Open-Vocabulary Video Relation Extraction

Wentao Tian,[1] Zheng Wang,[2] †Yuqian Fu,[1] Jingjing Chen,[1]†Lechao Cheng[3]

[1] Shanghai Key Lab of Intell. Info. Processing, School of CS, Fudan University
[2] College of Computer Science and Technology, Zhejiang University of Technology
[3] Zhejiang Lab
{wttian22@m., fuyq20@, chenjingjing@}fudan.edu.cn, zhengwang@zjut.edu.cn, chenlc@zhejianglab.com

Abstract

A comprehensive understanding of videos is inseparable from describing the action with its contextual action-object interactions. However, many current video understanding tasks prioritize general action classification and overlook the actors and relationships that shape the nature of the action, resulting in a superficial understanding of the action. Motivated by this, we introduce **O**pen-vocabulary **V**ideo **R**elation **E**xtraction (**OVRE**), a novel task that views action understanding through the lens of action-centric relation triplets. OVRE focuses on pairwise relations that take part in the action and describes these relation triplets with natural languages. Moreover, we curate the Moments-OVRE dataset, which comprises 180K videos with action-centric relation triplets, sourced from a multi-label action classification dataset. With Moments-OVRE, we further propose a cross-modal mapping model to generate relation triplets as a sequence. Finally, we benchmark existing cross-modal generation models on the new task of OVRE. Our code and dataset are available at https://github.com/Iriya99/OVRE.

Introduction

Videos contain abundant semantic information, including action, actors (e.g. humans, animals, objects, and other entities), and relationships between actors. To comprehend the dynamic and complex real-world situations depicted in videos, researchers have investigated a wide range of video comprehension tasks. These endeavors allow for the transition of understanding video content from broad semantic concepts to more detailed ones. Despite the variations, all of these tasks converge on a pivotal aspect : extracting semantic information within the videos and constructing a higher-level representation to facilitate comprehension.

Foundational tasks, such as action classification (**?**) and temporal action localization (**?**), primarily center on recognizing broad-level actions within videos, yet they often overlook the specific scenario in which these actions unfold. Consequently, these tasks often struggle to offer a profound understanding of the context and the specific actors that are part of actions. In essence, they concentrate solely on deciphering "what" action is transpiring and "when" it takes place, omitting the "who" and "how" aspects. On the other hand, video captioning (**?**), video grounding (**??**), and video-text retrieval (**?**) strive to encapsulate the videos' essence through textual descriptions by mapping them into a joint semantic space. Nevertheless, these textual descriptions often provide a detailed-level overview of the action context, lacking a nuanced comprehension of relations.

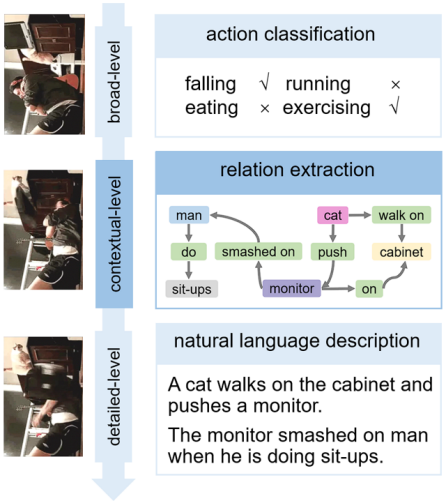

Figure 1: Open-vocabulary Video Relation Extraction enables a contextual-level comprehension of video content, bridging the gap between general action classification and precise language description.

To achieve a comprehensive comprehension of visual content, researchers have introduced Video Visual Relation Detection (VidVRD) tasks (**??**) specifically designed for video analysis. VidVRD tasks are geared toward identifying objects and the relations between them within videos. For instance, VidVRD (**?**) employs available video object detection data to non-exclusively label various simplistic relationships between objects. As depicted in Figure 2 (a), relationships like `touch` and `watch` are inherently *object-centric* relationships with limited implications for the `falling` action. Notably, Action Genome (**?**) is distinct in its dedication to an *action-centric* understanding of videos. It dissects actions into spatio-temporal scene graphs. However, it still

†Correspondence to: Zheng Wang, Jingjing Chen.
Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

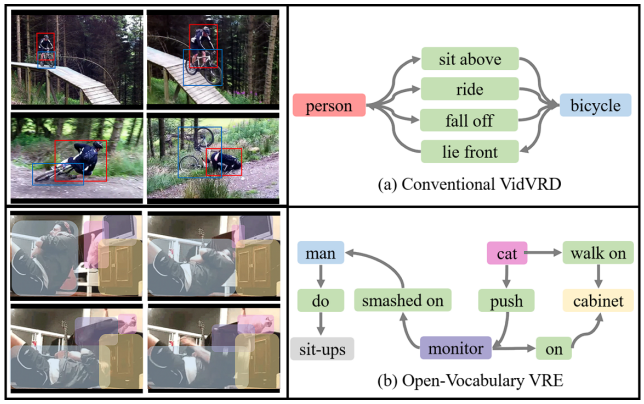

Figure 2: Two `falling` videos depicted in VidVRD and OVRE diverse in (a) salient objects interconnected throughout frames are exhaustively annotated with diverse relations in VidVRD; (b) OVRE builds a relation graph with various actors and relations that closely related to the falling action.

grapples with a challenge shared by VidVRD the constraint of a fixed vocabulary range. Consider the scenario shown in Figure 2 (a) involving a bicyclist falling from a steep bridge. Due to the vocabulary limitations, the description fails to encompass the bridge and its relationship with the biker, which is an essential element in understanding the falling action.

To mitigate the aforementioned limitations of VidVRD, we introduce **O**pen-vocabulary **V**ideo **R**elation **E**xtraction (OVRE). OVRE extracts all *action-centric* relation triplets from videos with language descriptions. Take Figure 2 (b) as an example: OVRE dissects the `falling` action into multiple relation components, such as <`cat`, `push`, `monitor`>and <`monitor`, `smashed on`, `man`>. Departing from VidVRD (**?**) that confined objects and relations to limited categories, we harness the immense potential of large Language Models (LLMs) to articulate action-aware relations using natural language. In essence, we undertake a two-fold process: first, using CLIP visual encoder to encapsulate video semantics; second, mapping these semantics into the linguistic realm for generating relation triplets through a pre-trained LLM. Employing this straightforward baseline, we achieve the ability to articulate relation triplets with unconstrained vocabularies.

To study OVRE, we present Moments-OVRE, an extensive dataset encompassing over 180K diverse videos. These videos are sourced from a subset of Multi-Moments in Time (**?**)(M-MiT), a multi-label video dataset with an average video duration of three seconds. In Moments-OVRE, we annotate both actors and relations as relation triplets that are most relevant to the action labels without vocabulary limitations. Note that "actor" does not refer only to human individuals but can also refer to objects, animals, or any entities that are actively involved in an action. Moments-OVRE confers several advantages: (I) Unrestricted Vocabulary Annotations: Moments-OVRE benefits from an expansive vocabulary encompassing diverse actors and relations, offering a more accurate portrayal of real-world scenarios. (II) Emphasis on Action-Centric Annotations: Our annotations exclusively focus on relation triplets related to actions/events depicted in the video. (III) Ample Dataset Scale: Moments-OVRE contains over 180K videos, establishing itself as the most extensive video relation extraction dataset known to date.

The main contributions are as follows:

• We introduce the novel Open-vocabulary Video Relation Extraction task, aiming to extract action-centric relations within videos.

• To foster ongoing research, we curate Moments-OVRE, an expansive dataset for video action comprehension. Comprising over 180k videos, Moments-OVRE is enriched with comprehensive open-vocabulary relationship annotations.

• We build a simple pipeline designed for generating action-centric relation triplets from raw videos. Extensive experimentation is conducted to validate assorted design considerations employed within the pipeline.

Related Works

Video understanding has long been an active research area in computer vision, encompassing a wide range of diverse aspects. Action recognition (**???**) categorizes the video content into a predefined set of action classes, temporal action localization (**??**) further emphasizes the specific time points at which these actions occur. Tasks such as video captioning (**?**), video question answering (**?**), and video-text retrieval (**?**) are dedicated to learning a detailed representation of semantic information. In the spectrum of these tasks resides an intermediary domain referred to as relation understanding, dedicated to capturing the diverse relations among objects. This contextual domain provides semantically rich intermediate representations for videos, crucial for comprehending interactions and behaviors depicted within them. VidVRD (**?**) and VidOR (**?**) are representative benchmarks for relation understanding, with finite object and predicate sets. Recently proposed task Open-VidVRD (**?**) manually splits the finite object and predicate sets into base and novel categories and builds the novel prediction setting. However, these existing VRD tasks focus on object-centric relation detection and struggle to bridge the gap between closed-label sets and natural languages, prompting us to propose video relationship extraction under open-vocabulary scenarios. Moreover, we aim to identify informative triplets that portray various actions in the video.

Contextual-level Visual Comprehension has been mainly explored in the image domain, including scene graph generation (**?**), visual semantic role labeling (**??**), and human-object interaction (**??**). These tasks have also been extended to the video domain, which requires information aggregation across frames. Existing works often treat video relation detection as a multi-stage pipeline, including object tracking, tracklet pair generation, and relation classification. These works focus on improving relation classification by leveraging contextual knowledge (**?**) or adding additional modules to enhance predicate representation (**?**) with generated tracklets. Recent trends lean towards one-stage models like VRDFormer (**?**), employing various queries to integrate spatio-temporal information, facilitating tracklet pair

Table 1: Representative video understanding datasets.

Dataset	Task	Annotation	#Video	#Subject/Object	#Predicate	Avg. Time (s)	Source
Kinetics400	Action Classification	Action Categories	306,245	-	-	10	YouTube
VideoLT			256,218	-	-	192	YouTube
MSR-VTT	Video Captioning / Retrieval	Textual Descriptions	7,180	-	-	20	Web
VATEX			41,269	-	-	10	YouTube
VidVRD	Video Visual Relation Detection	BBoxes, Relation Triplets	1,000	35	132	33	ILVSRC2016-VID
VidOR			10,000	80	50	35	YFCC100M
Action Genome			10,000	35	25	30	Charades
VidSitu	Video Semantic Role Labeling	Verbs, SRLs, Event Relations	29,220	Open	Open	10	Condensed-Movies
Moments-OVRE	**Open-Vocabulary Video Relation Extraction**	**Relation Triplets**	**186,943**	**Open**	**Open**	**3**	**M-MiT**

generation and relation classification concurrently. However, their scalability and generalization are hampered by datasets containing a restricted size of objects and relationships. Video captioning, aimed at generating open-ended descriptions, typically uses an encoder-decoder architecture. Over time, this setup has advanced to more sophisticated models (**???**). Recent strides in visual-language representation learning have led to models pre-trained on extensive paired data demonstrating promise in fine-tuning for various multi-modal tasks (**???**). Building on the progress in video captioning, we seek to leverage generative models for our OVRE task.

Datasets for video understanding have been introduced in support of multiple tasks over the past years. Kinetics400 (**?**) is a widely adopted category-balanced classification dataset. VideoLT (**?**), in opposition, follows a naturally long-tailed distribution for action recognition. MSR-VTT (**?**) is a widely adopted dataset for video captioning and video retrieval, it includes video clips from different domains, and each of them is annotated with approximately 20 natural sentences. Recently introduced VATEX (**?**) is a large-scale multilingual video description dataset. VidVRD (**?**) and VidOR (**?**) densely annotate objects and the most basic relations. Action Genome (**?**) focuses on action understanding by decomposing them into different relationships and ignoring others that are irrelevant to actions. VidVRD datasets have restricted their objects and relations to a fixed label set and thus are far away from reflecting the diverse relations in the real world. VidSRL (**?**) presents the sole open-vocabulary relation understanding task, while it primarily focuses on events. In contrast, OVRE places greater emphasis on actions, which is a fundamental ability for event comprehension. Detailed statistics for these datasets are listed in Table 1.

Video Relation Extraction Benchmark

Task Formulation Given a raw video input V, OVRE targets inferring its corresponding visual relationships R. Here, $V \in \mathbb{R}^{T\times H\times W\times C}$ denotes its frame sequence, where T, H, W, and C represent the number of frames, width, height, and channels respectively. The relation triplets are represented as $R = \{r_1, \cdots, r_K\}$, and each relation r_K is structured as a triplet: <subject, relation, object>.

Given that actions depicted in videos can encompass a wide spectrum of interactions, and distinct interactions might be common across various actions, our objective is to

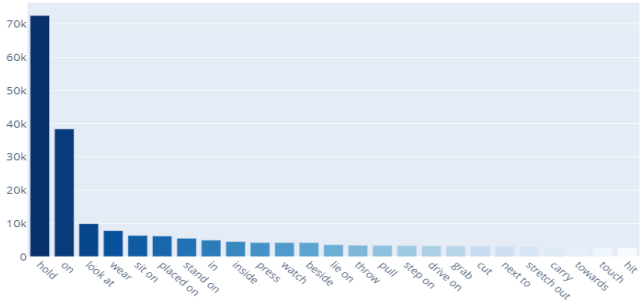

Figure 3: Counts of top 25 relationships in Moments-OVRE.

systematically extract all relationships associated with actions within the video context. Each r_i consists of a sequence of words $w = \{w_1, w_2, ..., w_N\}$. These individual r_i are then concatenated to form a token sequence $R = w_1, w_2, ..., w_M$, which serves as a representation of a sequence of relation triplets, collectively portraying all the action-centric relationships. The training objective is to maximize the likelihood of generating R given the video V:

$$\max_{\theta} \log p_\theta(R|V),$$

where θ denotes the model's trainable parameters. Our key idea is to leverage the powerful capability of generative models to address the OVRE problem. Consequently, the revised training objective can be formulated as follows:

$$\max_{\theta} \sum_{i=1}^{M} \log p_\theta(w_i|V, w_{i-1}).$$

OVRE involves generating relationships using nature language, rendering common metrics wildly used VidVRD metrics such as Precision@K and Recall@K unsuitable for our task. Inspired by the evaluation metrics in image captioning, we advocate for assessing the quality of generated triplets using Bleu, CIDEr, and METEOR scores. For the Bleu metric, we only consider B@1, B@2, and B@3, as each triplet contains at least three words. These metrics face challenges when applied directly to evaluating the overall relation triplets since the generated triplet sequence can be regarded as an unordered set, whose alignment with the ground truth triplets is unknown.

To match the generated triplet set with the ground truth, we first utilize a sentence embedding model SimCSE (**?**)

to quantify the text similarity between the generated and ground truth triplet set. Then we use these embeddings to acquire a similarity matrix S, where S_{ij} denotes the cosine similarity score of generated triplet $\hat{r}_i$ and ground truth triplet r_j. After obtaining S, we use the Hungarian algorithm to establish one-to-one matches between the triplets and apply the aforementioned metrics to these paired triplets. Note that the number of generated triplets may differ from that of ground truth triplets. When fewer triplets are generated, unmatched ground truth triplets receive zero scores.

Dataset Construction

We present the Moments-OVRE dataset tailored for OVRE. Notably, Moments-OVRE boasts distinct attributes such as diverse video content, an extensive compilation of video-triplet pairs, annotations embracing open vocabulary, and a focus on action-centric relationships. In this section, we will detail how we select representative videos, annotate relations, and devise data splits. Additionally, we thoroughly explore and analyze the statistical insights within the Moments-OVRE dataset.

Video Selection Since our task focuses on action-centric relationships, several requirements are posed for the selection of videos. First, we prefer videos with multiple actions. Second, videos should contain rich information, involving different scenes, objects, and events in the real world. We choose to annotate videos from Multi Moments in Time (M-MiT) (**?**) due to its multi-label nature, which allows a more nuanced comprehension of actions within intricate contexts. M-MiT offers several distinct advantages: (I) Unlike commonly utilized action datasets like Kinetics, M-MiT exhibits substantial intra-class diversity. This highlights the necessity of providing more intricate descriptions of relationships to capture variations between videos of the same action labels. (II) A majority of M-MiT videos have a duration of only 3 seconds. This temporal brevity encourages annotators to focus primarily on portraying the action itself, as opposed to the temporal context which is more prevalent in longer videos, such as those in YouCook and Howto100M datasets. Considering the long-tail distribution of action categories in M-MiT, we attempt to relatively balanced sampling videos from all classes. M-MiT has 292 action categories, we sample at least 660 videos per class and additionally select other random videos as some videos may be discarded during annotation.

Annotation Pipeline We ask annotators to perform an action-centric relation triplet annotation, which follows these steps: Given the potentially noisy nature of the large-scale M-MiT dataset, annotators first verify the correctness of the action labels. Videos with incorrect labeling are subsequently excluded. To perform action-centric annotation, both the video and the corresponding action labels are presented to the annotators. They first identify and annotate all pertinent objects, and then articulate the associated relationships among these objects. Annotators are instructed to provide descriptions for relation triplets solely when these relationships are relevant to the action labels. To illustrate, refer to Figure 1 (b): the annotation <`monitor`, `smashed on`, `man`>is required, whereas the annotation <`guitar`, `placed on`, `ground`>is invalid, as the latter constitutes background information rather than being directly tied to the `falling` action. Besides, we also manually review and correct the low-quality annotations.

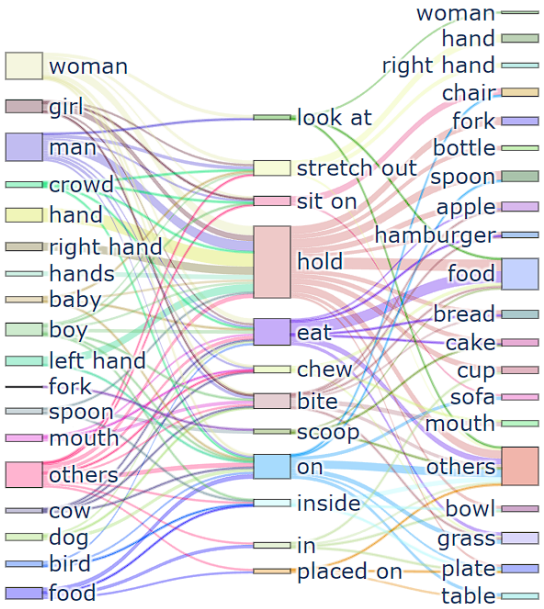

Figure 4: A weighted bipartite mapping of the top 12 most frequent relations in `eating` videos.

Dataset splits The data is partitioned into training and testing sets, resulting in 178,480 and 8,463 videos respectively. To ensure maximum consistency with the original splitting, all videos are selected exclusively from their designated sets within the M-MiT collection.

Dataset Analysis and Statistics

Overall, Moments-OVRE offers open-vocabulary annotations for 186,943 videos, encompassing a sum of 399,576 relation triplets. In Figure 3, we show the distribution of the top 25 most frequent relations in Moments-OVRE, which follows a natural long-tailed distribution. Furthermore, Figure 4 shows frequently described relation triplets of `eating` videos. Note that `hold` is a prevalent visual relation shared by various kinds of videos, such as `painting`, and `drinking` videos. A clue from this observation is that merely relying on a single relation is insufficient to infer the action categories from the videos, and a comprehensive understanding requires recognition of combinations of diverse relations.

Method

Overview An illustration of our method is provided in Figure 5. The overall framework mainly includes a Video Encoder, an Attentional Pooler, and a Text Decoder. The video is first transformed into a sequence of features via the video encoder. These features are further condensed into

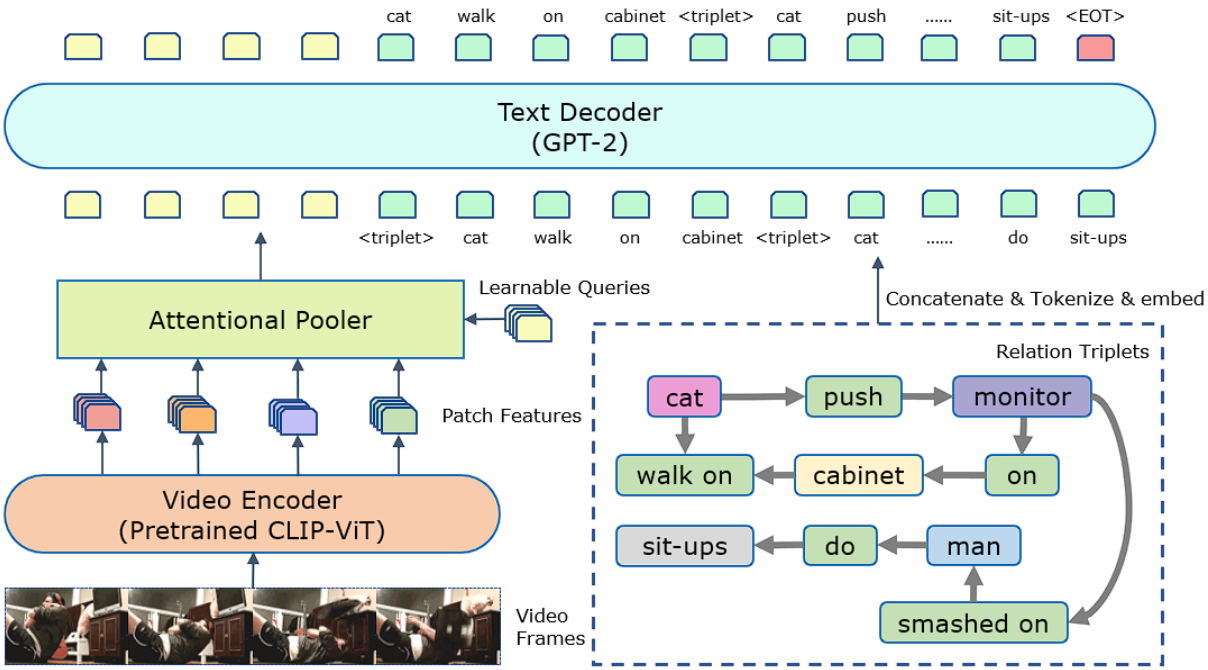

Figure 5: Overview of our model architecture, enabling simple relation generation with the powerful vision-language pre-trained model CLIP-ViT as the video encoder and the large language model GPT-2 as the text decoder.

prefix information by the attentional pooler, which is then fed into the text decoder for relation triplet generation.

Video Encoder To extract visual features from a video, we leverage the visual encoder of a pre-trained CLIP (**?**) model. Previous research has demonstrated its exceptional performance across various open-vocabulary recognition tasks (**???**). Our choice for visual feature extraction is CLIP-ViT, which utilizes ViT(**?**) to translate video frames into sequences of patches. These visual patches capture finer-grained information, making them an optimal fit for OVRE. However, the incorporation of the temporal dimension necessitates consideration of the resultant number of patches. Specifically, for a video $V \in \mathbb{R}^{T\times H\times W}$, this translates to $T \times \frac{HW}{P^2}$ patches.

Prefix Strategies With the features from V, our focus shifts towards transmitting them into the textual domain for relation generation. A simple solution is to directly feed all patches into the decoder, while the excessive number of patches not only leads to additional computational costs but also introduces a lot of redundant information. In light of this, we utilize an attentional pooler denoted as $F(\cdot)$, which employs a predetermined number of queries to extract meaningful features from all video patches: $q_1, ..., q_m = F(p_1, ..., p_n, q_1, ..., q_m)$, which aggregates spatial-temporal features into a more concise representation.

Attentional Pooler Drawing inspiration from the approach introduced in VideoCoCa (**?**), our framework incorporates an essential element known as the Attentional Pooler. This module takes learnable queries and patch features as its input, facilitating cross-attention mechanisms between them. The optimization of both the parameters of the Attentional Pooler and the queries enables the queries to gradually refine their ability to extract significant relationships from the patches. This design proves well-suited to the demands of OVRE. Remarkably, even with a single-layer transformer, our model showcases its ability to generate a diverse set of meaningful visual relationships.

Text Decoder We employ GPT-2 (**?**) as our generation model. Though relationships themselves are unordered, we observe that sorting them by Triplet Linearization (**?**) does not provide significant benefits in OVRE. Accordingly, all relation triplets of a video are simply concatenated into a text sequence by a separation token. Then, we map all the words into their corresponding tokens and pad them to the maximum length M to obtain a sequence of embeddings.

Experiments

Experiments Settings

Video and Text Preprocessing. For each input video, we first resize it to 224 × 224 and then extract 16 frames using uniform sampling, resulting in 786 patches for each video. We utilize RandAugment (**?**) as our data augmentation strategy. For the paired relation triplets, we concatenate the unordered relation triplets into a sequence using the special token `<triplet>`.

Training Setting. We train the generation model using cross-entropy loss and employ teacher forcing to accelerate the training process. All models are optimized using AdamW optimizer, with $\beta_1 = 0.9$, $\beta_2 = 0.999$, a batch size of 16, and weight decay of 1e-3. The initial learning rate is

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).

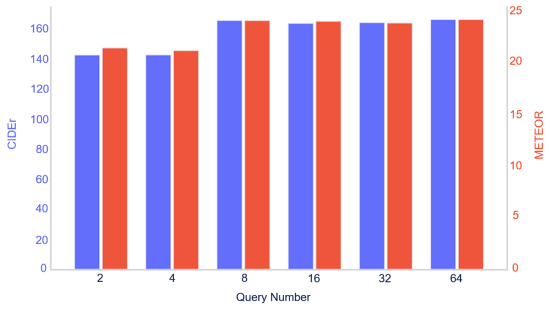

Figure 6: Impact of the query numbers on the performance. For each query number, we report CIDEr (blue) and METEOR scores (red) over the Moments-OVRE test set.

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).

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