

Supplementary Materials: Open-vocabulary Video Scene Graph Generation via Union-aware Semantic Alignment

Anonymous Authors

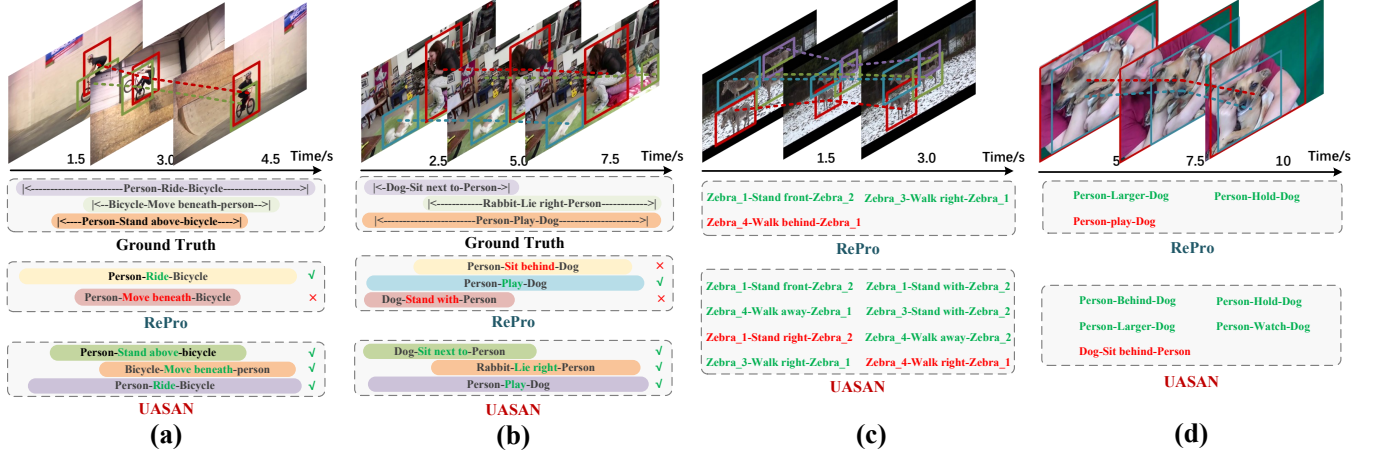


Figure 1: Qualitative results of our model and RePro [2]. We compared our proposed UASAN framework with RePro on Relation Detection task (RelDet) and Relation Tagging task (RelTag). Specifically, The results on RelDet task are in (a) and (b), while the results on RelTag task are in (c) and (d).

1 CODE AND PRE-TRAINED MODEL

The code and the pre-trained model can be found in the folder./code, and they can be used to reproduce our experimental results. Please refer to README for more details.

2 IMPLEMENTATION DETAILS

We utilize the same object trajectory data as [2]. Specifically, a Fast-RCNN [8]-based VinVL model [11] is first employed to detect objects with bounding boxes for each video frame. Then, Seq-NMS algorithm is utilized to generate class-agnostic object trajectories. We use a pre-trained ViT model [1] for trajectory feature extraction, and our bridge encoder is established based on a pre-trained Q-Former backbone [6]. Following [7, 9, 10], we generate visual relation triplets in short video segments, and merge the same relations with greedy relation association algorithm proposed by [10] during model inference. For VidVRD dataset, the base split have 25 object categories and 71 predicate categories, while the novel split have 10 object categories and 61 predicate categories. For VidOR dataset, the base split consists of 50 object categories and 30 predicate categories, while the novel split contains 30 object categories and 20 predicate categories. Please refer to [2] for detail base- and novel- splits. The hidden size d in our model is set to 512 and the length L of the extracted features is set to 32. L_{mot} is set to 2 and L_{rel} is set to 10. We use the Adam optimizer [4] to train our model. The learning rate is set to 10^{-4} for VidVRD and 5×10^{-5} for VidOR. The batch size is set to 8 for VidVRD and 4 for VidOR, and our model is trained 50 epochs on both VidVRD dataset and VidOR dataset. All our experiments are implemented in the PyTorch toolkit, and one NVIDIA GeForce RTX 3090 GPU is used.

3 PERFORMANCE COMPARISON

More experimental results compared with SOTA methods on VidOR dataset are shown in Table 1. When compared with ALPro [5] on all-split, our proposed UASAN outperforms it by gains of (9.20%, 12.00%) on R@50 and R@100 on SGCLs task, and UASAN also achieves margin improvements on PredCLs task. In addition, UASAN also outperforms VidVRD-II [9] with improvements of (0.87%, 1.45%) and (1.73%, 3.76%) on SGCLs and PredCLs tasks on novel-split, respectively. When compared with RePro [2], our model achieves improvements of (0.30%, 1.16%) on R@50 and R@100 metrics on SGCLs and 0.30% on R@100 metric on PredCLs on novel-split. Moreover, UASAN also surpasses RePro with an average of 0.52% on all-split. Note that RePro is trained with detected object trajectory annotations and manual annotations on VidOR dataset, while we only train our UASAN with a small amount of manually annotated object trajectories on VidOR dataset due to the incomplete released trajectory data of [2]. We can observe that our proposed method still maintains improvements in performance on most metrics. It demonstrates that our proposed method maintains the ability to recognize relations when confronted with challenges brought by more complex scenarios (VidOR dataset containing tens of times more data than VidVRD dataset), and it achieves improvements on almost all metrics on both novel-split and all-split when compared with SOTA methods.

4 QUALITATIVE RESULTS

Several visualization examples of our proposed UASAN framework are illustrated in Figure 1. Specifically, we evaluate the performance of UASAN and RePro [2] on both Relation Detection task (RelDet)

Table 1: Comparison with state-of-the-art methods on VidOR dataset.

| Models | Novel-split | | | | All-split | | | |
|-----------|--------------|--------------|--------------|--------------|---------------|---------------|---------------|---------------|
| | SGCls | | PredCls | | SGCls | | PredCls | |
| | R@50 | R@100 | R@50 | R@100 | R@50 | R@100 | R@50 | R@100 |
| ALPro | 3.17% | 3.74% | 8.35% | 9.79% | 0.95% | 1.32% | 2.61% | 3.66% |
| VidVRD-II | 1.44% | 2.01% | 4.32% | 4.89% | 9.40% | 12.78% | 24.81% | 34.11% |
| RePro | 2.01% | 2.30% | 7.20% | 8.35% | 10.03% | 12.91% | 27.11% | 35.76% |
| Ours | 2.31% | 3.46% | 6.05% | 8.65% | 10.15% | 13.32% | 27.36% | 37.06% |

and Relation Tagging task (RelTag) on VidVRD dataset. In RelDet task, both the precision of the predicted relation triplets and the localization of subject/object trajectories are considered, where the subject/object trajectories have sufficient voluminal Intersection over Union (vIoU) with those in ground-truth and the predicted triplet is the same as ground-truth. In RelTag task, the predicted result is correct if only the triplet to be the same as ground-truth. Figure 1(a),(b) shows the results on RelDet while Figure 1(c),(d) shows the results on RelTag. When evaluated on RelDet task, we can observe that our UASAN can clearly distinguish novel relation predicates from base ones, such as *stand above* and *move beneath* in Figure 1(a), while RePro fails to recognize novel predicate *stand above* and classifies it to a base category *move beneath*. In Figure 1(b), it is also clear that UASAN has the ability to predict novel predicates (i.e., *sit next to*) correctly, while RePro fails. Moreover, when evaluated on RelTag task, our proposed framework predicts more precise relation triplets with both base and novel predicates. As shown in Figure 1(c), UASAN can clearly recognize the objects of the same class and understand the relations between them. UASAN also predicts proper but rare relation predicate in a common subject-object pair, such as the *hold* for *person-hold-dog* in Figure 1(d), whereas RePro is easy to ignore such a triplet combination. The visualization results demonstrate the effectiveness of our proposed UASAN framework to predict relation predictions with both novel and base categories.

REFERENCES

- [1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xi-aohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
- [2] Kaifeng Gao, Long Chen, Hanwang Zhang, Jun Xiao, and Qianru Sun. 2023. Compositional prompt tuning with motion cues for open-vocabulary video relation detection. *arXiv preprint arXiv:2302.00268* (2023).
- [3] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).
- [4] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [5] Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven CH Hoi. 2022. Align and prompt: Video-and-language pre-training with entity prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4953–4963.
- [6] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*. PMLR, 19730–19742.
- [7] Xufeng Qian, Yueting Zhuang, Yimeng Li, Shaoning Xiao, Shiliang Pu, and Jun Xiao. 2019. Video relation detection with spatio-temporal graph. In *Proceedings of the 27th ACM International Conference on Multimedia*. 84–93.
- [8] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* 28 (2015).
- [9] Xindi Shang, Yicong Li, Junbin Xiao, Wei Ji, and Tat-Seng Chua. 2021. Video visual relation detection via iterative inference. In *Proceedings of the 29th ACM international conference on Multimedia*. 3654–3663.
- [10] 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.
- [11] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Making visual representations matter in vision-language models. *arXiv preprint arXiv:2101.00529* 1, 6 (2021), 8.