Supplementary Material of MedBLIP: Bootstrapping Language-Image Pretraining from 3D Medical Images and Texts

Qiuhui Chen[0000−0002−1210−1306] and Yi Hong✉[0000−0002−9065−3691]

Department of Computer Science and Engineering,
Shanghai Jiao Tong University, Shanghai 200240, China
✉yi.hong@sjtu.edu.cn

1 Confusion Matrix

Table 1 reports our NC/MCI/AD multi-classification performance on AIBL test dataset.

Table 1. Confusion matrix.

Actural \ Predicted	NC	MCI	AD
NC	181	13	6
MCI	40	156	4
AD	19	33	148

2 Ablations on the choice of image encoders

To assess the versatility of our method, we employ two distinct ViT image encoders: image encoder from EVA-CLIP and image encoder from BioMedCLIP [2]. As shown in Tab 2, the performance differences between two image encoders are marginal, illustrating that our method is agnostic to the choice of image encoders.

Table 2. Ablation on image encoders.

	AIBL	MIRIAD
EVA-CLIP	**80.8**%	71.0%
BioMedCLIP	80.6%	**71.4**%

3 Adding SOTA baseline

We compare our method with recent peer reviewed work Phi-2 [1] on AIBL and MIRIAD dataset.

Table 3. Quantitative comparison.

Methods		LM size	Learn	AIBL	MIRIAD
Phi-2	Text only		-	69.8%	54.3%
Ours w/ Phi-2	Frozen	2.7B	151M	78.3%	72.7%
	LoRA		159M	**81.3%**	**74.6%**

References

1. Javaheripi, M., Bubeck, S., Abdin, M., Aneja, J., Bubeck, S., Mendes, C.C.T., Chen, W., Del Giorno, A., Eldan, R., Gopi, S., et al.: Phi-2: The surprising power of small language models. Microsoft Research Blog (2023)
2. Zhang, S., Xu, Y., Usuyama, N., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., et al.: Large-scale domain-specific pretraining for biomedical vision-language processing. arXiv preprint arXiv:2303.00915 **2**(3), 6 (2023)