Contents

Supplementary Section 1. Pretraining dataset

We collected the pretraining data from two hospitals in China: Peking University Cancer Hospital & Institute (PKUCH) and Nanchang People's Hospital (NPH). PKUCH, located in Beijing (northern part of China), is a leading specialized cancer hospital renowned for its focus on cancer treatment, research, and education. NPH, located in Jiangxi Province (southern part of China), is a prominent comprehensive medical hospital. Notably, PKUCH not only treats local patients but also attracts patients from across the country which further enhances the diversity of the pretraining dataset. Scanning videos of breast examinations were collected from standard clinical workflows as videos provide more comprehensive information for pretraining BUSGen. Besides, we collected the corresponding ultrasound reports documented by radiologists and pathology results for patients who underwent biopsies or surgeries. We removed low-quality data where clinical information is incomplete or lesions could not be clearly visualized. We refer to this large-scale, high-quality pretraining dataset as "BUS-3.5M", as it contains 3,518,495 breast ultrasound images in total.

Supplementary Table 1: **Patient demographics of BUS-3.5M.**

Characteristics	**Count**
Patients	4,636
Normal patients	1,589 (34.3%)
Abnormal patients	3,047 (65.7%)
Age (mean)	47.19
Age (mean, benign)	42.50
Age (mean, malignant)	51.94
Age	
<40 years old	898 (19.4%)
40-49 years old	813 (17.5%)
50-59 years old	816 (17.6%)
60-69 years old	274 (5.9%)
≥70 years old	100 (2.2%)
Unknown	1,735 (37.4%)

Supplementary Table 2: **Lesion characteristics of BUS-3.5M.**

Characteristics	**Count**
Lesions	3,749
Biopsy-confirmed lesions	
All	1,387
Benign	694 (50.0%)
Malignant	693 (50.0%)
Lesion BI-RADS	
2	770 (20.5%)
3	1,462 (39.0%)
4A	602 (16.0%)
4B	329 (8.8%)
4C	486 (13.0%)
5	100 (2.7%)

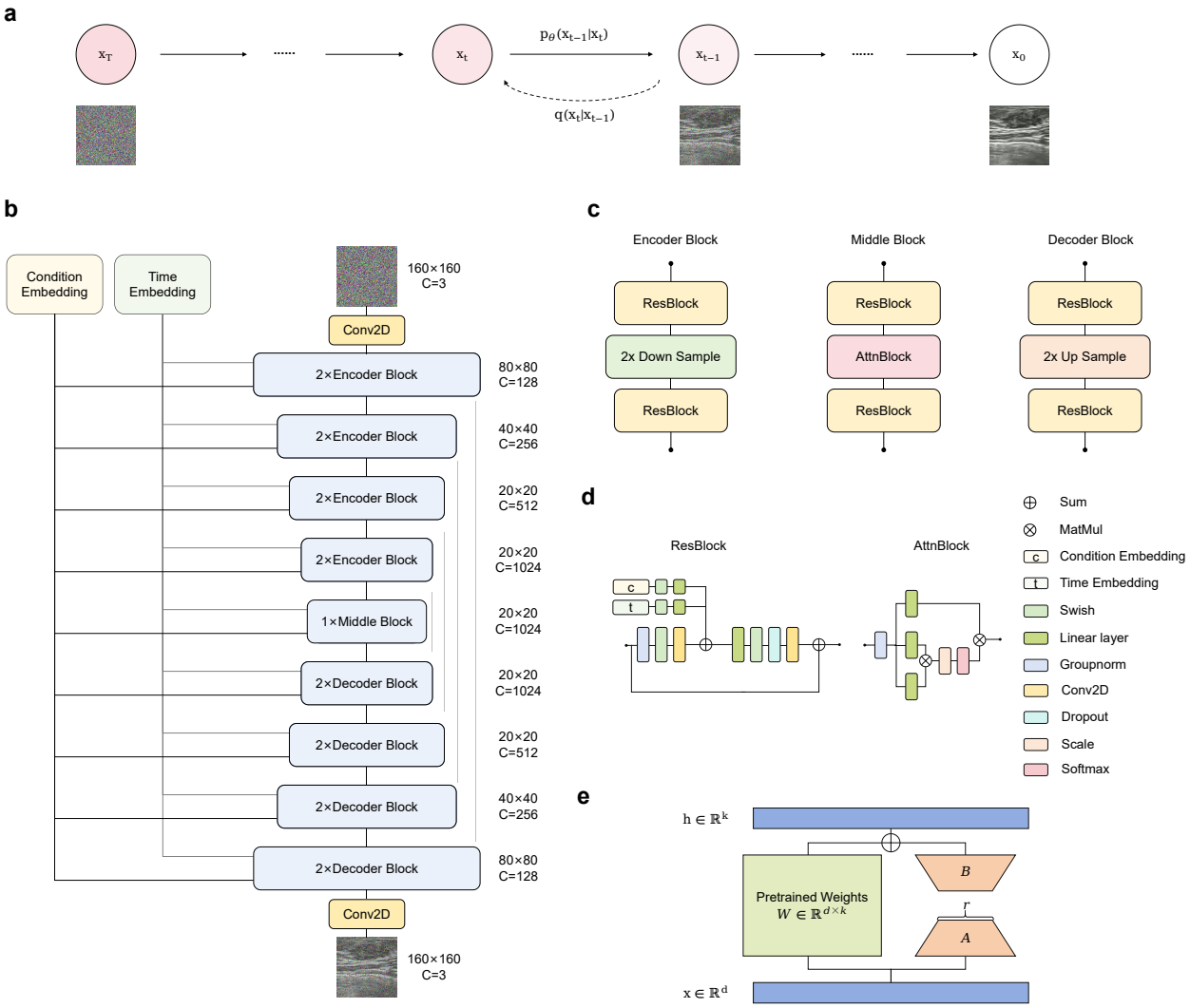
Supplementary Table 3: **Pathological subtypes of lesions in the BUS-3.5M.**

Characteristics	**Count**
Benign	694
Fibroadenoma (FA)	243 (17.5%)
Adenosis (AD)	102 (7.4%)
Mastitis (MST)	67 (4.8%)
Intraductal papilloma (IDP)	43 (3.1%)
Benign epithelial proliferation (BEP)	22 (1.6%)
Sclerosing adenosis (SAD)	16 (1.2%)
Atypical ductal hyperplasia (ADH)	13 (0.9%)
Mammary duct ectasia (MDE)	12 (0.9%)
Cyst (CYST)	9 (0.6%)
Benign phyllodes tumour (BPT)	7 (0.5%)
Radial scar (RS)	7 (0.5%)
Pseudoangiomatous stromal hyperplasia (PASH)	2 (0.1%)
Tubular adenoma (TA)	2 (0.1%)
Fat necrosis (FN)	1 (0.1%)
Granular cell tumour (GCT)	1 (0.1%)
Unspecified	147 (10.6%)
Malignant	693
Invasive breast carcinoma (IBC)	567 (40.9%)
Ductal carcinoma in situ (DCIS)	34 (2.5%)
Invasive micropapillary carcinoma (IMPC)	10 (0.7%)
Microinvasive carcinoma (MIC)	9 (0.6%)
Mucinous carcinoma (MC)	8 (0.6%)
Invasive lobular carcinoma (ILC)	7 (0.5%)
Carcinoma with apocrine differentiation (CAD)	6 (0.4%)
Solid papillary carcinoma in situ (SPCIS)	3 (0.3%)
Metaplastic carcinoma (MPC)	2 (0.1%)
Invasive solid papillary carcinoma (ISPC)	2 (0.1%)
Lobular carcinoma in situ (LCIS)	2 (0.1%)
Tubular carcinoma (TC)	1 (0.1%)
Lymphoma (LYM)	1 (0.1%)
Adenoid cystic carcinoma (ACC)	1 (0.1%)
Malignant phyllodes tumour (MPT)	1 (0.1%)
Paget's disease (PD)	1 (0.1%)
Unspecified	38 (2.7%)

Supplementary Table 4: **Ultrasound device types of the BUS-3.5M.**

Device type	**Manufacturer**	**Count**
Siemens-ACUSON-Oxana	Siemens	2,027 (34.3%)
GE-LOGIQ-E9	General Electric	1,165 (19.7%)
Esaote-MyLab90	Esaote	556 (9.4%)
Esaote-MyLabClassC	Esaote	551 (9.3%)
Philips-EPIQ7	Philips	383 (6.5%)
SonoScape-S60	SonoScape	296 (5.0%)
GE-EDVoluson-E8-Expert	General Electric	251 (4.2%)
Samsung-RS80A	Samsung	170 (2.9%)
SonoScape-Clinic	SonoScape	160 (2.7%)
TOSHIBA-Aplio-i700	TOSHIBA	128 (2.2%)
Mindray-M9	Mindray	71 (1.2%)
Canon-Aplio-i800	Canon	37 (0.6%)
Siemens-ACUSON-NX3-Elite	Siemens	36 (0.6%)
VINNO-G86	VINNO	26 (0.4%)
Mindray-Resona-R9	Mindray	18 (0.3%)
Mindray-NuewaI9	Mindray	11 (0.2%)
Samsung-Heraw10	Samsung	11 (0.2%)
Mindray-Resona7	Mindray	9 (0.2%)
Unspecified	-	1 (0.1%)

Supplementary Section 2. BUSGen pretraining



Supplementary Figure 1: **Overview of BUSGen pretraining and fine-tuning. a,** The denoising and noising process of diffusion model, where x_T represents the noisy image and x_0 is the original image. The model learns to denoise x_t to obtain x_{t-1} iteratively. **b,** The architecture of BUSGen, demonstrates the U-Net structure with incorporated condition embeddings and time embeddings at each layer. The structure includes multiple encoder blocks, a middle block, and decoder blocks, where each encoder block consists of ResBlocks and downsampling, the middle block contains ResBlocks and an AttnBlock and each decoder block comprises ResBlocks and upsampling. **c,** Encoder, Middle, and Decoder Blocks. **d,** ResBlock and AttnBlock. **e,** Illustration of the LoRA fine-tuning principle, where pretrained weights W are factorized into low-rank matrices A and B, enabling efficient adaptation of the model.

Supplementary Table 5: **Hyperparameters of BUSGen (diffusion model).**

Hyperparameter	**Value**
T (time steps)	500
Noise schedule (β_1, β_T)	(1.0×10^{-4}, 0.028)
Image	160× 160
Channel	128
Channel multiplier	[1, 2, 2, 2]
Dropout	0.15

Supplementary Table 6: **Hyperparameters of pretraining.**

Hyperparameter	Value
Epoch	70
Batch size	64
Optimizer	AdamW
Learning rate	6.25×10^{-6}
Weight decay	1.0×10^{-4}
Learning rate scheduler	Cosine Annealing
Gradient clipping	1.0

Supplementary Table 7: **Ablation study of pretraining settings.** We compared our pretraining settings with a web-data pretrained latent diffusion model: Stable Diffusion (SD-1.5) [1, 2]. For each setting, we sampled 100,000 images to train the diagnosis task.

Methods	Internal test	External test
Stable Diffusion	0.940 (0.924, 0.960)	0.929 (0.899, 0.961)
BUSGen	0.949 (0.933, 0.964)	0.944 (0.924, 0.970)

Supplementary Section 3. BUSGen adaptation

To generate images of downstream tasks, we fine-tuned the pretrained BUSGen on a small amount of downstream data. As shown in Supplementary Figure 1e, we froze pretrained parameters and employed the low-rank adapters (LoRA) [3] as tunable lightweight parameters to prevent overfitting during fine-tuning. The assumption of LoRA is that the change in weights during model adaptation has a low "intrinsic rank". LoRA injects trainable rank decomposition matrices into each layer during fine-tuning. Instead of updating the entire weight matrix W, LoRA updates two smaller matrices A and B such that $\Delta W = B \cdot A$. Compared to traditional fine-tuning, which updates the matrix W, LoRA fine-tuning learns only the low-rank update ΔW (i.e., A and B). During fine-tuning, we keep W fixed and the model learns the change in parameters, meaning $h = Wx + B(Ax)$, where x is the input vector and h is the output vector. After fine-tuning, the original W is updated to $W + \Delta W$.

Supplementary Table 8: **Hyperparameters of fine-tuning.**

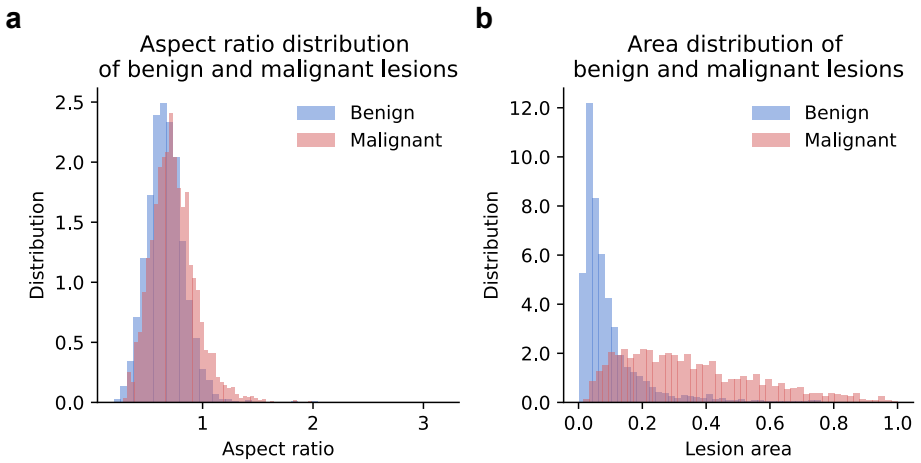
Hyperparameter	Value
LoRA rank	120
Epoch	70
Batch size	64
Optimizer	AdamW
Learning rate	6.25×10^{-6}
Weight decay	1.0×10^{-4}
Learning rate scheduler	Cosine Annealing
Gradient clipping	1.0

By selecting different conditions, we can control BUSGen to generate well-balanced and diverse data to enhance the generalizability of BUS-DMs for specific tasks. (1) We introduce the selection strategy for conditions, including pathology labels, device types, and lesion-bounding boxes. To ensure a balanced representation of the binary pathology conditions, we sample an equal number of benign and malignant images. (2) We uniformly sample across device types (Supplementary Figure 2) in the training set to ensure a well-balanced distribution of device types in the generated data. Different device types can lead to variations in image quality, color bias, and texture. Therefore, a well-balanced distribution of device types can improve the generalization of BUS-DM models across various ultrasound devices. (3) We densely sample bounding boxes of lesions from the prior distribution to enhance lesion area diversity. As shown in Supplementary Figure 3, the distribution of bounding boxes differs between benign and malignant lesions: the areas of malignant lesions are larger than those of benign ones. Therefore, we sampled bounding box for benign and malignant lesions separately from the prior distribution estimated from the training set.

Supplementary Table 9: **Hyperparameters of DPMSolver++.**

Hyperparameter	Value
Method	singlestep
T (time steps)	50
Order	3

Supplementary Figure 2: **Device types.** We show several device types in the training set. Different device types could lead to image variance in quality, color bias, and textures.



Supplementary Figure 3: **Distribution lesion bounding boxes. a,** Distribution of lesion aspect ratios of benign and malignant lesions. **b,** Distribution of lesion areas of benign and malignant lesions.

Supplementary Table 10: **Effectiveness of data cleaning.** For fair comparison, we sampled 1 million raw images without data cleaning, and trained the downstream model on this dataset for the diagnosis task.

Data clean	Internal test set	External test set
No	0.940 (95% CI 0.922−0.959)	0.936 (95% CI 0.906−0.969)
Yes	0.953 (95% CI 0.935−0.967)	0.951 (95% CI 0.921−0.975)

To ensure the quality of the generated data, we perform a data cleaning process to remove low-quality images whose ultrasound features are inconsistent with the pseudo labels (pathology condition used to generate the image). Incorrect pathology labels in the generated data can be particularly detrimental to the training of BUS-DMs. To automatically identify images that are likely to have incorrect pathology labels, we train filters to predict the pathology label for each image. Thereby, we remove images whose pseudo labels are inconsistent with the predicted labels. The training set of filters is noise-free, which consists of the collected real data with biopsy-confirmed pathology labels. After data cleaning, approximately 10% of the generated data are removed. As shown in Supplementary Table 10, this cleaning process leads to a notable improvement in the performance of the BUS-DMs, as it ensures that the generated training data are of high quality.

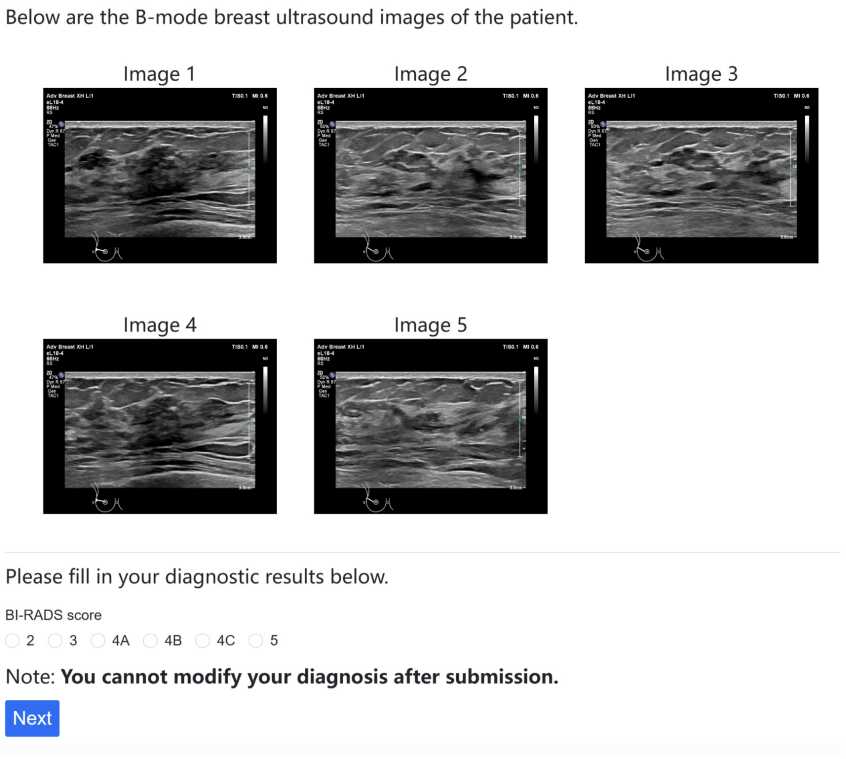
Supplementary Section 4. Baseline-CLIP implementation

To evaluate the adaptivity of BUSGen, we compared BUS-DMs with classification-based foundational models (CLIP [4]) trained on real data, referred to as "Baseline-CLIP", which were pretrained on the PMC-OA dataset [5] containing 1.65 million image-caption pairs. These 1.65 million data samples were extracted from papers from PubMedCentral's OpenAccess [6] and tackled the subfigure-subcaption correspondence. We transferred them to the breast ultrasound domain by training on image-report pairs of the BUS-3.5M dataset. The ultrasound report contained information written by radiologists to describe the main features of lesions and breasts. Also, we incorporated pathology reports (if existed) to introduce the pathological information. During the training process, we froze the text encoder of the PMC-OA data pretrained CLIP model to preserve rich language-based knowledge embeddings. Finally, we adapted the Baseline-CLIP model to downstream tasks through (few-shot) fine-tuning, where we also froze the text encoder of the CLIP model.

Supplementary Section 5. Reader study

Supplementary Table 11: **Reader Experience in the Study.** Summary of the participating radiologists' expertise, including their estimated annual volume of breast ultrasound interpretations and years of professional experience. All participants specialized in breast imaging.

Reader ID	Estimated reads per year	Years of experience
R1	2,000	26
R2	1,850	19
R3	2,150	14
R4	2,700	11
R5	1,500	9
R6	800	7
R7	1,200	4
R8	790	4
R9	780	3



Supplementary Figure 4: **Reader study interface.** Radiologists were required to assign BI-RADS predictions of given breast ultrasound images.

A reader study was conducted on two diagnosis tasks of benign-DCIS and benign-malignant classification. In Supplementary Table 11, we listed the experience of readers (n=9). To save readers' time, we directly incorporated specially collected DCIS data into the benign-malignant classification dataset to form the reader study dataset. The benign-malignant classification data are prospectively collected from consecutive patients who underwent biopsies. As biopsies were conducted for patients with suspiciously malignant lesions, the benign lesions were much less than the malignant ones in this external test set. Since the distribution of this data differs from the distribution in real-world clinical settings, we specifically remind doctors to evaluate each case individually without considering other cases.

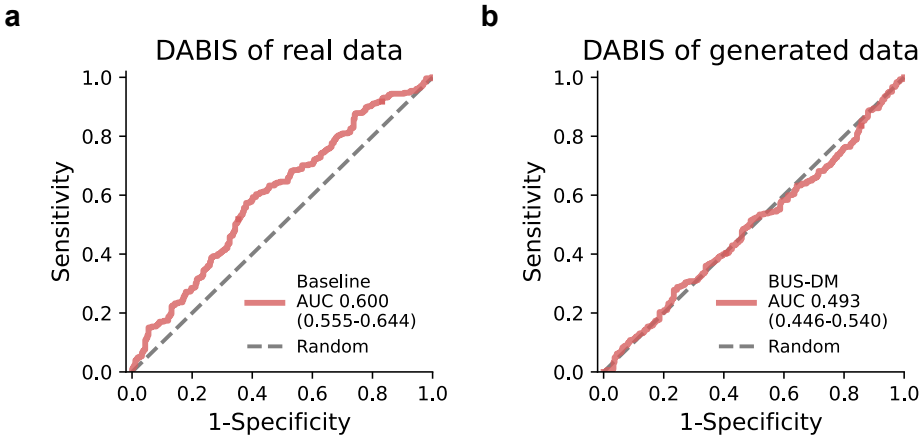
A web interface was employed to facilitate interaction during the reader study (Supplementary Figure 4). (1) Each reader independently analyzed breast ultrasound lesions and submitted their evaluations. To maintain study integrity and prevent information leakage, the participants had no prior access to the data used in the study. Readers were explicitly instructed to evaluate each lesion independently and were informed that predictions could not be revised after submission. (2) To assess the effectiveness of AI assistance, readers were then given the results from the BUS-DM model. Readers were required to analyze these results and update their predictions accordingly. We provided information on the sensitivity and specificity of BUS-DM at various thresholds based on the internal test set, helping readers understand the model's predicted probability scores.

Supplementary Table 12: **Detailed results of reader study.** We provide the changes of BI-RADS scores with the BUS-DM assistance on the benign and malignant lesions on the benign-malignant and benign-DCIS test set. We also provide the number of lesions whose BI-RADS changes between BI-RADS 3 and 4A+, indicating the change of biopsy recommendation.

		Benign (n=63)		**Malignant** (n=164)		**DCIS** (n=133)	
	Adjustment	**BI-RADS**	**Biopsy**	**BI-RADS**	**Biopsy**	**BI-RADS**	**Biopsy**
R1	Upgrade	1	0	7	0	16	1
	Downgrade	5	3	0	0	2	1
R2	Upgrade	1	1	17	1	13	7
	Downgrade	8	7	10	0	8	6
R3	Upgrade	1	0	6	0	18	2
	Downgrade	1	0	3	1	7	5
R4	Upgrade	2	1	10	1	14	3
	Downgrade	19	13	15	4	16	9
R5	Upgrade	7	0	28	0	50	4
	Downgrade	3	2	1	0	2	1
R6	Upgrade	0	0	3	0	8	3
	Downgrade	4	2	2	1	3	3
R7	Upgrade	4	0	46	0	49	1
	Downgrade	5	1	1	0	0	0
R8	Upgrade	2	0	11	2	24	3
	Downgrade	4	3	2	0	6	3
R9	Upgrade	2	0	40	4	44	6
	Downgrade	12	7	13	1	10	4

In Supplementary Table 12, we provide details on how BUS-DM assisted radiologists in their diagnoses. (1) We listed the changes in BI-RADS scores (including "Downgrade", and "Upgrade") made by each reader with the help of BUS-DM on the external test set. For benign lesions, the number of downgraded lesions was equal to or greater than the number of upgraded ones, except for reader R5. For malignant lesions (excluding the specific subset of 133 DCIS lesions), the number of upgraded lesions was greater than the number of downgraded ones, except for reader R4. For DCIS lesions, the number of upgraded lesions consistently exceeded the number of downgraded ones. (2) Additionally, we provide data on the number of lesions changing between BI-RADS 3 and 4A+, indicating the change in biopsy recommendation. For benign lesions, the number of lesions downgraded from BI-RADS 4A+ to 3 was equal to or greater than the number of those upgraded from BI-RADS 3 to 4A+. For malignant lesions, the number of lesions upgraded from BI-RADS 3 to 4A+ was equal to or greater than the number of those downgraded from BI-RADS 4A+ to 3, except for readers R3, R4, and R6. For DCIS lesions, the number of lesions upgraded from BI-RADS 3 to 4A+ was consistently greater than or equal to the number of those downgraded from BI-RADS 4A+ to 3, except for readers R3 and R4. These results indicate that readers enhanced their BI-RADS predictions for both benign and malignant lesions, and the improvements were fairly consistent.

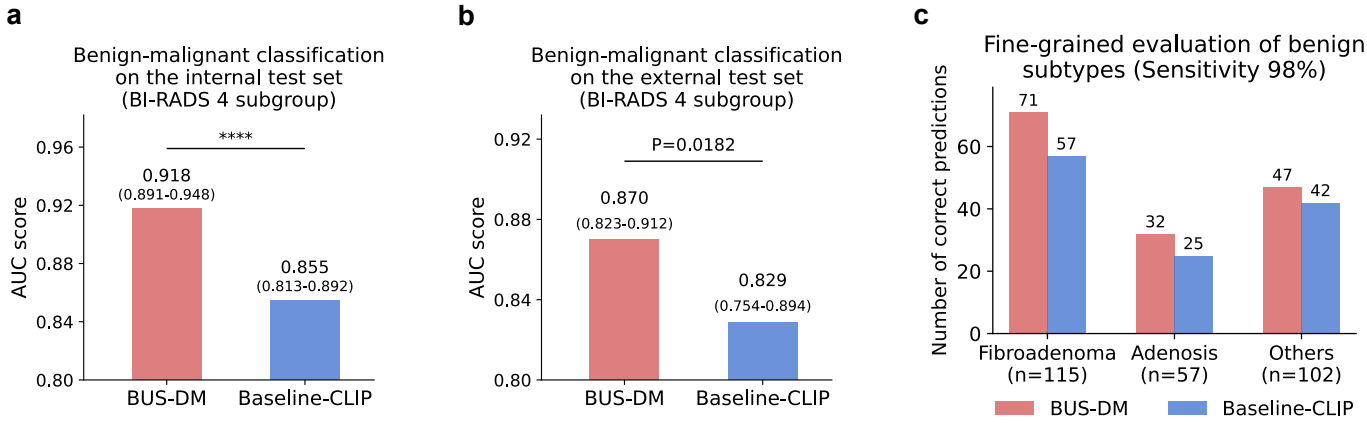
Supplementary Section 6. BUSGen enhances generalization ability



Supplementary Figure 5: **DABIS score of real data and generated data.**

Shortcut learning is a phenomenon where deep learning models learn to solve a task based on spurious correlations in the data, but not the causal features of the task. Following the approach proposed by the previous work [7], we quantify the degree of data acquisition-induced shortcut learning (DABIS), with larger AUC_{DABIS} score indicating higher degree of shortcut learning and worse generalization ability. We demonstrate that real collected data (AUC_{DABIS}: 0.600; 95% CI 0.555-0.644) induced more spurious correlations than BUSGen generated data (AUC_{DABIS}: 0.493; 95% CI 0.446-0.540). As an AUC of 0.50 indicates random predictions without learning any shortcuts, these results show that BUSGen could effectively prevent DAB-induced shortcut learning. Supplementary Figure 5 shows the DABIS score of real data and generated data.

Supplementary Section 7. Subgroup analysis of the diagnosis task



Supplementary Figure 6: **Subgroup analysis of the diagnosis task. a**, Comparison of BUS-DM (red) with Baseline-CLIP (blue) for benign-malignant classification of internal BI-RADS 4 lesions. BUS-DM achieved a higher AUC of 0.918 (95% CI: 0.891–0.948) compared to Baseline-CLIP with an AUC of 0.855 (95% CI: 0.813–0.892; P-value<0.0001). **b**, Comparison of BUS-DM (red) with Baseline-CLIP (blue) for benign-malignant classification of external BI-RADS 4 lesions. BUS-DM achieved a higher AUC of 0.870 (95% CI: 0.823–0.912) compared to Baseline-CLIP with an AUC of 0.829 (95% CI: 0.754–0.894; P-value=0.0182). **c**, The number of correct predictions for each benign subtype at a sensitivity of 98%. BUS-DM consistently outperformed Baseline-CLIP across different subtypes. ****P-value<0.0001.

Supplementary Section 8. Clustering results of prognosis tasks

Supplementary Table 13: **Evaluation of clustering of TNBC and ALN classification embeddings.**

TNBC	NMI	ARI	SIL
Baseline-CLIP	0.035	0.098	0.137
BUSGen-DM	0.117	0.274	0.160

ALN	NMI	ARI	SIL
Baseline-CLIP	0.003	0.009	0.030
BUSGen-DM	0.382	0.514	0.286

For evaluating the clustering accuracy of predicting TNBC and ALN status, we employed metrics including normalized mutual information (NMI), adjusted Rand index (ARI) and silhouette coefficient (SIL). For the calculation of the NMI and ARI, we employ K-means as the clustering method. Below are the definitions and computation methods for these metrics.

Normalized Mutual Information Assume two label assignments of the same N objects, U and V. Their entropy is the amount of uncertainty for a partition set is defined by $H(U) = -\sum_{i=1}^{|U|} P(i)\log(P(i))$ and $H(V) = -\sum_{i=1}^{|V|} P'(i)\log(P'(i))$ where $P(i) = |U_i|/N$ is the probability that an object picked at random from U falls into class U_i. Likewise for $P'(i) = |V_i|/N$. The mutual information $\mathrm{MI}(U,V)$ is defined as $\mathrm{MI}(U,V) = \sum_{i=1}^{|U|}\sum_{j=1}^{|V|} P(i,j)\log\left(\frac{P(i,j)}{P(i)P'(j)}\right)$. The normalized mutual information is defined as:

$$\mathrm{NMI}(U,V) = \frac{\mathrm{MI}(U,V)}{\mathrm{mean}\left(H(U), H(V)\right)}. \tag{1}$$

Adjusted Rand Index Let C be a ground truth label assignment and K be the clustering, and a is defined as the number of pairs of elements that are in the same sets in C and in the same set in K, b is the number of pairs of elements that are in different sets in C and in different sets in K. The Rand index is given by $\mathrm{RI} = \frac{a+b}{C_2^{n_{samples}}}$. And the ARI is given by:

$$\mathrm{ARI} = \frac{\mathrm{RI} - E[\mathrm{RI}]}{\max(\mathrm{RI}) - \mathrm{E[RI]}}. \tag{2}$$

where E[RI] is the expected Rand Index (RI) of random labelings.

Silhouette Coefficient score Let a be the mean distance between a sample and all other points in the same class, and b be the mean distance between a sample and all other points in the next nearest cluster. The Silhouette Coefficient score for a set of samples is given as the mean of each sample's Silhouette Coefficient where the Silhouette Coefficient s for a single sample is then given as

$$s = \frac{b-a}{\max(a,b)}. \tag{3}$$

As illustrated in Supplementary Table 13, the BUS-DM had higher NMI, ARI and SIL scores, showing its higher generalization ability than Baseline-CLIP.

Supplementary References

[1] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[2] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[3] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

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

[5] Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2023.

[6] Bethesda (MD): National Library of Medicine. Pmc open access subset, 2003.

[7] Cathy Ong Ly, Balagopal Unnikrishnan, Tony Tadic, Tirth Patel, Joe Duhamel, Sonja Kandel, Yasbanoo Moayedi, Michael Brudno, Andrew Hope, Heather Ross, et al. Shortcut learning in medical ai hinders generalization: method for estimating ai model generalization without external data. *NPJ Digital Medicine*, 7(1):124, 2024.