei, R., Wang, Z. & Qian, L. Ultrasonic characteristics of medullary thyroid carcinoma: differential from papillary thyroid carcinoma and benign thyroid nodule. *Ultrasound Q.* **37**, 329–335 (2021).
49. Yang, Z. et al. Automated diagnosis and management of follicular thyroid nodules based on the devised small-dataset interpretable foreground optimization network deep learning: a multicenter diagnostic study. *Int. J. Surgery* **109**, 2732–2741 (2023).
50. Alexander, E. K. & Cibas, E. S. Diagnosis of thyroid nodules. *Lancet Diabetes Endocrinol.* **10**, 533–539 (2022).
51. Yao, S. et al. Human understandable thyroid ultrasound imaging AI report system—A bridge between AI and clinicians. *Iscience* **26**, https://doi.org/10.1016/j.isci.2023.106530 (2023).
52. Zhou, T. et al. US of thyroid nodules: can AI-assisted diagnostic system compete with fine needle aspiration? *European Radiol.* **34**, 1324–1333 (2024).

53. Alabousi, M. et al. Diagnostic test accuracy of ultrasonography vs computed tomography for papillary thyroid cancer cervical lymph node metastasis: A systematic review and meta-analysis. *JAMA Otolaryngolo. Head Neck Surg.* **148**, 107–118 (2022).
54. Yao, S. et al. Thyroid Cancer Central Lymph Node Metastasis Risk Stratification based on homogeneous positioning deep learning. *Research* **7**, 0432 (2024).
55. Lee, S., Shin, J. H., Han, B.-K. & Ko, E. Y. Medullary thyroid carcinoma: comparison with papillary thyroid carcinoma and application of current sonographic criteria. *Am. J. Roentgenol.* **194**, 1090–1094 (2010).
56. Wang, Y. et al. Clinical evaluation of malignancy diagnosis of rare thyroid carcinomas by an artificial intelligent automatic diagnosis system. *Endocrine* **80**, 93–99 (2023).
57. Zhu, J.-Y., He, H.-L., Jiang, X.-C., Bao, H.-W. & Chen, F. Multimodal ultrasound features of breast cancers: correlation with molecular subtypes. *BMC Med. Imaging* **23**, 57 (2023).
58. Rädsch, T. et al. Labelling instructions matter in biomedical image analysis. *Nat. Mach. Intell.* **5**, 273–283 (2023).
59. Komatsubara, K. M. & Carvajal, R. D. The promise and challenges of rare cancer research. *Lancet Oncol.* **17**, 136–138 (2016).
60. Huang, E. P. et al. Criteria for the translation of radiomics into clinically useful tests. *Nat. Rev. Clin. Oncol.* **20**, 69–82 (2023).
61. Hedlin, E. et al. Unsupervised semantic correspondence using stable diffusion. In *Advances in Neural Information Processing Systems.* (2024).
62. Tessler, F. N. et al. ACR thyroid imaging, reporting and data system (TI-RADS): white paper of the ACR TI-RADS committee. *J. Am. Coll. Radiol.* **14**, 587–595 (2017).
63. Zhang, L., Rao, A. & Agrawala, M. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 3836–3847 (2023).
64. Ho, J. & Salimans, T. Classifier-Free Diffusion Guidance. Preprint at https://doi.org/10.48550/arXiv.2207.12598 (2021).
65. Song, J., Meng, C. & Ermon, S. Denoising Diffusion Implicit Models. *In Proc. International Conference on Learning Representations* (2021).
66. Jocher, G., Chaurasia. A. & Qiu. J. "YOLO by Ultralytics". https://github.com/ultralytics/ (2023).
67. Sobel, I. & Feldman, G. A 3 x 3 Isotropic Gradient Operator for Image Processing. *Pattern Classification and Scene Analysis* 271–272 (1973).
68. Chowdhury, G. G. *Introduction to Modern Information Retrieval*, (Facet Publishing, 2010).
69. Dai, Y., Gieseke, F., Oehmcke, S., Wu, Y. & Barnard, K. Attentional feature fusion. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* 3560–3569 (2021).
70. Dai, F. fangdai-dear/Tiger-Model: Tiger-Model v1.0.0. Zenodo. https://doi.org/10.5281/zenodo.15175057 (2025).
71. Cohen, J. P. et al. TorchXRayVision: A library of chest X-ray datasets and models. In *International Conference on Medical Imaging with Deep Learning* 231–249 (PMLR, 2022).

Acknowledgements

This work was supported in part by funding from the National Natural Science Foundation of China (Grant No. 62406191, S.Y.; Grant No. 82302231, Y.Z.), SJTU Star Program (YG2024QNA02, S.Y.), the SFYF at SJTU (24×010500175, S.Y.), the Fundamental Research Funds for the Central Universities (YG2025LC03, S.Y.), Shanghai Municipal Education Commission (No. 2024AIYB010, S.Y.), Science and Technology Commission of Shanghai Municipality (STCSM) (Grant No. 23JS1400700, H.L.), Science and Technology Commission of Shanghai Municipality (STCSM) (Grant No. 23JS1400800, H.L.), Neil Shen's SJTU Medical Research Fund (H.L.) and SJTU Transmed Awards Research (STAR) Grant No. 20210106 (H.L.). The content of this article does not reflect the view of the funding sources.

Author contributions

F.D., S.Y., M.W., Y.Z., X.Q., P.S., C.Q., J.Y., G.S., J.S., M.F.W., Y.W., Z.Y., F.S., J.S., X.W., W.C., X.Z., H.L., S.Y., and F.D. developed the concept for the manuscript. S.Y., F.D., X.Z., and P.S. contributed to the drafting of the manuscript. F.D. P.S., J.Y., and X.Q. designed the model and analysis the data. M.W. organized the doctors to complete the evaluation test. Y.Z., C.Q., G.S., J.S., M.F.W., Y.W., and Z.Y., F.S., J.S., X.W., participated in the collection and annotation of data. W.C. contributed to provide medical data and clinical advice. F.D. and S.Y. contributed equally to this work. All authors critically revised 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/s41467-025-59478-8.

Correspondence and requests for materials should be addressed to Siqiong Yao, Fenyong Sun, Wei Cai, Xingcai Zhang or Hui Lu.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.