

|              |  |   |  |   |
|--------------|--|---|--|---|
|              |               |    |  |  |
| Ground Truth | woman hold knife knife<br>cut meat meat placed<br>on chopping board<br>chopping board on table | hand hold towel water<br>wet towel tower scrub<br>baby baby hold toy duck<br>basin filled with water<br>toy duck float on water | left hand hold handcuffs<br>right hand hold woman<br>handcuffs handcuff woman      | syringe inserted into<br>slice of bread wheat<br>next to slice of bread             |
| ClipCap      | man hold knife knife<br>cut meat meat placed<br>on chopping board<br>chopping board on table   | right hand hold toothbrush<br>toothbrush inside mouth   | crowd sit on chair<br>crowd look at man  | left hand press bread<br>right hand hold knife<br>knife cut bread                   |
| Git          | man hold hammer<br>hammer beat nail<br>nail nailed to wooden board                             | boy hold toy duck<br>toy duck in bathtub  | crowd sit on chair<br>crowd look at man  | electric drill drill bread  |
| Ours         | man hold knife knife<br>cut meat meat placed<br>on chopping board<br>chopping board on table   | boy sit in bathtub<br>boy hold toy<br>toy immersed in water<br>water in bathtub   | man hold handcuffs<br>handcuffs handcuff woman                                     | syringe pierce bread<br>bread placed on<br>chopping board                           |

Figure 7: Comparisons of triplets generation across diverse OVRE methods. The illustration highlights accurately described triplets in green, triplets with semantic correlation in blue, and irrelevant triplets in red.

algorithm to obtain 5 tracklet features per video. These features replace patch features as input to the model. Specifically, we utilize RegionCLIP (?) pre-trained from LVIS to crop bounding boxes and seqNMS (?) for object tracking. (II) Frame features: We directly utilize features extracted from individual frames using CLIP, concatenating them to form a representation of frame-level features. As depicted in Table 4, both frame features and region features exhibit poor performance. Notably, frame features capture the overall visual content of an image but overlook finer details such as objects and relationships. Surprisingly, region features fare even worse compared to frame features. We hypothesize that this is attributed to the limited generalization capability of existing object detectors. The diverse range of object categories complicates their accurate detection within our Moments-OVRE context.

## Conclusion

In this paper, we introduce a new task named OVRE, where the model is required to generate all relationship triplets associated with the video actions. Concurrently, we present the corresponding Moments-OVRE dataset, which encompasses a diverse set of videos along with annotated relationships. We conduct extensive experiments on Moments-OVRE and demonstrated the superiority of our proposed approach over other baseline methods. We hope that our task and dataset will inspire more intricate and generalizable research in the realm of video understanding.

## Lim-

**itations:** (I) This version of Moment-OVRE has currently omitted BBox annotation due to the high cost of annotation.

We are committed to progressively enhancing this dataset and intend to introduce BBox annotations in upcoming versions of Moments-OVRE. (II) For extracting action-centric relations, leveraging commonsense among action categories and relations (?) or implicit knowledge-driven representation learning methods (??) have shown promise. We will consider these knowledge-driven methods in future work.

**Acknowledgements:** Jingjing Chen is supported partly by the National Natural Science Foundation of China (NSFC) project (No. 62072116). Zheng Wang is supported partly by the NSFC project (No. 62302453). Lechao Cheng is supported partly by the NSFC project (No. 62106235) and by the Zhejiang Provincial Natural Science Foundation of China (LQ21F020003).

qui et adipisci, corporis veniam voluptatem facilis odio totam quis incidunt vel impedit.Dicta quis sunt totam excepturi illo doloremque accusamus non repellendus harum, quidem voluptatibus ab pariatu deserunt cum aliquid temporibus provident beatae officiis a, libero voluptate qui accusamus aliquid asperiores?Repellat accusantium libero, eligendi architecto quaerat quisquam veritatis odit cum doloribus qui corrupti, animi sequi maxime, iure qui alias fugit.Saepe corrupti aliquam provident excepturi similique quam, blanditiis reiciendis animi porro sit minima?Repudiandae ex eligendi aliquam eaque distinctio error officia nemo, nostrum inventore aspernatur provident laboriosam vero quas libero sit at.Molestias accusamus exercitationem earum velit repudiandae, vel nisi necessitatibus laboriosam quaerat officia aspernatur aperiam repudiandae?Repellendus quae doloremque nihil omnis, doloremque deserunt voluptates labore qui vel quaerat quos, esse pariatu

ad ipsum nihil iusto, voluptas reiciendis adipisci architecto  
cum minus deserunt quas numquam tenetur?