

# Supplementary Material for Design as Desired: Utilizing Visual Question Answering for Multimodal Pre-training

Tongkun Su<sup>1,2,\*</sup>, Jun Li<sup>3,\*</sup>, Xi Zhang<sup>1,2</sup>, Haibo Jin<sup>5</sup>, Hao Chen<sup>5</sup>, Qiong Wang<sup>1</sup>,  
Faqin Lv<sup>4</sup>, Baoliang Zhao<sup>1,\*\*(✉)</sup>, and Ying Hu<sup>1,\*\*(✉)</sup>

<sup>1</sup> Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences  
{tk.su, bl.zhao, ying.hu}@siat.ac.cn

<sup>2</sup> University of Chinese Academy of Sciences

<sup>3</sup> Technical University of Munich, Munich Center for Machine Learning

<sup>4</sup> Southern Medical University

<sup>5</sup> The Hong Kong University of Science and Technology

<p>The breast soft tissues demonstrate clear visualization of all layers. The skin and subcutaneous fat layers show no obvious abnormal echogenicity. Bilateral glandular structures appear disordered with heterogeneous internal echogenicity. A hypoechoic nodule is observed in the left breast at the [size] o'clock position from the nipple, measuring approximately [size] in size, exhibiting clear boundaries and regular morphology. CDFI demonstrates no detectable blood flow signals. In the right breast, multiple hypoechoic nodules are observed, with the largest one located at the [size] o'clock position from the nipple, measuring approximately [size] in size, displaying clear boundaries and regular morphology. CDFI demonstrates no detectable blood flow signals within these nodules. Bilateral axillary scanning reveals no enlarged lymph nodes.</p>	<p><b>Coarse</b></p>	<p>Q: Generate ultrasound report.</p>		
		<p>A: (The original report showing on the left side.)</p>		
	<p><b>Medium</b></p>		<p>Q: Left breast:</p>	
			<p>A: In the lower pole of the left lobe, a hypoechoic nodule is visible, measuring approximately [size], displaying clear boundaries and a regular shape. CDFI shows detectable blood flow signals within the nodule.</p>	
			<p>Q: Right breast:</p>	
			<p>A: In the right breast, multiple hypoechoic nodules are observed, with the largest one located at the [size] o'clock position from the nipple, measuring approximately [size] in size, displaying clear boundaries and regular morphology. CDFI demonstrates no detectable blood flow signals within these nodules.</p>	
	<p><b>Fine</b></p>		<p>Q: Bilateral axillary:</p>	
			<p>A: No obvious enlargement of lymph nodes detected.</p>	
			<p>Q: Are there any nodules in the images?</p>	<p>Q: Where is the nodule located in the images?</p>
			<p>A: Yes.</p>	<p>A: Bilateral breast.</p>
		<p>Q: Are there dilated mammary ducts in the images?</p>	<p>Q: Where are the dilated mammary ducts located?</p>	
		<p>A: No.</p>	<p>A: [Empty]</p>	

**Fig. 1.** Example of the breast VQA design. [Empty] indicates that this question will not be treated as pre-training data.

\* These authors contributed equally to this work.

\*\* Corresponding authors.

**Table 1.** Configuration of pre-training.

Configuration	Multimodal pretraining	Report generation
Optimizer	AdamW	AdamW
Learning rate	2e-5	2e-5
Weight decay	0.05	0.05
Learning rate scheduler	Linear warmup+ cosine annealing	Linear warmup + cosine annealing
Initial learning rate	1e-8	1e-8
Warmup periods	40% of training time	40% of training time
Early stop	5	5
Batch size(B)	25	25
Buffer size(N)	100	100
Query(m)	32	32
Epoch	30 for ViT and 50 for ResNet	50

**Table 2.** Finetuning configuration of different downstream vision tasks.

Configuration	Classification	Detection	Segmentation
Optimizer	AdamW	AdamW	AdamW
Early stop	10	10	10
Epoch	50	50	50
Method	Linear probe	YOLOv3	SETR/UNet
Learning rate	5e-4	5e-4	2e-4
Weight decay	1e-6	1e-6	0.05
Batch size	48	16	8

**Table 3.** Quantitative result of multimodal pre-training. The results have been presented in the main text. “-” indicates that it is not suitable for this situation.

Method	Classification(AUC%)		Detection(AP%)		Segmentation(DICE%)	
	BUSI	AUITD	BUSI	DDTI	BUSI	DDTI
Random(ViT)	56.4	81.3	-	-	38.1	58.1
Random(Res)	61.5	81.3	51.5	13.9	58.0	64.7
ImageNet(ViT)	84.5	82.5	-	-	63.9	61.5
ImageNet(Res)	82.9	82.2	<b>66.7</b>	50.0	49.0	61.1
GloRIA(Res)	85.5	80.2	54.9	21.1	63.7	63.8
MGCA(ViT)	82.9	80.2	-	-	61.2	68.7
MGCA(Res)	82.9	82.2	55.5	10.5	59.2	68.8
MRM(ViT)	69.2	81.3	-	-	61.1	<b>73.1</b>
<b>Ours(ViT)</b>	<b>88.9</b>	<b>83.3</b>	-	-	63.5	70.2
<b>Ours(Res)</b>	84.6	82.2	62.1	<b>57.9</b>	<b>65.6</b>	70.4