Qeios

Peer Review

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

Werner Bailer[1]

1. Joanneum Research Forschungsgesellschaft mbH, Graz, Austria

The paper aims to provide descriptions or questions to answers to video segments with low latency, enabling real-time interaction with a video. This is a relevant task where there are gaps in the literature.

However, there are some choices in the design of the proposed approach that seem arbitrary and not well motivated.

* First, the approach uses fixed-length segments, and from the demo, it seems that in many cases, new descriptions/answers are also provided in more or less fixed intervals (even if they are identical to the previous one). Using a content-based segmentation, e.g., using shots and possibly subshots (if needed), seems to be a more logical approach in order to find the time when a new description is needed.

* Second, if the authors aim for the lowest possible latency, why do they train with insertion points that are never in the first half of the segment? In a more or less static shot, it may already be feasible to provide a useful description earlier.

* Third, the authors claim an advantage over VideoLLM-online is their use of heads for information and relevance - but both values are labelled in training in a quite arbitrary manner, completely ignoring where in the video new information is provided. Determining appropriate training values could, for example, either be determined by analysing saliency in the video or generating descriptions of subsegments of different lengths using LLMs and determining their overlap and differences.

doi.org/10.32388/B7WJ8B

* It is not clear how much the three temporal video grounding datasets contribute to the success of the training; they might be more important for choosing the informative and relevance information than the provided ground truth. This should be studied further.

* It is not clear why the task of creating descriptions or answers at low latency is mixed with a dialogue scenario that includes pausing the video and resuming after a message, and a specific interaction format. It seems generating and updating descriptions or multiple answers to a question along the video is a relevant task on its own.

There are some points that should be addressed in the evaluation:

* For the zero-shot description task, it is not clear if the results of MMDuet are generated at the end of the segment or the earliest possible time.

* As the authors claim that VideoLLM-online is a closely related approach, the authors should evaluate against it.

Declarations

Potential competing interests: No potential competing interests to declare.