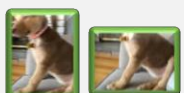
Image-Text Pairs: 6.5M

① Masked Token Loss ② Contrastive Loss

Image-Text Representation

(A **dog** is sitting on a **couch** , **Dog** **Couch** , )

Word-Tag-Region Triplet

Understanding

- VQA
- GQA
- NLVR2
- Image-Text Retrieval
- Text-Image Retrieval

Generation

- Image Captioning
- Novel Object Captioning

Pre-training → Fine-tuning