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

Thomas Hummel*[1,2], Shyamgopal Karthik*[1,2], Mariana-Iuliana Georgescu[2], Zeynep Akata[2,3]

[1]Tübingen AI Center, University of Tübingen
[2]Helmholtz Munich, MCML [3]TU Munich
{thomas.hummel, shyamgopal.karthik}@uni-tuebingen.de

In this appendix, we report results when applying re-ranking to other methods in Section 1, show failure cases and future directions of TFR-CVR in Section 2, and report results with TFR-CVR on WebVid-CoVR-Test [4] in Section 3. Furthermore, we provide more details on EgoCVR's diversity (Section 4.1), present the instruction prompts given to the LLM models to create the text modification (Section 4.2), to create the target captions for TF-CVR (Section 4.3) and to perform the temporal event detection analysis on the CVR benchmarks (Section 4.4), as well as provide additional qualitative examples (Section 5).

1 Re-Ranking Applied to Other Methods

We show the results for applying re-ranking using the same LanguageBind encoder for all the methods below. While the re-ranking improves $\text{BLIP}_{\text{CoVR}}$, the overall performance and improvement are much larger for TFR-CVR.

Table 1: Results on EgoCVR in terms of R@1, R@5 and R@10 on the global setting with and without applying re-ranking. We also report the mean recall change when applying the re-ranking.

Method	w/o re-ranking			w/ re-ranking			Δ R@{1,5,10}
	R@1	R@5	R@10	R@1	R@5	R@10	
CLIP [3]	7.5	33.6	**55.6**	7.5	33.5	54.4	0.4 ↓
BLIP [1]	8.7	32.9	52.8	8.7	33.8	54.2	0.8 ↑
$\text{BLIP}_{\text{CoVR}}$ [4]	5.4	15.2	24.3	10.4	31.9	52.2	16.5 ↑
TFR-CVR	4.4	12.9	18.3	14.1	39.5	54.4	**24.1** ↑

2 Failure Cases and Future Directions

Due to the modular approach of TFR-CVR, we can break down the failure cases. As shown in Table 4 in the main paper, employing the ground-truth (GT) captions

* Denotes equal contribution

achieves only an R@1 of 51.7% with a gallery of 7 candidates (Local). Therefore, the biggest source of improvement on EgoCVR would be from applying stronger text-to-video retrieval models. We can also trace back errors to erroneous video captions. In Figure 1, we show an example of the wrong retrieval caused by the text-to-video retrieval method at the bottom. In addition, we highlight an error caused by the video captioning method on the top. Here, the *"pot"* was mistaken for a *"bowl"* during the captioning, leading to the retrieval of a video that mainly depicts a bowl.

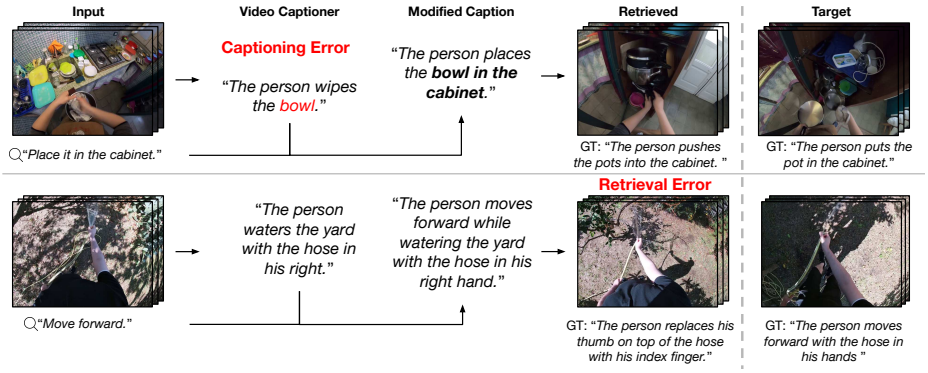


Fig. 1: Qualitative depiction of failure cases of TFR-CVR. The modular approach of TFR-CVR allows us to trace back failure cases mostly to two main sources: video *captioning errors* (top) and text-to-video *retrieval errors* (bottom).

3 Results on WebVid-CoVR Benchmark

We additionally show results of TFR-CVR on the WebVid-CoVR-Test [4] benchmark in Table 2. For TFR-CVR, we employ LanuguageBind [5] as the visual and textual encoder, since it was trained on different types of videos, unlike EgoVLPv2 [2] which is restricted to egocentric videos. We observe that TFR-CVR is able to achieve the best results among all methods that do not explicitly train on the WebVid-CoVR training set. The performance (R@1 of 51.7%) is also quite competitive with the state-of-the-art result obtained by $\text{BLIP}_{\text{CoVR}}$ [4] (R@1 of 53.1%). We also continue to notice a constant benefit from applying the re-ranking strategy. After applying the visual-only filtering of LanguageBind (which in itself is a weak model scoring an R@1 value of 43.2%), the performance of the text-based TF-CVR method improves from R@1 of 48.4% to 51.7%.

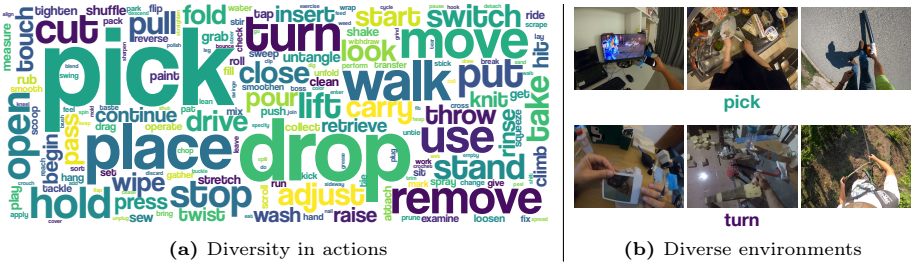
Table 2: Results on WebVid-CoVR-Test [4] in terms of Recall@1, Recall@5 and Recall@5. We report results with models evaluated in the Zero-Shot setting [1, 3, 5] (including our TF-CVR and `TFR-CVR`) and with models specifically trained for this data set [4].

Method	Zero-Shot	Fusion Strategy	R@1	R@5	R@10
Random	✓	-	0.08	0.23	0.35
CLIP [3]	✓	Avg	44.4	69.1	77.7
BLIP [1]	✓	Avg	45.5	70.5	79.5
LanguageBind [5] (stage 1; visual)	✓	-	43.2	66.3	75.2
TF-CVR	✓	Captioning	48.4	73.7	**81.9**
`TFR-CVR` (LanguageBind → TF-CVR) ($n_c = 20$)	✓	Captioning	**51.7**	**75.3**	80.7
$\text{BLIP}_{\text{CoVR}}$ [4]	✗	Cross-Attention	53.1	79.9	86.9

4 Additional Details to EgoCVR

4.1 Dataset Diversity

EgoCVR entails 179 different actions across 47 distinct scenarios (top-6: *crafting, cooking, cleaning, construction, carpenter, car mechanic*). In Figure 2, we show (i) a word cloud illustrating the various actions included in EgoCVR, and (ii) exemplary of diversity in action environments ("*pick*" appears in 30 and "*turn*" in 9 scenarios). EgoCVR offers wide-ranging environments, objects and actions.



(a) Diversity in actions **(b)** Diverse environments

Fig. 2: Diversity of actions and environments in EgoCVR.

4.2 Creating Video Modification Instructions

In this section, we give more details on how we create video modification instructions. We automatically generate instructions from query and target clip narrations. More specifically, given the narrations of the source and the target video, the goal is to generate the instruction (modification) text. To achieve this, we use 15 in-context examples showing how this could be done effectively, focusing only on the modification and not the source and target captions itself.

Dataset Generation (Instruction) Prompt

I have 2 videos. Given a brief description of the source and the target video, write an instruction that describes the transformation from the source to the target. The caption you generate should only talk about the necessary modifications. Keep the instruction as short as possible, and focus always on the action. Mention objects only when absolutely necessary. You should not describe objects common to both descriptions, instead use pronouns. Describe only the transformation required. Use the examples below for reference.

Source Narration: #C C picks up the jug.
Target Narration: #C C cleans the jug.
Instruction: The person is cleaning.

Source Narration: #C C picks a spanner from the table with his right hand.
Target Narration: #C C picks a gasket from a table with his right hand.
Instruction: Gasket being picked up.

Source Narration: #C C picks up the wood from the shelf with his left hand
Target Narration: #C C detaches a wood from the wooden structure with his right hand
Instruction: Person uses the other hand and detaches.

Source Narration: #C C fixes the bolt on the motorbike with his right hand.
Target Narration: #C C holds the part of the motorbike with his left hand.
Instruction: Person holds it with the other hand.

Source Narration: #C c climbs up the steps.
Target Narration: #C c climbs down the steps
Instruction: The person climbs down.

Source Narration: #C C Wipes as paint brush with a paper towel.
Target Narration: #C C dips brush in water.
Instruction: Dip the object in water.

Source Narration: #C C pours the water in the shoe.
Target Narration: #C C rinses the shoe.
Instruction: Rinse it instead.

Source Narration: #C C puts electric shoe cleaner on the sink
Target Narration: #C C puts shoe in the sink
Instruction: Same action with a shoe.

Source Narration: #C C puts down penetrant oil
Target Narration: #C C sprays the oil
Instruction: Spray it.

Source Narration: #C C moves the scissors aside
Target Narration: #C C moves the coins aside
Instruction: Change it to coins.

Source Narration: #C C fixes the lawn mower basket
Target Narration: #C C holds the lawn mower basket
Instruction: Hold it instead.

Source Narration: #c c sits on the mat
Target Narration: #C C kneels on the carpet
Instruction: Kneels instead.

Source Narration: #C C drops the plate of food on the sink slap.
Target Narration: #C C picks a plate from the sink slap.
Instruction: Pick it up.

Source Narration: #C C holds the basket of flowers on the floor with her left hand.
Target Narration: #C C puts the white flowers on the tray with her right hand.
Instruction: Transfer it to the tray.

Source Narration: #C C scrapes carrot remains from the grater into the brown bowl
Target Narration: #C C picks carrot from the brown bowl with her right hand
Instruction: Pick it up from the bowl.

4.3 Creating Target Caption

We present the prompt utilized to obtain a valid target caption from the query video caption and the instruction text. Our goal is to take a video caption along with an instruction specifying some changes and then generate a valid target caption. Similar to the dataset generation process, this also uses a few in-context examples to improve the quality of the generated captions.

TF-CVR Prompt

I have a video. Given a brief description of the source video and a instruction that modifies it, write a description of the target video. Keep the modified description as short as possible, while being complete. Mention objects only when absolutely necessary. Use the examples below for reference.

Source Narration: #C C picks up the jug.
Instruction: The person is cleaning.
Target Narration: #C C cleans the jug.

Source Narration: #C C picks a spanner from the table with his right hand.
Instruction: Gasket being picked up.
Target Narration: #C C picks a gasket from a table with his right hand.

Source Narration: #C C picks up the wood from the shelf with his left hand.
Instruction: Person uses the other hand and detaches.
Target Narration: #C C detaches a wood from the wooden structure with his right hand.

Source Narration: #C C fixes the bolt on the motorbike with his right hand.
Instruction: Person holds it with the other hand.
Target Narration: #C C holds the part of the motorbike with his left hand.

Source Narration: #C c climbs up the steps.
Instruction: The person climbs down.
Target Narration: #C c climbs down the steps.

Source Narration: #C C Wipes as paint brush with a paper towel.
Instruction: Dip the object in water.
Target Narration: #C C dips brush in water.

Source Narration: #C C pours the water in the shoe.
Instruction: Rinse it instead.
Target Narration: #C C rinses the shoe.

Source Narration: #C C puts electric shoe cleaner on the sink.
Instruction: Same action with a shoe.
Target Narration: #C C puts shoe in the sink.

Source Narration: #C C puts down penetrant oil.
Instruction: Spray it.
Target Narration: #C C sprays the oil.

Source Narration: #C C moves the scissors aside.
Instruction: Change it to coins.
Target Narration: #C C moves the coins aside.

Source Narration: #C C fixes the lawn mower basket.

Instruction: Hold it instead.
Target Narration: #C C holds the lawn mower basket.

Source Narration: #C C sits on the mat.
Instruction: Kneels instead.
Target Narration: #C C kneels on the carpet.

Source Narration: #C C drops the plate of food on the sink slap.
Instruction: Pick it up.
Target Narration: #C C picks a plate from the sink slap.

Source Narration: #C C holds the basket of flowers on the floor with her left hand.
Instruction: Transfer it to the tray.
Target Narration: #C C puts the white flowers on the tray with her right hand.

Source Narration: #C C scrapes carrot remains from the grater into the brown bowl.
Instruction: Pick it up from the bowl.
Target Narration: #C C picks carrot from the brown bowl with her right hand.

4.4 Analysing Modification Instructions for Temporal vs. Object Events

We analyse video modification instructions for both WebVid-CoVR [4] and our EgoCVR benchmark to study the type of modifications existing in the data sets. To categorize modification instructions as temporal or object-centric, we employ GPT-4. In the prompt, apart from a description of the task itself, we also provide a few in-context examples shown below.

Temporal Event Detection Prompt

I have an instruction to modify a video. Looking at just the instruction, you should decide whether the instruction is focused on temporal events such as actions, or if it is just focused on objects. The answer you generate should only be "yes" or "no". Use the examples below for reference.

Instruction: have him fishing
Answer: yes

Instruction: turn it red on a watercolor stain.
Answer: no

Instruction: make the sax player into a drummer
Answer: no

Instruction: the girl is crying
Answer: yes

Instruction: change the meat to prawns
Answer: no

Instruction: remove the man.
Answer: no

Instruction: dip it into the paint.
Answer: yes

Instruction: cut the carrot instead.
Answer: no

Instruction: insert it into the roof.
Answer: yes

Instruction: pick it up instead.
Answer: yes

5 Additional Qualitative Examples

We illustrate a few more examples extracted from `EgoCVR` in Figure 3. It is clearly visible that the data set contains a wide variety of activities in different environments, while the examples typically focus on action focused changes.

We also illustrate resulting order of the videos obtained after re-ranking performed by `TFR-CVR` in Figure 4 and Figure 5. We observe that the correct video is not in the first position in the first stage, however, after performing the second stage, the correct video is in the top position.

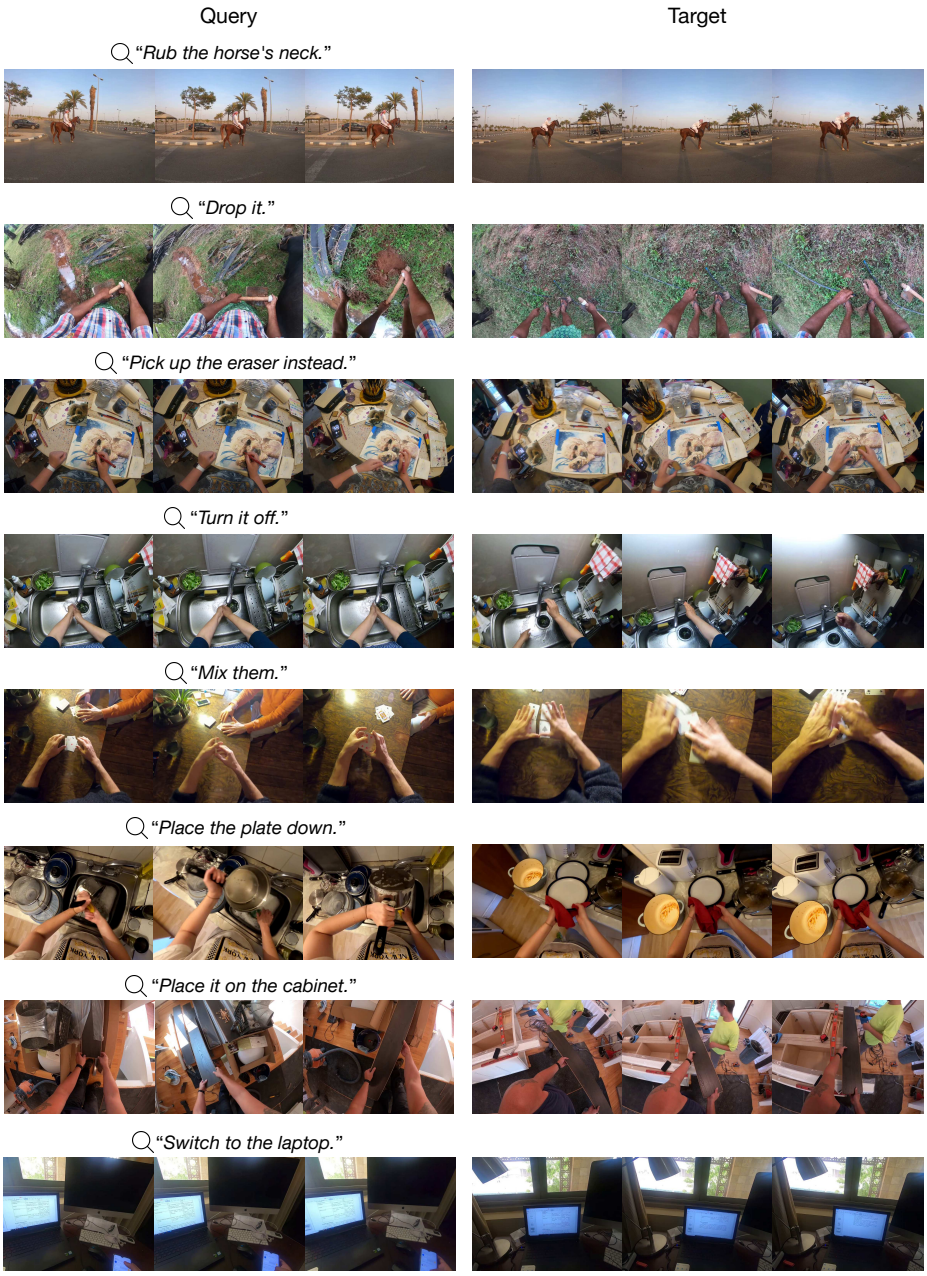


Fig. 3: Additional examples from our `EgoCVR` benchmark. We present the input video, the text modification and the target video. The dataset is diverse, covering various types of scenes and activities.

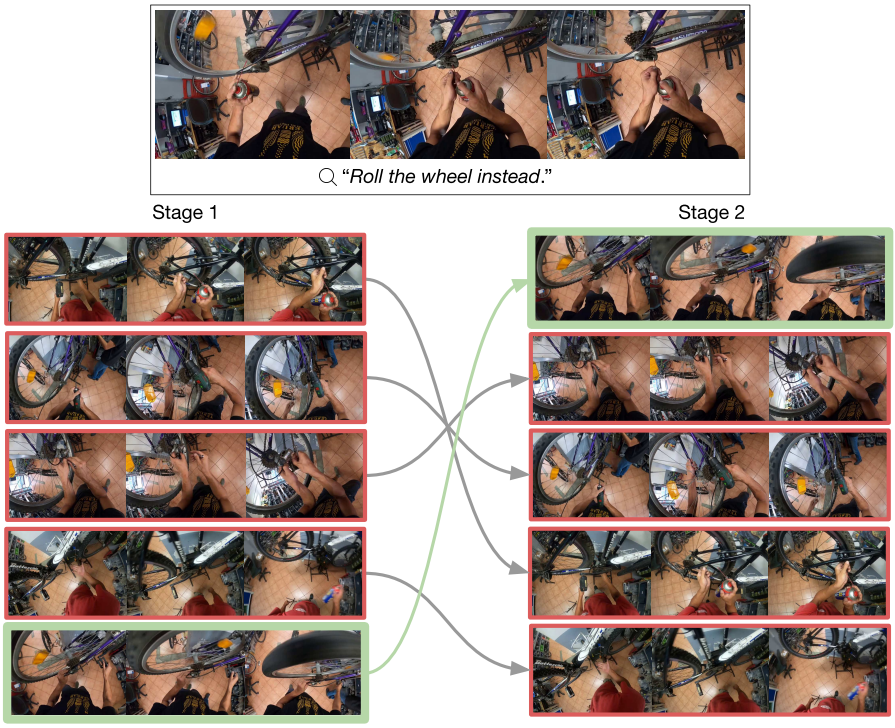


Fig. 4: Qualitative example for the first and the second stage ranking results of our `TFR-CVR` method for the query instruction "`Roll the wheel instead.`". The arrows indicate how the ranking was changed after re-ranking. The correct video is showcased in green.

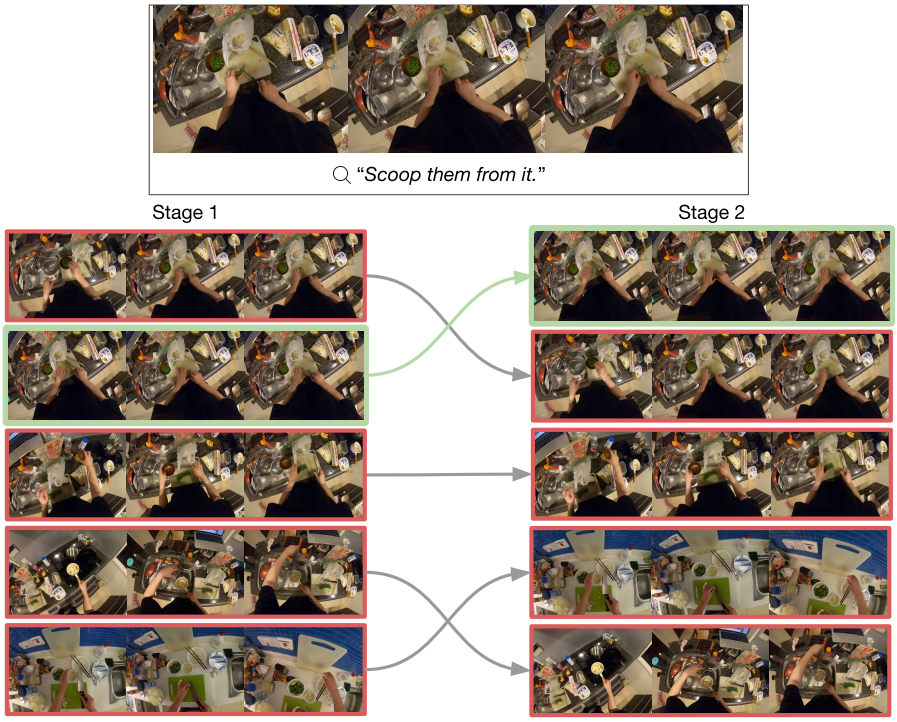


Fig. 5: Qualitative example for the first and the second stage ranking results of our `TFR-CVR` method for the query instruction "`Scoop them from it.`". The arrows indicate how the ranking was changed after re-ranking. The correct video is showcased in green.

References

1. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: ICML (2022) 1, 3
2. Pramanick, S., Song, Y., Nag, S., Lin, K.Q., Shah, H., Shou, M.Z., Chellappa, R., Zhang, P.: EgoVLPv2: Egocentric video-language pre-training with fusion in the backbone. In: ICCV (2023) 2
3. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021) 1, 3
4. Ventura, L., Yang, A., Schmid, C., Varol, G.: CoVR: Learning composed video retrieval from web video captions. In: AAAI (2024) 1, 2, 3, 6
5. Zhu, B., Lin, B., Ning, M., Yan, Y., Cui, J., Wang, H., Pang, Y., Jiang, W., Zhang, J., Li, Z., et al.: LanguageBind: Extending video-language pretraining to n-modality by language-based semantic alignment. In: ICLR (2024) 2, 3