

Image-Text Pairs: 6.5M

- ① Masked Token Modeling
- ② Contrastive Loss

Image-Text Representation



Pre-training



Fine-tuning

Understanding

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

Generation

- Image Captioning
- Novel Object Captioning

Fig. 1. OSCAR pipeline. The model takes a triplet as input sequence, is pre-trained with two objectives (a masked token loss over words & tags, and a contrastive loss between tags and others), and fine-tuned for 5 understanding and 2 generation tasks (detailed in Section 4).