S1 Supplementary Material

Table S1. Data Efficiency Evaluation on CheXpert [13]. We evaluate the full-finetuning performance of the proposed method with multiple baselines on the CheXpert-5×200 [13] datasets with different ratios of training data (1%/10%/100%). A more robust pre-trained model should be able to generalize easily to the target task even with a small amount of training data. We highlight the top result in bold and the second-best with an underline.

Method	**CheXpert 5 × 200** [13]					
	1%		10%		100%	
	FT-Acc	FT-AUC	FT-Acc1	FT-AUC	FT-Acc	FT-AUC
Random-ViT [20]	21.42	58.81	20.02	54.96	20.32	63.68
ImageNet-ViT [20]	35.54	67.68	45.15	75.52	56.46	85.71
CLIP-ViT-BERT [21]	23.82	62.78	40.74	71.20	44.84	77.59
GLoRIA-R50 [12]	48.45	77.08	53.05	83.34	57.56	87.11
MGCA-ViT [26]	45.55	75.85	54.65	82.54	56.96	86.33
MRM-ViT [29]	49.35	78.84	<u>55.86</u>	<u>86.06</u>	56.56	87.41
MedCLIP-Swin [27]	<u>50.65</u>	**80.60**	52.95	82.91	57.46	<u>87.85</u>
Ours-Prefix [15]	45.35	77.17	55.46	83.67	<u>61.16</u>	87.73
Ours-IA3 [16]	46.65	75.69	53.65	82.23	61.06	86.81
Ours-LoRA [11]	**51.15**	<u>80.10</u>	**56.96**	**86.28**	**63.96**	**88.22**

Table S2. Data Efficiency Evaluation on RSNA [24]. We evaluate the full-finetuning performance of the proposed method with multiple baselines on the out-of-domain RSNA [24] datasets with different ratios of training data (1%/10%/100%). A more robust pre-trained model should be able to generalize easily to the target task even with a small amount of training data. Our accuracy drops by <3% when using 1% training data compared to 100% training data. We highlight the top result in bold and the second-best with an underline.

Method	**RSNA** [24]					
	1%		10%		100%	
	FT-Acc	FT-AUC	FT-Acc	FT-AUC	FT-Acc	FT-AUC
Random-ViT [20]	62.15	66.23	71.76	78.65	72.70	79.89
ImageNet-ViT [20]	71.71	78.11	76.09	83.97	77.44	85.24
CLIP-ViT-BERT [21]	66.48	71.45	76.09	76.64	77.08	83.53
GLoRIA-R50 [12]	74.79	82.13	76.29	83.28	78.55	87.15
MGCA-ViT [26]	74.22	82.13	76.37	83.02	79.79	88.11
MRM-ViT [29]	72.98	80.93	76.37	84.70	78.77	86.63
MedCLIP-Swin [27]	75.55	83.41	77.33	85.79	78.80	87.36
Ours-Prefix [15]	76.40	83.99	78.38	85.55	79.34	88.52
Ours-IA3 [16]	74.52	82.26	77.59	85.47	79.99	88.59
Ours-LoRA [11]	**77.53**	**84.81**	**78.38**	**86.22**	**80.36**	**88.72**

Table S3. Model Trainable Parameter Size. We list the total trainable model size, trainable vision encoder size, and trainable language model size for each baseline and our method.

Model Name	**Total Trainable Size**	**Vision Encoder Size**	**Language Model Size**
ViT-B-14 [20]	90.42M	90.42M	-
CLIP-ViT-BERT [21]	153.59M	90.42M	63.16M
ConVIRT-R50-BERT [28]	108.13M	25.08M	83.05M
GLoRIA-R50-BERT [12]	108.14M	25.08M	83.05M
MGCA-ViT [26]	168.86M	85.80M	83.05M
MRM-ViT [29]	168.85M	85.79M	83.05M
MedCLIP-Swin-BERT [27]	110.98M	27.91M	83.05M
Ours-ViT-GPT2(Prefix) [15]	93.04M	90.42M	2.62M
Ours-ViT-GPT2(IA3) [16]	90.99M	90.42M	0.57M
Ours-ViT-GPT2(LoRA) [11]	93.70M	90.42M	3.27M