21 December 2024 · CC-BY 4.0

Qeios

Peer Review

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

Werner Bailer[1]

1. DIGITAL - Institute for Digital Technologies, Joanneum Research Forschungsgesellschaft mbH, Graz, Austria

The paper proposes an approach for improving moment retrieval based on multimodal LLMs by improving temporal localisation and adaptively compressing time to better handle context limitations.

The approach appears to be sound, and the results are interesting. However, there are some clarifications and improvements needed:

The name of the method is somewhat misleading. LLaVA seems to make reference to "LLaVA: Large Language and Vision Assistant" (Liu et al., NeurIPS 2023), but in fact, it seems that the paper is unrelated to this method. It is thus suggested to find a less misleading name.

In Section 3.1, there are some questions related to Q-Former. First, a reference to the method used must be provided. Second, as the method is query-based, it should be stated clearly which stages in the processing pipeline are query-agnostic and which are query-dependent. It should also be clarified whether the inference time includes this processing of the video or not.

The discussion about handling the different cases of the sampling rate and handling rounding errors is not clear. Why would a single sufficiently precise time addressing, such as milliseconds, not solve the problem?

The equations for K^v and N^v in sec. 3.3 are not clear. First, it is not defined over which range j resp. l are running, and second, j seems to have a different meaning in the two grouped equations.

The results provided in Table 1 show indeed the good performance of the method. However, the authors should also consider SG-DETR ("Saliency-Guided DETR for Moment Retrieval and Highlight Detection", Oct. 2024) in their comparison.

doi.org/10.32388/NNYIX6

Fig. 5 is unclear; both axes have no scale. In addition, a grid might improve readability.

The authors should clarify how they selected the hyperparameters and which shares of the dataset were used to select the hyperparameters. Also, Table 7 reports results for Charades-STA. Have the same hyperparameters been used for both datasets?

The authors state that the code will be released upon acceptance. As the review process is not double-blind, there is no reason not to make the code available for review.

There are some formatting errors in the PDF, such as missing spaces between text and variables (e.g., query q on p. 5, to d on p. 6), and symbols with subscript and fractions being printed above the line (e.g., p. 6).

Declarations

Potential competing interests: No potential competing interests to declare.