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

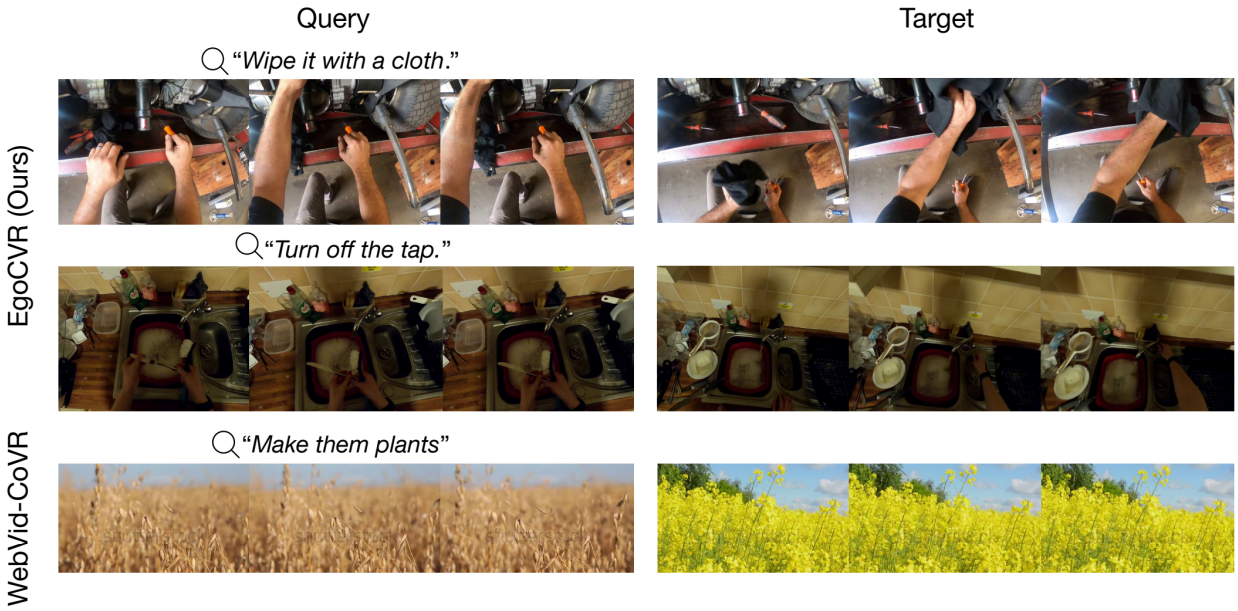
Thomas Hummel, Shyamgopal Karthik, Mariana-Iuliana Georgescu, Zeynep Akata

Accepted for ECCV 2024 (Paper) (GitHub)

The paper introduces EgoCVR, an evaluation benchmark for fine-grained CVR that focuses on temporal video understanding rather than object modifications. The authors also introduce a training-free method that combines LLM-based text refinement and visual filtering for composed video retrieval.

Motivation

85% of WebVid-CoVR-Test samples involve object-centered modifications, such as color or shape changes, which can be solved with image-based methods and do not require temporal understanding. Also, videos in the WebVid-CoVR-Test dataset were paired through single-word differences in video captions, which does not properly test models on diverse and complex modifications.



Method

Evaluation Dataset

1. 155k annotated clips were extracted from 1,250 long videos in the Ego4D dataset.
2. The clips were filtered for temporal overlap, resulting in 9k distinct video clips.
3. Pairs were manually identified from video clips in the same video, looking for similarity in captions, specifically a single action or object change, with an emphasis on temporal modifications, resulting in 2,295 high-quality pairs.
4. GPT-4 was few-shot prompted to generate the modification text between the video pair.

5. The authors include similar video clips that were from the same long video and don't have the same action as the target as a distraction.

TFR-CVR

1. The source video is captioned and then combined with the modification text using an LLM to generate a target video caption.
2. The target video caption is used to perform text-based retrieval.
3. The video database is filtered for visual similarities with the source video to ensure videos are both textually relevant and visually similar.

Method	Video Model	Textual Input	Visual Input	Fusion Strategy	Global R@1	Global R@5	Global R@10	Local R@1	Local R@2	Local R@3
Random	✗	✗	✗	-	0.01	0.05	0.1	25.3	38.2	50.7
CLIP	✗	✓	✗	-	0.7	1.7	2.7	33.5	48.8	61.8
BLIP	✗	✓	✗	-	0.4	1.4	2.7	32.5	46.9	59.7
EgoVLPv2	✓	✓	✗	-	1.7	3.9	7.2	41.0	57.3	69.0
LanguageBind	✓	✓	✗	-	0.9	2.7	4.2	34.2	51.1	64.1
CLIP	✗	✗	✓	-	7.4	33.2	55.3	26.1	43.4	57.7
BLIP	✗	✗	✓	-	6.5	32.6	55.3	26.5	43.7	57.5
EgoVLPv2	✓	✗	✓	-	7.6	32.5	49.6	27.5	44.3	59.1
LanguageBind	✓	✗	✓	-	6.1	33.1	53.4	26.1	42.9	57.7
CLIP	✗	✓	✓	Avg	7.5	33.6	**55.6**	26.4	43.7	57.9
BLIP	✗	✓	✓	Avg	8.7	32.9	52.8	29.5	45.9	61.0
EgoVLPv2	✓	✓	✓	Avg	9.5	34.9	52.1	30.7	51.3	66.0
LanguageBind	✓	✓	✓	Avg	6.1	33.2	53.5	26.1	43.1	57.8
$\text{BLIP}_{\text{CoVR}}$ [51]	✗	✓	✓	Cross-Attention	5.4	15.2	24.3	33.1	49.5	62.9
$\text{BLIP}_{\text{CoVR-ECDE}}$ [47]	✗	✓	✓	Cross-Attention	6.0	16.3	24.8	33.4	49.3	63.0
CIReVL [28]	✗	✓	✓	Captioning	2.0	6.8	10.6	33.6	49.7	61.4
`TFR-CVR` (Ours)	✓	✓	✓	Captioning	**14.1**	**39.5**	54.4	**44.2**	**61.0**	**73.2**

Limitations

1. The benchmark does not include a training set.
2. The dataset only consists of egocentric videos and may not generalize well to third-person videos.
3. LLM-generated captions occasionally misrepresented the video content.