Spatial-Temporal Transformer for Dynamic Scene Graph Generation Supplementary Material

Yuren Cong[1], Wentong Liao[1], Hanno Ackermann[1], Bodo Rosenhahn[1], Michael Ying Yang[2]
[1]TNT, Leibniz University Hannover, [2]SUG, University of Twente

Appendix

In this supplementary material, we provide additional implementation details for our method in Sec. 1 of this appendix. In Sec. 2, we present detailed analysis of the Action Genome dataset [1]. In Sec. 3, we show additional qualitative results. Failure cases of our method are shown in Sec. 4.

1. Implementation Details

In this section, we present some implementation details that were omitted in the main paper for brevity.

Box Function f_{box} It transforms the bounding boxes of the subject and object to the $256 \cdot 7 \cdot 7$ feature map. Following [6], the bounding boxes of the subject and object are firstly converted to a binary spatial mask of size $2 \cdot 27 \cdot 27$ which indicates the location of the subject and object in the frame. By forwarding the spatial mask into a convolutional network (see Fig. 1), the location representation is computed which can be added to the $256 \cdot 7 \cdot 7$ feature map of the union box.

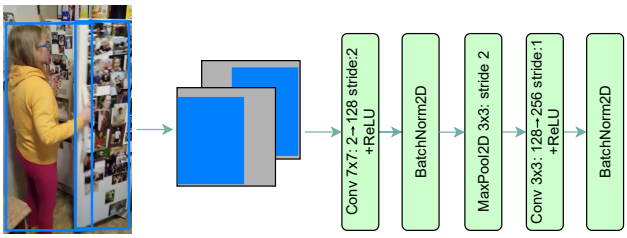

Figure 1: Illustration of the box function f_{box}

Queries and Keys in the Temporal Decoder For the i-th batch in the decoder layers, the queries $\boldsymbol{Q}$ and keys $\boldsymbol{K}$ are computed by adding the learned frame encoding $\boldsymbol{E}_f = [\boldsymbol{e}_1, \ldots, \boldsymbol{e}_\eta]$ to $\boldsymbol{Z}_i = [\boldsymbol{X}_i, \ldots, \boldsymbol{X}_{i+\eta-1}]$. Note that $\boldsymbol{E}_f$ and $\boldsymbol{Z}_i$ have the same length. $\boldsymbol{X}_i = \{\boldsymbol{x}_i^1, \ldots, \boldsymbol{x}_i^{K(i)}\}$ denotes all the relationship representations in the i-th frame. Here we use braces to emphasize that there is no order between relationships in the same frame and $\boldsymbol{X}_i$ is still a matrix (tensor) in our PyTorch code. Therefore, the first elements of $\boldsymbol{Q}$ and $\boldsymbol{K}$ can be formulated as:

$$\boldsymbol{q}_1 = \boldsymbol{k}_1 = \boldsymbol{e}_1 + \boldsymbol{X}_i = [\boldsymbol{x}_i^1 + \boldsymbol{e}_1, \ldots, \boldsymbol{x}_i^{K(i)} + \boldsymbol{e}_1] \quad (1)$$

which means the same encoding is added to the relation representations in the same frame.

Object Classification FasterRCNN [5] based on ResNet101 outputs a `2048-d` feature vector and a class distribution for each object proposal box. With multiplying the class distribution by the linear matrix $W_e \in \mathbb{R}^{36\times200}$, a `200-d` semantic embedding is computed. Meanwhile, the `4-d` box coordinate is forwarded into a feed-forward network (see Fig. 2) to achieve a `128-d` position embedding. We concatenate the feature vector, semantic embedding and position embedding, then project the concatenated vector to a `37-d` distribution (including the class *background*) with two linear layers and a ReLU function in between.

Data Pre-processing When performing down-sampling in the backbone, the visual information of ultra-small objects is damaged. In the experiments for SGCLS/SGDET, we only keep bounding boxes with short edges larger than 16 pixels as [2] did.

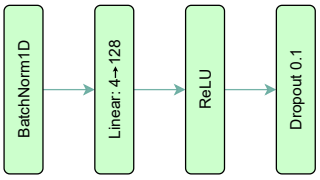

Figure 2: The box coordinate is forwarded into the feed-forward network to compute the position embedding.

2. Benchmark from Action Genome

In the Action Genome (AG) dataset [1], each human-object pair is annotated with three types of relationships, namely *attention*, *spatial*, and *contact* relationships where *attention* and *contact* relationships are formulated in the order of `<person-predicate-object>`, and *spatial* relationships are in the order of `<object-`

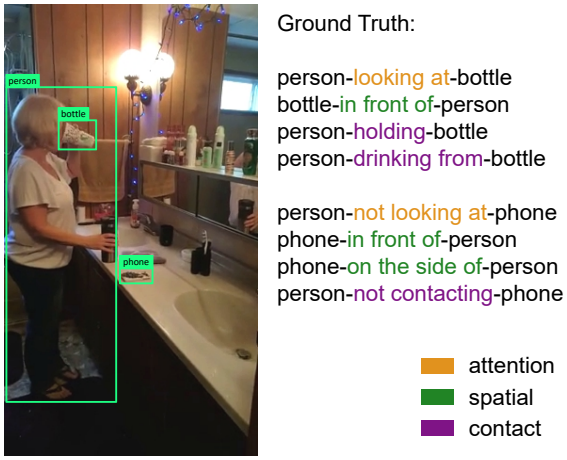

Figure 3: An example of the data annotation in Action Genome dataset.

predicate-person>. Note that the *spatial*, and *contact* relationships can be annotated with multiple labels in Action Genome dataset. An annotation example is shown in Fig. 3.

A benchmark following **With Constraint** is provided by [1]. However, their evaluation code and object detector have not been released. We also evaluate several advanced image-based models. Although the ranking of the model performances is consistent with [1] (VRD [4]<Motif Freq [6]<MSDN [2]<RelDN [7]), the values of $Recall@K$ are different. PredCLS-$R@K$ ($K = [10, 20, 50]$) computed by us are generally much higher, *e.g.*, PredCLS-$R@20$ from us $= 69.5$ whereas PredCLS-$R@20$ from [1] $= 49.4$ for RelDN [7]. The reason for the difference was found after discussing with the authors of [1]. Each person-object pair is allowed to have either an *attention* or *contact* relationship in [1]. Instead of, we allow each person-object to have:

- <person-attention relationship-object>
- <object-spatial relationship-person>
- <person-contact relationship-object>

for **With Constraint** so that $attention$ and $contact$ relationships can be detected simultaneously. Each human-object pair is allowed to have more than one $spatial$ or $contact$ relationship when the confidence score is higher than the threshold (0.9) following **Semi Constraint**. For **No Constraint**, the most confident top-K relationships are chosen no matter what kind of relationship. SGCLS/SGDET-$R@K$ ($K = [10, 20, 50]$) from [1] are slightly higher than ours. We argue that their object detector has a better performance which is crucial for SGCLS/SGDET. Note that person boxes in the ground truth are annotated by the detector from [1] in the present version of the Action Genome dataset.

Furthermore, there are two kinds of $Recall@K$ metrics in [1]: image-wise and video-wise. The video-wise $Recall@K$ is not adopted in our work because the only difference is whether the per-frame measurements in each video are first averaged.

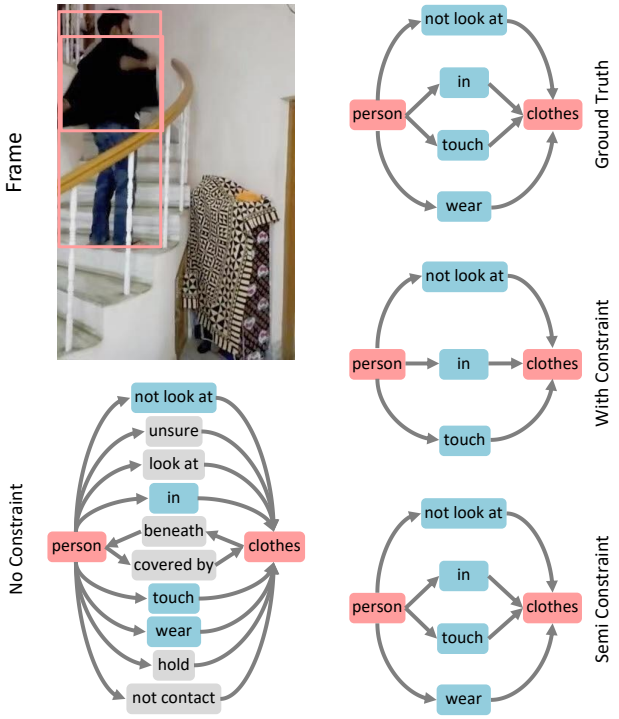

Figure 4: Qualitative instances in the PredCLS setting following different strategies. Light blue relationships are true positives predicted by STTran at the $R@10$ setting while gray are false positives. The graph from **Semi Constraint** is identical to the ground truth, whereas there are several false positives in the graph from **No Constraint** without restriction.

3. Additional Results

We also report the average precision of predicates AP_{pred} to evaluate the performance for single relationships. The AP_{pred} evaluates the average precision of the predicates where the subject and object boxes are given. The 10 most frequently occurring relationships in Action Genome dataset (2 $attention$, 4 $spatial$ and 4 $contact$ relationships) are evaluated with our model and GPS-Net [3], which performs best in the image-based scene graph generation methods. The results are shown in the Table 1. Compared with GPS-Net, our model has a great advantage in predicting $attention$ relationships with temporal dependencies and also performs better for $spatial$ relationships. However,

Method	AP_{pred}										
	not looking at	looking at	in front of	on the side of	beneath	hebind	holding	not contacting	touching	sitting on	*mean*
GPS-Net[3]	64.94	49.81	90.14	38.08	**88.98**	77.45	**88.38**	**81.30**	37.26	88.36	*70.47*
STTran	**79.73**	**67.07**	90.14	**40.52**	88.93	**81.01**	85.29	81.29	**37.50**	**90.67**	***74.22***

Table 1: The average precision of predicates AP_{pred} for the top-10 frequent relationships including 2 $attention$, 4 $spatial$ and 4 $contact$ relationships. We compare our model with GPS-Net [3] which performs best on the Action Genome dataset among the image-based baselines. With temporal dependencies, STTran has a great advantage in predicting $attention$ relationships and also performs better for $spatial$ relationships. For $contact$ relationships, GPS-Net outperforms STTran on the prediction of $holding$ and $not\ contacting$. The last column is the mean of AP_{pred} for these 10 relationships.

GPS-Net outperforms STTran on the prediction of $holding$ and $not\ contacting$ for $contact$ relationships.

Different performance of 3 generation strategies are demonstrated in Fig. 4. For **With Constraint**, $wearing$ is abandoned since only one contact relationship is allowed between each object pair. Although **No Constraint** allows multi-label prediction, the result contains a lot of noise when there are few pairs in the frame, especially bounding boxes are given in PredCLS and SGCLS.

Additional qualitative results for dynamic scene graph generation from the video are shown in Fig. 5. The dynamic scene graphs are generated with the top-10 confident predictions with different Strategies in the SGDET task. The green boxes denote the undetected truths. The melon and gray colors indicate true positive and false positive respectively. Correct relationships are colored with light blue whereas relationships not in the ground truth are colored with gray. In the video the person sitting on the bed holds the medicine and bottle. Then she takes the medicine and drinks water from the bottle.

4. Failure Cases

In order to clarify the limitation of the model, we analyze the results and summarize the following most common failure cases:

1. The object is not detected (IoU$<$ 0.5), particularly small objects such as $phone$ and $medicine$.
2. The predictions do not match the ground truth relationships which are annotated by mistake.
3. The relationship is ambiguous and difficult to be identified even by humans.
4. The model predicts the wrong majority relationship instead of the correct minority relationship.

The failure cases are shown in Fig. 6. We conjecture that Failure 1 can be improved by a better object detector. Failure 2 and Failure 3 are caused by the human-labeled annotations. Failure 4 is caused by the imbalanced relationship distribution both in the dataset and in the real world.

References

[1] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10236–10247, 2020. 1, 2

[2] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. Scene graph generation from objects, phrases and region captions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1261–1270, 2017. 1, 2

[3] Xin Lin, Changxing Ding, Jinquan Zeng, and Dacheng Tao. Gps-net: Graph property sensing network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3746–3753, 2020. 2, 3

[4] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European conference on computer vision*, pages 852–869. Springer, 2016. 2

[5] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015. 1

[6] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5831–5840, 2018. 1, 2

[7] Ji Zhang, Kevin J Shih, Ahmed Elgammal, Andrew Tao, and Bryan Catanzaro. Graphical contrastive losses for scene graph parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11535–11543, 2019. 2

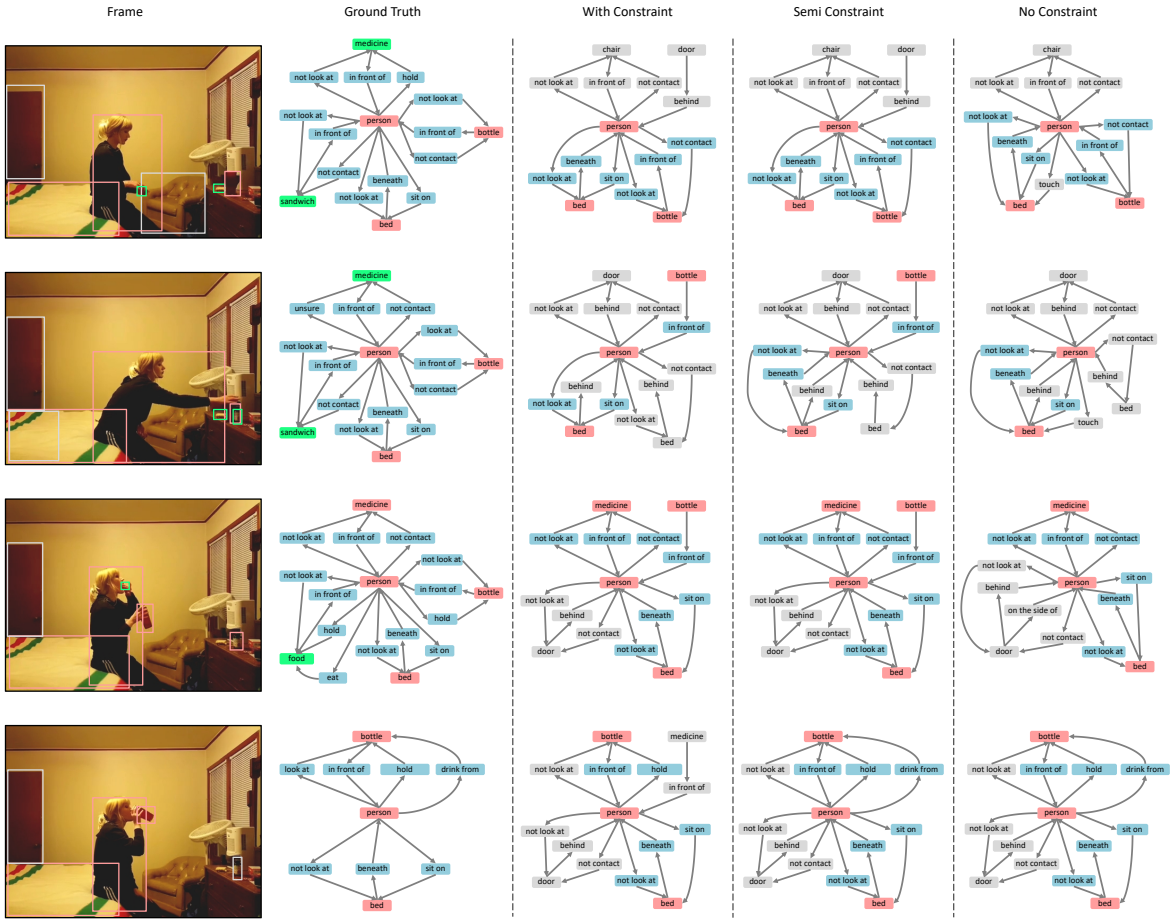

Figure 5: Qualitative results for dynamic scene graph generation. The scene graphs are generated with the top-10 confident predictions with different Strategies in the SGDET task. The green boxes denote the undetected ground truth. The melon and gray colors indicate true positive and false positive respectively. Correct relationships are colored with light blue whereas relationships not in the ground truth are colored with gray. In the video the person sitting on the bed holds the medicine and bottle. Then she takes the medicine and drinks water from the bottle.

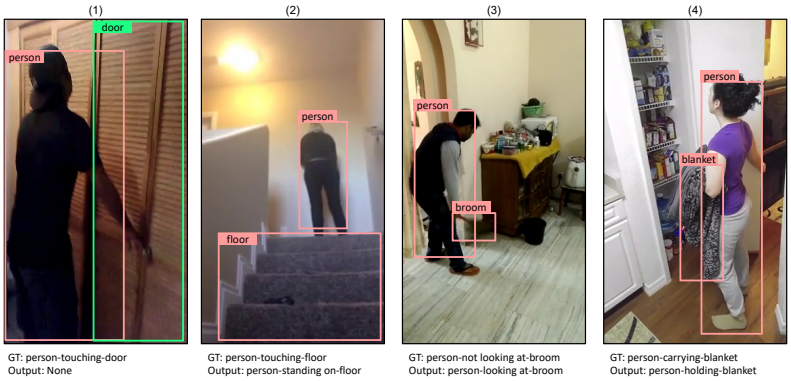

Figure 6: Instances of the most common failure cases. (1) The door is not detected by the object detector and there is no corresponding relationships in the output. (2) STTran predicts that the person is standing on the floor while the ground truth is incorrect. (3) Although the prediction from STTran is wrong, it is difficult for humans to identify whether the person is looking at the broom or not. (4) *carrying* which occurs less frequently in Action Genome is predicted as *holding* with a similar meaning and a higher frequency.