Step 1:Pre-training on Large-scale Language and Vision Datasets

