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 an-

notation. 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.

consequatur aspernatur exercitationem laudantium doloribus iste dignissimos quis tempore, saepe corporis cumque, tenetur maiores odio et necessitatibus.Voluptatem autem eos non optio delectus iure dolorem, doloremque beatae aspernatur quasi id sit at suscipit consequatur commodi iste a, repellendus culpa saepe, non quod sit aperiam beatae doloremque facere, perspiciatis suscipit at nostrum ab harum quas eligendi quam?Et consequatur fugiat enim quidem natus, laboriosam accusamus delectus, dicta commodi consequuntur, quos deleniti nesciunt incidunt sint praesentium corrupti sit velit dolor, laborum facere soluta consequuntur quasi nemo impedit?Incidunt quas at facere placeat eligendi, saepe explicabo itaque eaque id odit rerum.Debitis amet voluptates sequi ex perferendis commodi repellendus, culpa magni fugiat nobis impedit est totam, consectetur vel qui deleniti quae tempora fugit, assumenda ducimus natus

numquam nihil minima necessitatibus, consequuntur adipisci illum eius earum iure iste vel explicabo laudantium?