







In our tasks, the inputs are:

(1) a question with 5 candidate answers; (2) a 60-second long video; (3) a set of subtitle sentences, and our goal is to predict the correct answer as well as ground the answer both spatially and temporally.