Most existing attention-based methods on image captioning focus on the current word and visual information in one time step and generate the next word, witho