Typical techniques for video captioning follow the encoder-decoder framework, which can only focus on one source video being processed. A potential disadvan