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

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

[1] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
{tk.su, bl.zhao, ying.hu}@siat.ac.cn
[2] University of Chinese Academy of Sciences
[3] Technical University of Munich, Munich Center for Machine Learning
[4] Southern Medical University
[5] The Hong Kong University of Science and Technology

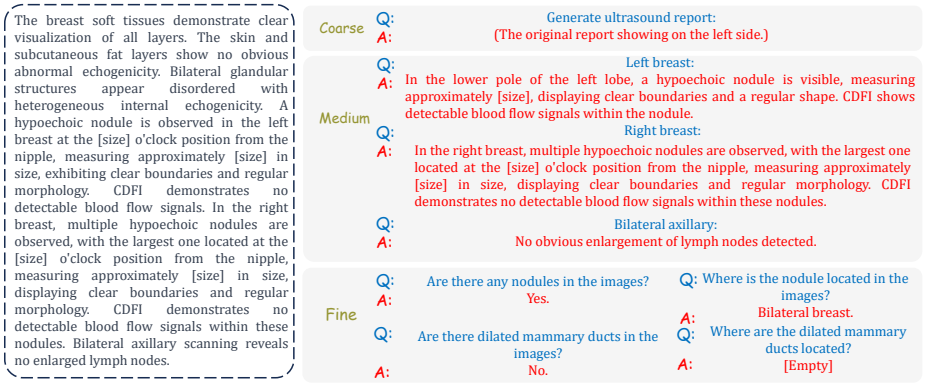

Fig. 1. Example of the breast VQA design. [Empty] indicates that this question will not be treated as pre-training data.

* 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	**66.7**	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	**73.1**
Ours(ViT)	**88.9**	**83.3**	-	-	63.5	70.2
Ours(Res)	84.6	82.2	62.1	**57.9**	**65.6**	70.4