ABSTRACT

The present disclosure provides a multi-modal pre-training method and apparatus. The method including: sampling a video in a video-text pair to obtain a first video frame sequence; performing word segmentation processing on a text in the video-text pair to obtain a first word segmentation sequence; masking on the first video frame sequence to obtain a second video frame sequence; masking on the first word segmentation sequence to obtain a second word segmentation sequence; encoding the first video frame sequence to obtain a first video feature, and encoding the first word segmentation sequence to obtain a first word segmentation feature; encoding the second video frame sequence to obtain a second video feature, and encoding the second word segmentation sequence to obtain a second word segmentation feature; performing multi-modal pre-training by using the first video feature, the first word segmentation feature, the second video feature and the second word segmentation feature.