

Supplementary Materials for VrdONE: One-stage Video Visual Relation Detection

Anonymous Authors

1 OVERVIEW

This supplementary material serves to provide further explanations and results to complement the main manuscript. Initially, we discuss how to add additional visual features to our model in Section 2. Subsequently, in Section 3, we present more implementation details regarding how to train our model and how to detect video relations during inference. Further ablations are outlined in Section 4. Finally, we include extra visualization examples in Section 5.

2 ADDING ADDITIONAL FEATURES

Our VrdONE utilizes a pretrained object detector [2] to extract visual features $F = \{f_1, f_2, \dots, f_N\}$ for entities, together with their corresponding spatial positions $\Theta = \{\theta_1, \theta_2, \dots, \theta_N\}$, where N denotes the total entity count in a single video. For each subject-object pair, we process their features $(f_s, \theta_s, f_o, \theta_o)$ using the Bilateral Spatiotemporal Aggregation (BSA) Module to fully perceive spatiotemporal interactions and integrate the dual entities into unified features. Subsequently, a one-stage relation detector is applied to encode the features across multiple scales and decode them for both relation classification and temporal boundary localization simultaneously.

In TABLE 1 of the main manuscript, we augment our VrdONE model with additional visual features extracted from CLIP [5]. This process is done before the BSA, which merges visual features from different sources to enhance their representation. By cropping the original images based on θ_i , we employ CLIP image encoder to encode the cropped images into powerful visual feature representations f_i^c . A simple fusion method is applied by a multilayer perceptron (MLP), which is

$$f_i := \text{MLP}(\text{Concat}(f_i, f_i^c)), \quad (1)$$

where $\text{Concat}(\cdot, \cdot)$ is concatenation along feature dimensions, and $:=$ means the old ones will be replaced by the new ones. Fused features replace the original ones and undergo subsequent processing by the BSA and one-stage relation detector. Incorporating extra visual features demonstrates further improvement. More designs of feature integration are reserved for future exploration.

3 MORE IMPLEMENTATION DETAILS

In this section, we first introduce more training and inferencing details on how we conduct the experiments on VidOR and Imagenet-VidVRD.

3.1 VidOR

Parameter Settings. We initialize the feature dimension C as 512, which is then projected to 256 at the beginning of the decoder. The maximum length of overlapped subject-object durations l_{so} is set to 512. For those overlapped durations that are less than 512, we mask the extra parts to prevent attention calculation. The Multiscale Transformer Encoder incorporates 3 blocks, along with

the output from the Subject-Object Synergy (SOS) block, resulting in a 4-layer feature pyramid. With a downsampling ratio of 2, the feature pyramid comprises lengths of [512, 256, 128, 64] respectively. The decoder consists of 4 layers, with the number of queries N_q set to 9. All attention heads in both the encoder and decoder are 8. The dimension of the hidden layer in the Feedforward Network (FFN) following local attention calculation is $4 \times$ the dimension of the input features.

When calculating a Local Attention, the attention range for each token is restricted within a window size k_w , as well as the kernel size of the **Conv1D** layer before Local Attention. We set the window size k_w to 9. Since there are 2 SOS layers and 3 downsampling layers before decoding, the attention range of each token will be [9, 81, 512, 512, 512] after the layers, respectively. The attention range is restricted by the maximum token length. As query tokens are parallel to each other and lack strict sequential order, when computing local cross attention in the decoder, we set the kernel size of **Conv1D** for the queries to 1.

Training Details. We sample frames uniformly with a stride of 4. When the overlapped length between subject and object l_{so} exceeds 512, we randomly truncate it to 512. Relations occurring within the truncated l_{so} , with durations less than half of the original length, are excluded from model learning. Additionally, We omit the learning of some pairs with more relations than N_q . Employing a batch size of 48, we conduct training for 10 epochs. Before Local Attention and MLP computation, LayerNorm [1] is implemented. Drop-out and Drop-path [4] rates are specified as 0 and 0.1, respectively. Training of VrdONE employs the AdamW [21] optimizer with a learning rate of 2×10^{-4} , where the learning rate for each parameter group follows a cosine annealing schedule, eventually dropping to 2×10^{-5} . Warmup and Exponential Moving Average (EMA) techniques are employed to enhance and stabilize the training process. No positional encoding is added to the learnable query tokens, consistent with the approach in [7]. The parameters λ_{cls} , λ_{mf} , λ_{md} are assigned values of 2, 2, and 5, respectively.

Inference Details. During inference, we keep the entities with confidence scores > 0.4 and enumerate the remaining entities to compose all possible subject-object pairs. Pairs with no overlapping length are discarded. All pairs within a video are divided into two groups based on their lengths: those with lengths ≤ 512 and the left. For the former, the same processing steps as during training are applied. For the latter, the l_{so} is set to the maximum length among all pairs. Following prior studies [3, 6], we retain top-6 predicate classification results for each detected instance. For each video, we retain the top 200 results across all detected instances. Foreground probability > 0.5 in the localization masks indicates where relation instances happen. The temporal boundaries of the relation instances are indicated by the first positive masks and the last positive masks. Typically, no further processing of the localization masks is necessary, as we find that filling holes or

Table 1: Comparison of oracle detection results. The “-oracle” postfix means the model is trained with ground-truth detection trajectories and entities’ categories provided.

Method	Relation Detection						Relation Tagging					
	mAP	Δ mAP	R@50	Δ R@50	R@100	Δ R@100	P@1	Δ P@1	P@5	Δ P@5	P@10	Δ P@10
Imagenet-VidVRD												
VidVRD [6]	8.58		5.54		6.37		43.00		28.90		20.80	
VidVRD-oracle [6]	15.53	+6.95	12.51	+6.97	16.55	+10.18	43.50	+0.50	29.70	+0.80	23.20	+2.40
VrdONE	31.33		18.20		21.61		80.50		59.40		44.17	
VrdONE-oracle	43.15	+11.82	30.67	+12.47	38.30	+16.69	82.50	+2.00	62.10	+2.70	46.55	+2.38
VidOR												
VrdONE	11.86		11.13		14.21		66.11		54.92		43.90	
VrdONE-oracle	41.75	+29.89	35.93	+24.80	47.79	+33.58	85.10	+18.99	71.37	+16.45	58.43	+14.53

Table 2: Ablation of the window size k_w , whose values are varying within [5, 11].

k_w	Relation Detection			Relation Tagging		
	mAP	R@50	R@100	P@1	P@5	P@10
5	11.77	11.11	14.00	67.31	54.84	44.14
7	<u>11.83</u>	11.04	<u>14.14</u>	66.11	55.23	43.87
9*	11.86	11.13	14.21	66.11	54.92	43.90
11	11.86	<u>11.12</u>	14.05	<u>66.59</u>	54.75	43.42

Table 3: Ablation of the positional representations θ s.

θ^a	θ^r	Relation Detection			Relation Tagging		
		mAP	R@50	R@100	P@1	P@5	P@10
–	–	10.84	10.46	13.45	65.62	54.46	43.24
–	✓	11.45	10.75	13.85	<u>66.11</u>	54.36	<u>43.82</u>
✓	–	<u>11.83</u>	<u>11.06</u>	<u>13.98</u>	66.91	54.64	43.60
✓	✓	11.86	11.13	14.21	<u>66.11</u>	54.92	43.90

applying morphological methods (e.g. erosion or dilation) yields minimal improvements.

3.2 Imagenet-VidVRD

When shifts to Imagenet-VidVRD, some settings of the hyperparameters are slightly different. For Imagenet-VidVRD, we set l_{so} to 96, resulting in feature pyramid lengths of [96, 48, 24, 12]. The window size k_w is set to 7, with corresponding attention ranges of [7, 49, 96, 96, 96]. We employ a sampling stride of 1, a batch size of 24, and conduct training for 12 epochs. During inference, we retain the top-8 predicate classification results for each detected instance.

4 ADDITIONAL EXPERIMENTS

In this section, we provide more experimental evidence for exploring the potential of our framework under the limited quality of pretrained trackers. Extra ablations are also conducted to verify the influences of the varying kernel size k_w , positional representations θ^a and θ^r , and the weight factors λ_{cls} , λ_{mf} , and λ_{md} using in Eq. 19.

Tracklets. Before proceeding to our VrdONE, the input videos are pre-processed by pretrained object detector, which results in lists of detection bounding boxes for each entity. These bounding boxes are further used for consolidating precise spatial context information into the feature of entity tracklets.

We empirically find that the accuracy of the detection results significantly affects the upper bound of our model’s performance. Therefore, we propose to provide the ground-truth bounding boxes and the corresponding class label of the entity trajectories to our model. As shown in Table 1, after perceiving accurate tracklets, our model has witnessed a giant progress on both Imagenet-VidVRD and VidOR. The performances of our VrdONE on both relation detection and temporal localization are significantly boosted. In the Imagenet-VidVRD dataset, the gains of our model on 5 out of the 6 metrics surpass VidVRD [6] significantly. A similar trend is also observed in the VidOR dataset, with tremendous improvements of 29.89 on mAP and 24.8 on R@50.

The averaging temporal length of videos in VidOR is much longer than those in Imagenet-VidVRD, which incurs a more difficult detection for trackers, leading to worse performance. However, even in such a challenging scenario, our model can still perform precise relation classification and temporal localization assisted by accurate tracklets. These experimental results imply the potential of our proposed VrdONE facilitated by the advancement of modern object trackers.

Window Size of Local Attention. In Table 2, we estimate the influences of varying the window size k_w . In all our experiments, we select k_w as 9 for more balanced performances on RelDet and RelTag.

Positional Representation. In our spatiotemporal synergistic learning, incorporating the positional representations θ^a and θ^r can help the model perceive the spatial variances of relations between independent entities. We illustrate the results of ablation on the absolute positional representations θ^a and the relative positional representations θ^r in Table 3. Both θ^a and θ^r are crucial in promoting the detecting ability, while θ^a exhibits more progress. These results highlight the influence of temporal changes of positions for relation classification, demonstrating the rationality and necessity of our spatiotemporal learning in video relation detection.

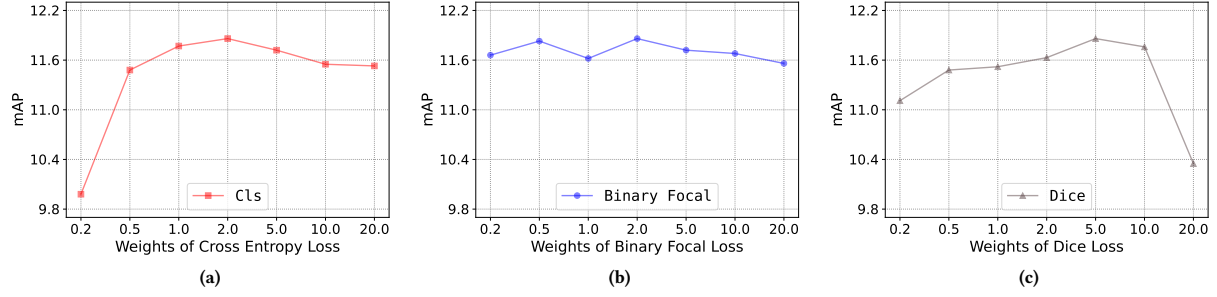


Figure 1: Ablations on varying weight factors of λ_{cls} , λ_{mf} , and λ_{md} , which are applied to (a) Cross Entropy Loss for relation classification, (b) Mask Focal Loss and (c) Dice Loss for relation localization, respectively.

Loss Factors. In all our experiments, we set the λ_{cls} , λ_{mf} , and λ_{md} as 2, 2, and 5 respectively based on the performance on mAP. In this section, we ablate the loss factors of λ_{cls} , λ_{mf} , and λ_{md} by varying one of them and setting the rest two frozen.

The results are depicted in Fig. 1. The curves validate that the performance of our model remains robust when the loss factors vary within a reasonable range. Otherwise, either λ_{cls} is too small or λ_{dice} is too large will lead to drastic drops in performance.

5 VISUALIZATION

Additional qualitative results are illustrated in Fig. 2 and Fig. 3. We show the spatiotemporal form of the detection results on both Imagenet-VidVRD (Fig. 2(a)) and VidOR (Fig. 2(b)) datasets. We find that our method can accurately learn the categories of relations, especially those related to spatial positions, demonstrating the effectiveness of our spatiotemporal learning. Moreover, it is obvious in this form that our method is hindered heavily by the tracker. For example, in Fig. 2(a), our method accurately finds that the **preson** is ridding the **bicycle** in the whole tracklets. However, the overlapped trajectories are only within [1, 163], decreasing the accuracy of temporal boundary predictions. So as to the “**sofa-beneath-child**” int Fig. 2(b).

In Fig. 3, we illustrate more classification results of our model. Even though the number of queries of each pair is limited, the number of accumulated detection results is sufficient for the whole video.

REFERENCES

- [1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).
- [2] Yihong Chen, Yue Cao, Han Hu, and Liwei Wang. 2020. Memory enhanced global-local aggregation for video object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10337–10346.
- [3] Kaifeng Gao, Long Chen, Yulei Niu, Jian Shao, and Jun Xiao. 2022. Classification-then-grounding: Reformulating video scene graphs as temporal bipartite graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19497–19506.
- [4] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. 2016. Fractalnet: Ultra-deep neural networks without residuals. *arXiv preprint arXiv:1605.07648* (2016).
- [5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

- [6] Xindi Shang, Tongwei Ren, Jingfan Guo, Hanwang Zhang, and Tat-Seng Chua. 2017. Video visual relation detection. In *Proceedings of the 25th ACM international conference on Multimedia*. 1300–1308.
- [7] Chen-Lin Zhang, Jianxin Wu, and Yin Li. 2022. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*. Springer, 492–510.

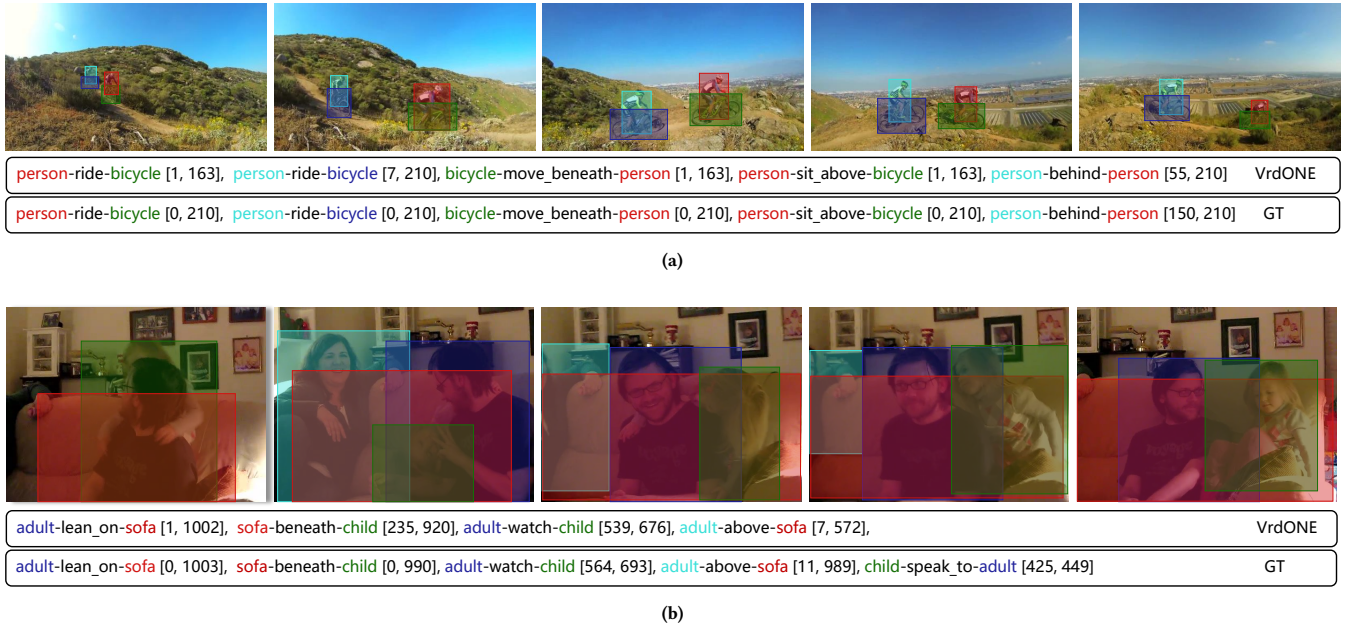


Figure 2: Additional qualitative results on Imagenet-VidVRD (above) and VidOR (below). The results are visualized in spatiotemporal forms to include all elements of video relations. (E.g., subject class, object class, subject trajectory, object trajectory, predicate class, predicate temporal boundary.) The numbers in the brackets are the start frames and end frames of current relations.

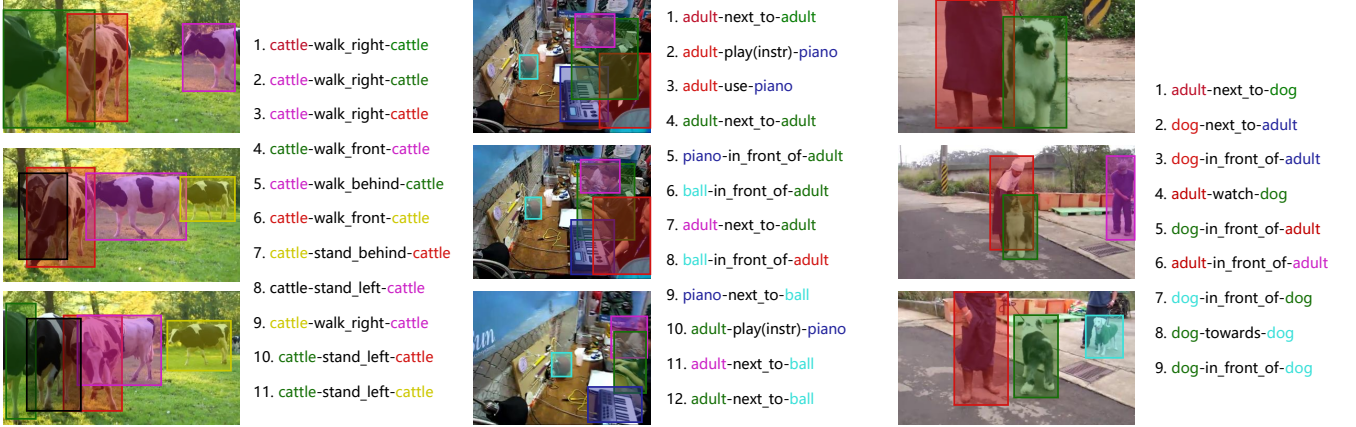


Figure 3: Visualizations of the relation detection results in more details.