

Procedure	Dataset/Model					
	ADHD-200				MCIC	OASIS
	MCIAT ₁	MCIAT ₂	MCIAT ₃	MCIAT ₄	MCIAT ₃	MCIAT ₃
Batch size	64	64	64	64	12	16
Optimizer	AdamW	AdamW	AdamW	AdamW	AdamW	AdamW
pre-training epochs	2000	2000	2000	2000	3000	3000
LR	×	×	×	×	×	×
Base LR	1.0E-05	5.0E-05	1.0E-05	8.0E-05	1.0E-04	1.0E-04
LR decay	cosine	cosine	cosine	cosine	cosine	cosine
Layer-wise decay	×	×	×	×	×	×
Weight decay	0.05	0.05	0.05	0.05	0.05	0.05
Warmup epochs	5	5	5	5	5	5
Label smoothing	0.1	0.1	0.1	0.1	0.1	0.1
Dropout	×	×	0.1	0.1	0.1	0.1
Gradient Clip.	×	×	×	×	×	×
Mixup alpha	0.8	0.8	0.8	0.8	0.8	0.8
Init scale	0.001	0.001	0.001	0.001	0.001	0.001