EgoCVR: An Egocentric Benchmark for Fine-Grained Composed Video Retrieval

Thomas Hummel, Shyamgopal Karthik, Mariana-Iuliana Georgescu, Zeynep Akata

Accepted for ECCV 2024 (Paper) (GitHub)

The authors resolve issues with the WebVid-CoVR-Test dataset, specifically most of the modification texts involve a color, shape, or object change which doesn't require temporal understanding, and that captions are only considered similar if they have a one word difference. Also, they introduce a training-free method that achieves stronger performance on their new evaluation dataset.

Motivation

85% of samples in the WebVid-CoVR-Test dataset focused on object-centered modifications, such as a color, shape, or object change, which doesn't require temporal understanding and only requires image-level understanding. Also, the WebVid-CoVR-Test dataset was created by searching for single-word differences in video captions, which does not properly evaluate the generalizability to different prompt formats.

Method

Evaluation Dataset

1. 1,250 long videos are taken from the Ego4D FHO dataset.
2. 155k annotated clips are extracted from the videos, ranging from 2 to 8 seconds.
3. The clips are filtered for temporal overlap, resulting in 9k distinct clips.
4. Pairs are manually identified from video clips in the same video, looking for similarity in captions, except for a single action or object change.
5.

Limitations