### Contribution of each member

|  |  |
| --- | --- |
| **Group Member** | **Contribution** |
| Sha Guo(郭莎) | To capture short-term and long-term motion structure, develop a optical flow modality module, which uses a SOTA pre-trained deep learning based optical flow model PWC-Net to calculate dense motion optical flow between sampled frames. Then extend RGB input modality to RGB and optical flow field by simply adding their embedding together before sending them to Swin architecture. By combing these two inputs, build the better video caption system, which has numerous potential applications in real-world problems. |
| He Huang(黄河) | Look into the sparse masking strategy and propose a window sparse mask.  **To do:** build an input-feature based attention generation network with Gumbel softmax to produce sparse and variant attention mask |
| Zehong Ma(马泽红) | To adaptively select the sparse frames from videos, propose an adaptive frame sampling module. The sampling module calculates the difference between two adjacent frames which are preprocessed by an 7\*7 average kernel as motion representation and then leverage a motion-uniform sampling strategy based on the cumulative motion distribution to ensure the sampled frames evenly cover all the important segments with high motion salience.  To take full advantage of the great multimodal representations in visual-language pretraining, replace the encoder of SWINBERT with CLIP and proposed a new model called VC-CLIP. Each selected frame of video is encoded into a visual token by CLIP visual encoder. Each word is separately encoded into a non-contextual textual token by textual encoder in CLIP. The multimodal transformer take as input the visual and textual tokens encoded by CLIP, which makes the multimodal transformer mainly focus on the interaction of multimodal representation. |

### Project link: