
Text Pre-training
5.2T data
Pure Text Data
ViT Training
2.0T -> 0.1T data
CoCa-loss with tiny language decoder -> align to LLM
resumes LR scheduler
Joint Pre-training
1.4T data
Up to 40% Multimodal Data
Progressive Multimodal Ratio
Joint Cooldown
0.6T data
High-quality Text & Multimodal Data
Re-warmup to higher LR
resumes LR scheduler
Joint Long-context
0.3T data
Long Text & Long Video & Long Doc
RoPE base: 50,000 -> 800,000