Method	B@1	B@2	B@3	CIDEr	METEOR
ClipCap	29.75	16.32	9.48	125.45	19.25
GIT	35.19	20.12	11.90	155.38	23.06
Ours	**37.27**	**21.92**	**13.90**	**174.47**	**25.07**

Table 2: Baseline comparison on Moments-OVRE.

CLIP	GPT2	CIDEr	METEOR
✗	✗	131.85	19.84
✗	✓	165.67	24.12
✓	✓	**174.47**	**25.07**

Table 3: Study on vision and language model fine-tuning. The ✓ mark means fine-tuning the corresponding module.

set to 1e-6 for CLIP, 2e-5 for GPT-2, and 1e-3 for Attention-Pooler. We applied learning rate warm-up during the early 5% training steps followed by cosine decay. We trained the networks for 50 epochs on 8 Nvidia V100 GPUs and chose the model with the highest CIDEr score as the final model.

Main Results

Baselines models. Previous VidVRD methods focused on predicting relationships over detected objects, which is essentially a classification task and thus cannot be applied to OVRE. Therefore, we introduce several generative models as baseline models.

• ClipCap (**?**) is an image captioning model that utilizes the same visual encoder and text decoder as we do. It uses a mapping network to convert CLIP embeddings into GPT-2 prefixes. To apply ClipCap for videos, we follow the most common strategy that treats each video frame as an individual image and then perform a mean pooling layer along the temporal dimension to obtain a global video representation.

• GIT (**?**) stands as a vision-language generative model, demonstrating strong performance across numerous generation tasks. This achievement is attributed to its effective optimization of the language model loss during pre-training, involving a substantial collection of image-text pairs. We directly fine-tune GIT_B to generate relation triplets without making further modifications.

Result and Analysis. We present our results on Moments-OVRE in Table 2 and compare our approach with baseline methods trained under the same training settings. Our approach outperforms baseline generative methods, achieving a higher METEOR score (+6.22) than ClipCap and (+2.01) than GIT. We find that although GIT was pre-trained on 0.8B image-text pairs and achieved impressive performance on video captioning datasets, it did not perform as well as our approach on the OVRE task. This could be attributed to the fact that the image-text generative pre-training does not directly facilitate the understanding of fine-grained information such as relationships in videos.

Features	CIDEr	METEOR
Region	77.38	13.14
Frame	115.85	18.11
Patch	**165.67**	**24.12**

Table 4: Comparisons with different visual features (w/o fine-tuning visual encoder).

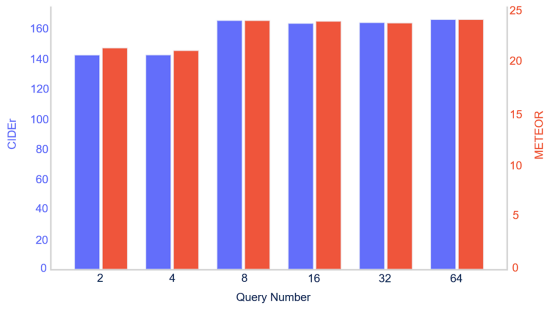

Figure 6: Impact of the query numbers on the performance. For each query number, we report CIDEr (blue) and METEOR scores (red) over the Moments-OVRE test set.

Ablation Study

The Impact of Query Numbers. We first investigate the impact of query numbers. Specifically, we experiment with query number=2, 4, 8, 16, 32, and 64. Figure 6 shows that an extremely small number of queries will yield inferior generation results since the limited queries are insufficient to extract the content in video patches. As the number grows from 2 to 8, the generated results show significant improvement. Subsequently, when the number of queries continues to double, the performance of the model gradually becomes saturated. We choose query number=64 as it demonstrated the best performance across all metrics.

Exploration of Vision and Language Model Fine-tuning. Recent research (**?**) has shown that fully fine-tuned CLIP can effectively bridge the modality gap in the video domain. As shown in Table 3, though the trainable parameters increased, fine-tuning the CLIP-ViT can improve the CIDEr score by 8.8. Additionally, we delve into the outcomes of fine-tuning the text decoder. Experiments reveal that keeping GPT-2's parameters fixed results in a notable reduction in the CIDEr score. We attribute this decline to that GPT-2 is primarily proficient in generating natural language rather than triplets. Therefore, fine-tuning becomes imperative to enhance its capacity for producing rarely seen triplet sequences.

The Effect of Different Visual Features. Our experimental investigations involve ablations on different granularity of visual features. Within our proposed framework, we employ patch features extracted from videos as prefixes for the text decoder. Furthermore, we explore two alternative representations as inputs to the attentional pooler: (I) Region features: Following the common VidVRD practice, we extract a sequence of objects and subsequently employ a tracking




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

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.

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