Qeios

Peer Review

Review of: "LLaVA-MR: Large Language-and-Vision Assistant for Video Moment Retrieval"

Zekun Yang[1]

1. Nagoya University, Japan

1. The research motivation is not clear. Where were the applications of LLM before? Why does the author think LLM is suitable for video moment retrieval?
2. It can be understood that a key frame captures dynamic variation, while a non-key frame carries redundant information. The key/non-key is supposed to be pre-defined; otherwise, it may have a negative effect on the results. But in this work, such frames are divided by the Informative Frame Selection module. Is it possible to give more details? Like, how is the accuracy of such a definition determined?
3. The use of LLM is a main contribution in this work, but the author does not claim which LLM he/she is using and why he/she selected this LLM model.
4. Experiment conditions are not well explained. In what environment did the author conduct the experiment?

Declarations

Potential competing interests: No potential competing interests to declare.

doi.org/10.32388/AVGoUF