

Peer Review

Review of: "VideoLLM Knows When to Speak: Enhancing Time-Sensitive Video Comprehension with Video-Text Duet Interaction Format"

Haozhao Wang¹

1. Huazhong University of Science and Technology, China

This paper focuses on the video-text LLM processing task, especially for the real-time task, which has been seldom considered before.

Strengths:

1. The task is new and important.
2. The idea of inserting the text into multiple segments makes sense and is well motivated.
3. The improvement of the experimental results is significant.

Weaknesses:

1. The writing is not formal and needs improvement.
2. There are too many heuristic hyperparameters. For example, this paper inserts a text when there is a 50% change in the video, which is too heuristic.
3. The technical contribution may not be sufficient. The position of insertion needs more consideration.
4. The organization can be improved. Data construction should be in an isolated section instead of the method.

Declarations

Potential competing interests: No potential competing interests to declare.