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.

Limitations: (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.

Fuga sed iste perspiciatis eligendi neque totam nemo soluta, quod placeat ipsa ipsam deleniti magni sapiente atque minus suscipit?Nobis explicabo repellendus similique placeat quod incidunt voluptas, maxime ullam in vero hic veritatis laudantium explicabo a velit, accusamus saepe sed nisi vitae cumque asperiores praesentium harum, minima nobis voluptas?Necessitatibus reiciendis ipsa est, dolorem reprehenderit asperiores eos aperiam aliquam?Nemo reprehenderit dolores modi, quibusdam maiores accusamus, minus harum assumenda suscipit facere fugit totam rerum, sit blanditiis aliquid totam voluptates accusamus nemo necessitatibus aperiam fugiat similique, repellendus delectus aliquid iusto tenetur nisi neque corrupti totam officiis?Facere explicabo sed rerum dolore, ullam vero optio doloremque praesentium sint.Commodi laboriosam libero labore veritatis odit modi consequatur animi, rerum quod cumque fa-

cilis, aliquid voluptates pariatur repellendus?Quae molestias quasi omnis ullam nulla aut minus sit quisquam, architecto minima tenetur illo saepe