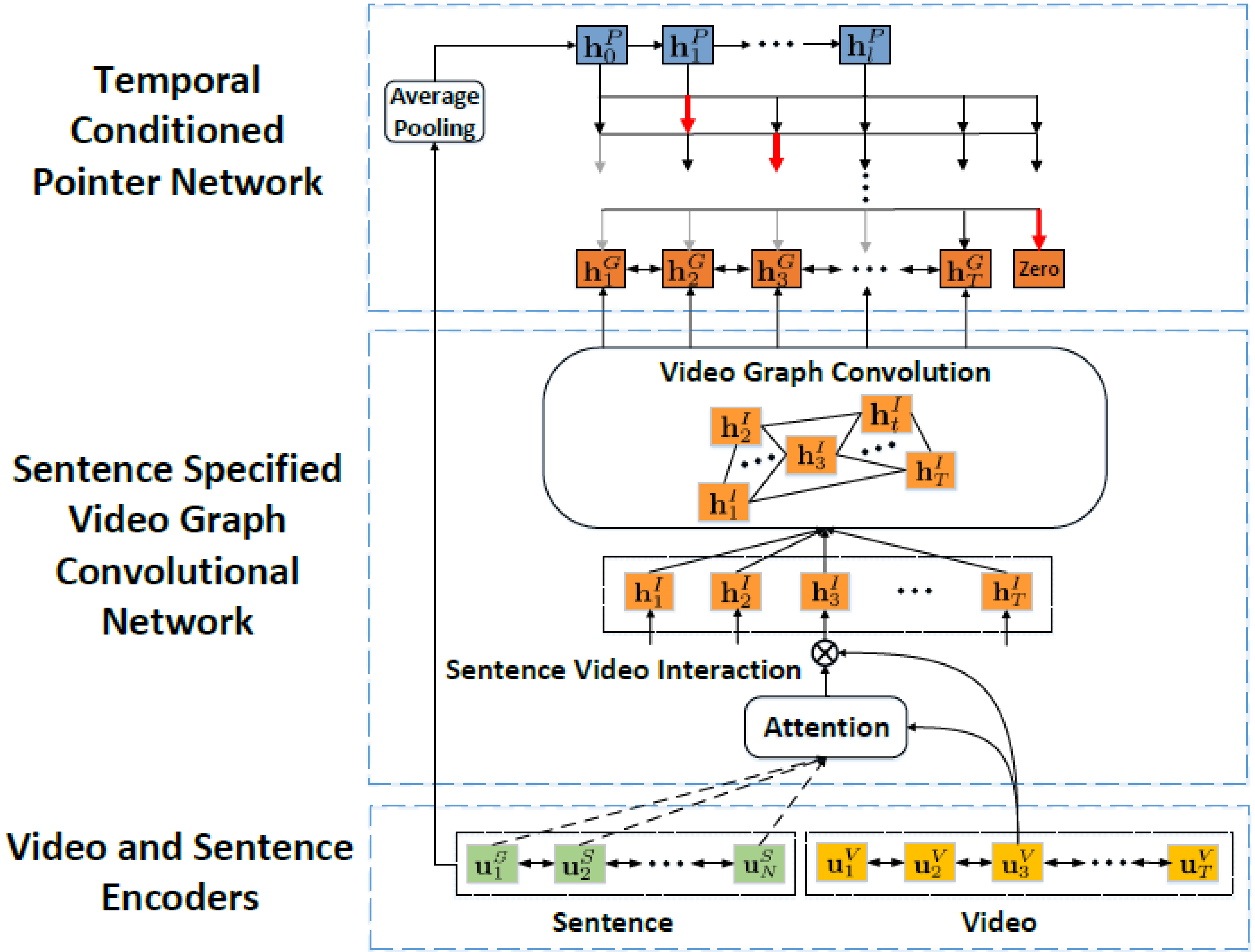

Figure 2: The architecture of our GTP model, which consists of three modules. First, the video and sentence encoders aggregate the contextual evidences from the video clip representations and word embeddings of the sentence query, respectively. Second, the sentence specified video graph convolutional network establishes the fine-grained word-by-clip interaction between the sentence and video, and leverages a GCN to further exploit the sentence specified video clip relationships. Finally, the temporal conditioned pointer network predicts and concatenates the video clips to yield the video thumbnail in a sequential manner.