Supplementary Material for
VRDFormer: End-to-End Video Visual Relation Detection with Transformers

Sipeng Zheng
Renmin University of China
zhengsipeng@ruc.edu.cn

Shizhe Chen
Inria
shizhe.chen@inria.fr

Qin Jin*
Renmin University of China
qjin@ruc.edu.cn

Section A presents more detailed descriptions of the inference procedure; Section B provides implementation details of the VRDFormer; Section C presents additional ablation experiments on the large-scaled VidOR dataset. Finally, Section D shows more qualitative examples.

A. Inference Details

Given a tracklet pair in the memory, we create an interactiveness curve for each relation class (as illustrated in Figure 3(a)), which reflects the probability of a certain relation class in the tracklet pair over time. Therefore, we create a total of N_{rel} such curves given a tracklet pair, where N_{rel} denotes the number of relation classes in the dataset. We then generate relation instances according to these N_{rel} curves for each tracklet pair similar to [3]. To be specific, we slice the interactiveness curve into different temporal regions based on a threshold β (as illustracted in Figure 3(b)), and β is uniformly sampled from $(0.3, 0.7)$ with a step of 0.05. In this way, we obtain the valid relation temporal regions in each threshold interval (the blue shaded area). Next, we merge the valid temporal regions in each threshold interval to generate relation instances. Assuming the sequence of valid temporal regions in a specific threshold interval is $\{l_1, l_2, ...l_i, ...\}$. We keep merging two adjacent temporal regions l_i and l_{i+1} until the ratio of the total valid duration to the total merged duration is below a certain threshold η. Each merged temporal region (green box as illustrated in Figure 3(c)) represents a relation instance proposal. We then use Non-Maximum Suppression (NMS) to filter out highly overlapped proposals (as illustrated in Figure 3(d)). Finally, we generate relation instances from each interactiveness curve for each tracklet pair in the video, and select top K instances according to the product of subject, relation and object probability, $P = P^s * P^r * P^o$.

B. Implementation Details

We augment the video frames by random cropping, random horizontal-flip and random resizing, where the maxi-

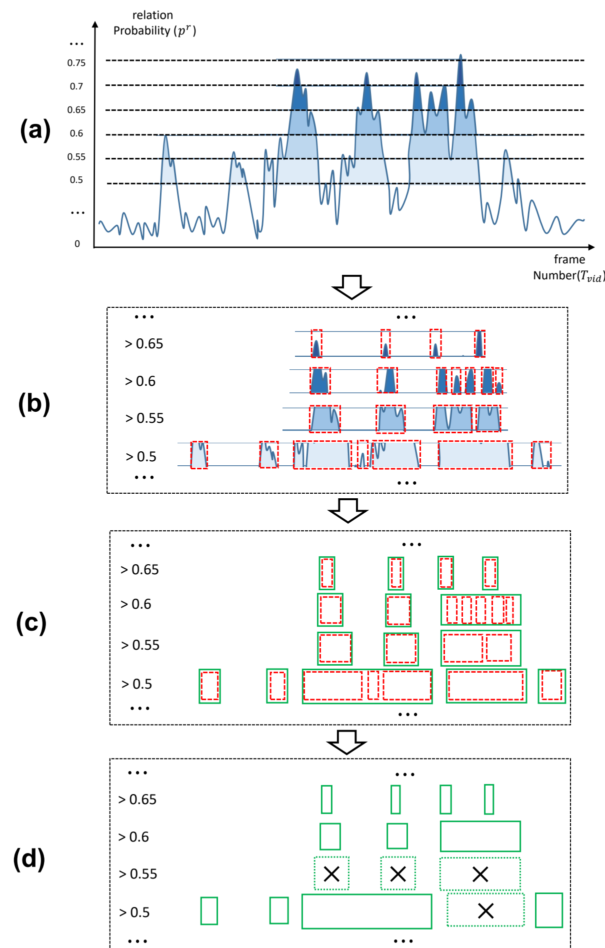

Figure 1. Illustration of our inference procedure: **(a)** Given a tracklet pair, a unique interactiveness curve for each relation class is created. **(b)** We use different threshold β to slice the interactiveness curve in order to generate potential valid temporal regions. **(c)** Our model merges adjacent valid temporal regions into relation instance proposal. **(d)** We use Non-Maximum Suppression (NMS) to filter out highly overlapped relation instance proposals. Relation instances marked by "×" are filtered out.

*Qin Jin is the corresponding author.

Table 1. Ablations of recurrent queries and re-activate strategy on VidOR (Q1).

	Re-Activate	Recurrent Query	Relation Detection		Tracklet Pair Detection	
			mAP	R@50	R@50	R@100
1	×	×	9.21	9.08	17.86	20.38
2	×	✓	10.46	10.25	19.12	21.94
3	✓	✓	11.19	11.05	19.73	23.58

Table 2. Ablations of joint training of object detection and relation classification on VidOR (Q2).

	joint train	Relation Detection		Tracklet Pair Detection	
		mAP	R@50	R@50	R@100
1	×	10.38	10.21	18.32	21.95
2	✓	11.19	11.05	19.73	23.58

mum size of each frame does not exceed 1280 pixels. We also augment the data by randomly sampling negative tracklet pairs related to "no-interaction" during training. The ratio between positive and negative samples are set as 1: 1.5. In addition, each self-attention layer [2] in the transformer contains 8 attention heads. We apply the deformable attention layer [4] for cross attention and set the total sampled key point number as 4. The subject and object bounding box MLP heads have 3 linear layers with ReLU activation, while the subject class, object class, interactiveness and relation heads only have 1 linear layer. The scaling factors of μ_{box}, μ_{cls} and μ_{intr} are set as 3.5, 1, 1. Our method uses an interactive threshold θ_{intr} to filter out negative tracklet pair proposals. To avoid the undesirable bias that one score of the subject, relation or object is significantly smaller than the other two, a tracklet pair proposal is considered as positive only when all of them are larger than 0.3. During training, we jointly train the model with Task I and Task II. In implementation, we use one mini-batch to train Task I and then another mini-batch to train Task II, which is the so-called 'alternately training' in the main paper. For the tagging task, as the groundtruth tracklets are provided, we use the groundtruth in training instead of predictions in Task I.

C. Additional Ablation Study on VidOR

In addition to the ablation experiments on the ImageNet-VidVRD dataset presented in the main paper, we carry out additional ablation study on the large-scaled VidOR dataset as well, which contains more dynamic and complex scenes of relation instances compared to ImageNet-VidVRD. Experiments from Table 1 to Table 5 corresponds to the same questions Q1 to Q6 in the main paper. We reach similar conclusions on VidOR.

Table 3. Ablations of different number of queries, where N_q denotes the number of static queries on VidOR (Q3).

	N_q	Relation Detection		Tracklet Pair Detection	
		mAP	R@50	R@50	R@100
1	20	7.24	8.60	14.95	16.42
2	50	9.63	9.28	16.58	18.72
3	100	11.19	11.05	19.73	23.58
4	200	10.71	10.56	19.28	22.94
5	300	10.45	10.27	18.45	21.82

Table 4. Ablations of different strategies to aggregate temporal contexts for relation tracklets on VidOR (Q4).

	Aggregation	Relation Detection		Relation Tagging	
		mAP	R@50	P@1	P@5
1	Mean	10.68	10.56	59.92	46.68
2	LSTM	10.82	10.72	60.69	47.42
3	Self Att	11.19	11.05	63.71	51.07

Table 5. Ablations of transformer components on VidOR, where "Cross" and "Self" denote cross- and self-attention in transformer decoder (Q6).

	Dec		Relation Detection		Relation Tagging	
	Cross	Self	mAP	R@50	P@1	P@5
1	×	✓	8.57	8.75	54.08	42.85
2	✓	×	7.85	8.22	52.32	41.37
3	✓	✓	11.19	11.05	63.71	51.07

Table 6. Ablations of different length for temporal aggregation on VidOR (Q5).

	T length	Relation Detection		Relation Tagging	
		mAP	R@50	P@1	P@5
1	1	10.26	10.21	58.12	44 85
2	4	10.42	10.37	58.85	45.36
3	8	10.61	10.51	59.48	45.96
4	32	11.19	11.05	63.71	51.07

D. Additional Qualitative Examples

In Figure 2, we illustrate the impact of spatio-temporal contexts for object localization. It shows that the object localization can be improved by using spatio-temporal contextualized information. For example, our model successfully detects the occluded bicycles (Figure 2(a)) according to their locations in previous frames. Meanwhile, our model is able to localize some challenging objects such as skateboard (Figure 2(b)) through spatial contexts, such as the adult on it. However, the VidVRD baseline [1] which relies on isolated object detection fails to detect relations such as child-ride-bicycle (Figure 2(a)) or adult-above-skateboard (Figure 2(b)).

Figure 2. Visualization of the spatio-temporal contexts for object localization: **(a)** temporal contexts help to localize the bicycles in the last frame; **(b)** spatial contexts enable our model to detect the bounding box of skateboard given the adult on it.

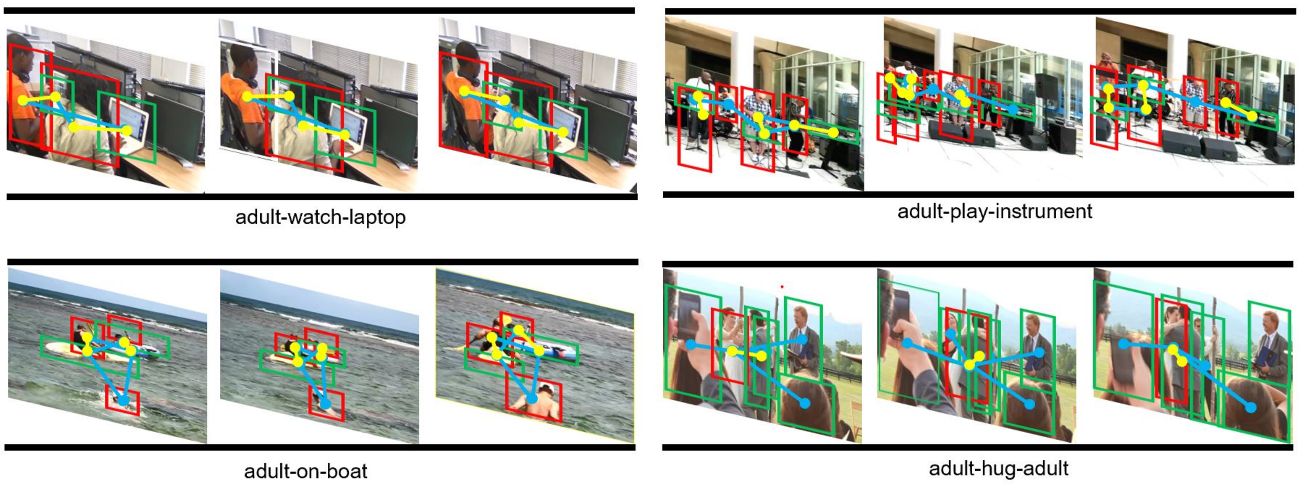

Figure 3. Visualization of our query-based relation instance generation, where red and green denote the subject and object respectively. Our model captures semantically meaningful relation instances denoted by yellow lines and filters out negative proposals denoted by blue lines in complex scenes at the same time.

Figure 3 visualizes the effects of our query-based relation instance generation. Our model is capable of capturing positive relation instances from noisy negative proposals even in complex scenes, such as the multi-person scenes.

References

[1] Xindi Shang, Tongwei Ren, Jingfan Guo, Hanwang Zhang, and Tat-Seng Chua. Video visual relation detection. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1300–1308, 2017. 2

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 2

[3] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2914–2923, 2017. 1

[4] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2