Classification-Then-Grounding: Reformulating Video Scene Graphs as Temporal Bipartite Graphs
********** Supplementary Document **********

Kaifeng Gao[†], Long Chen[‡,*], Yulei Niu[‡], Jian Shao[†], Jun Xiao[†]
[†]Zhejiang University, [‡]Columbia University
{kite_phone,jshao,junx}@zju.edu.cn {zjuchenlong,yn.yuleiniu}@gmail.com

Appendix

This supplementary document is organized as follows:

- More explanations about the temporal bipartite graph are in Sec. A.
- More detailed implementation details are in Sec. B.
- Statistics about predicates with multiple instances for multi-instances grounding are in Sec. C.
- More qualitative results are in Sec. D.
- Potential negative societal impact are in Sec. E.

A. More Explanation about Temporal Bipartite Graph

As shown in Figure 1, compared to existing video scene graphs (a), our temporal bipartite graph (b) have multiple advantages: 1) avoids exhaustively enumerating all entity pairs for predicate prediction; 2) is easier to model entity pairs with multiple predicates; and 3) can be easily extended to more general relations with more semantic roles (*e.g.*, `instrument` [9]).

B. More Implementation Details

Tracklet Detector. We utilized the video object detector MEGA [1] with backbone ResNet-101 [2] to obtain initial frame-level detection results, and adopted deepSORT [8] to generate object tracklets. This detector was trained on a video set and an image set. The video set is a set of downsampled videos from the training set, and the downsample rate were set to 5 and 32 for VidVRD and VidOR, respectively. The image set consists of images with the same categories as the VidSGG dataset, which are selected from the training and validation set of MSCOCO [4] (for VidOR), or MSCOCO plus ILSVRC2016-DET [6] (for VidVRD).

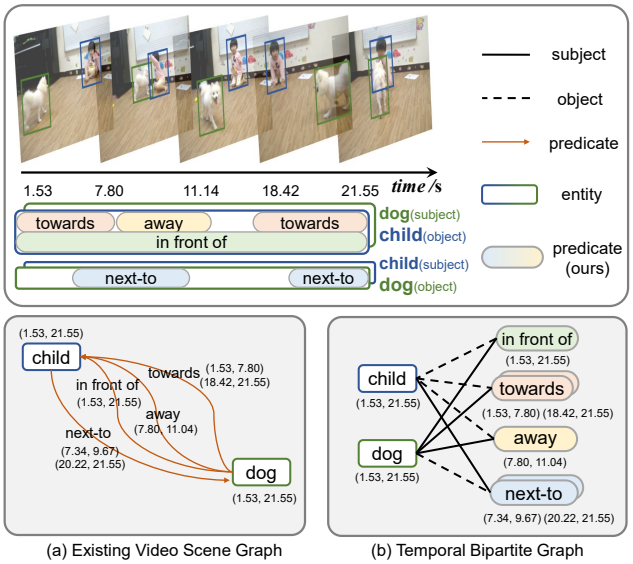


Figure 1. Existing video scene graph vs. temporal bipartite graph.

Parameter Settings. The dimension of bounding box RoI feature d_v was set to 1024, which was determined by MEGA. The dimensions of video I3D feature d_I and word embedding d_w were set to 1024 and 300, respectively. The hidden dimensions d_q and d_e were set to be 512. The output length l_e of pooling operation was 4. The non-linear transforms F_s, F_o are both MLPs. All the MLPs are two-layer FC networks with ReLU and the hidden dimension was 512. All bounding box coordinates and time slots are normalized to the range between (0, 1) w.r.t video size and video length, respectively. The loss factor was set as $\lambda = 30$. The parameters of DEBUG are set to the default settings [5].

Training Details. We trained our model for the classification stage and grounding stage separately. For the classification stage, the model was trained Adam [3] for total 80/60 epochs with batch size 8/4 for VidVRD and VidOR,

*Long Chen is the corresponding author. This work started when LC at Zhejiang University, and YN at Nanyang Technological University.

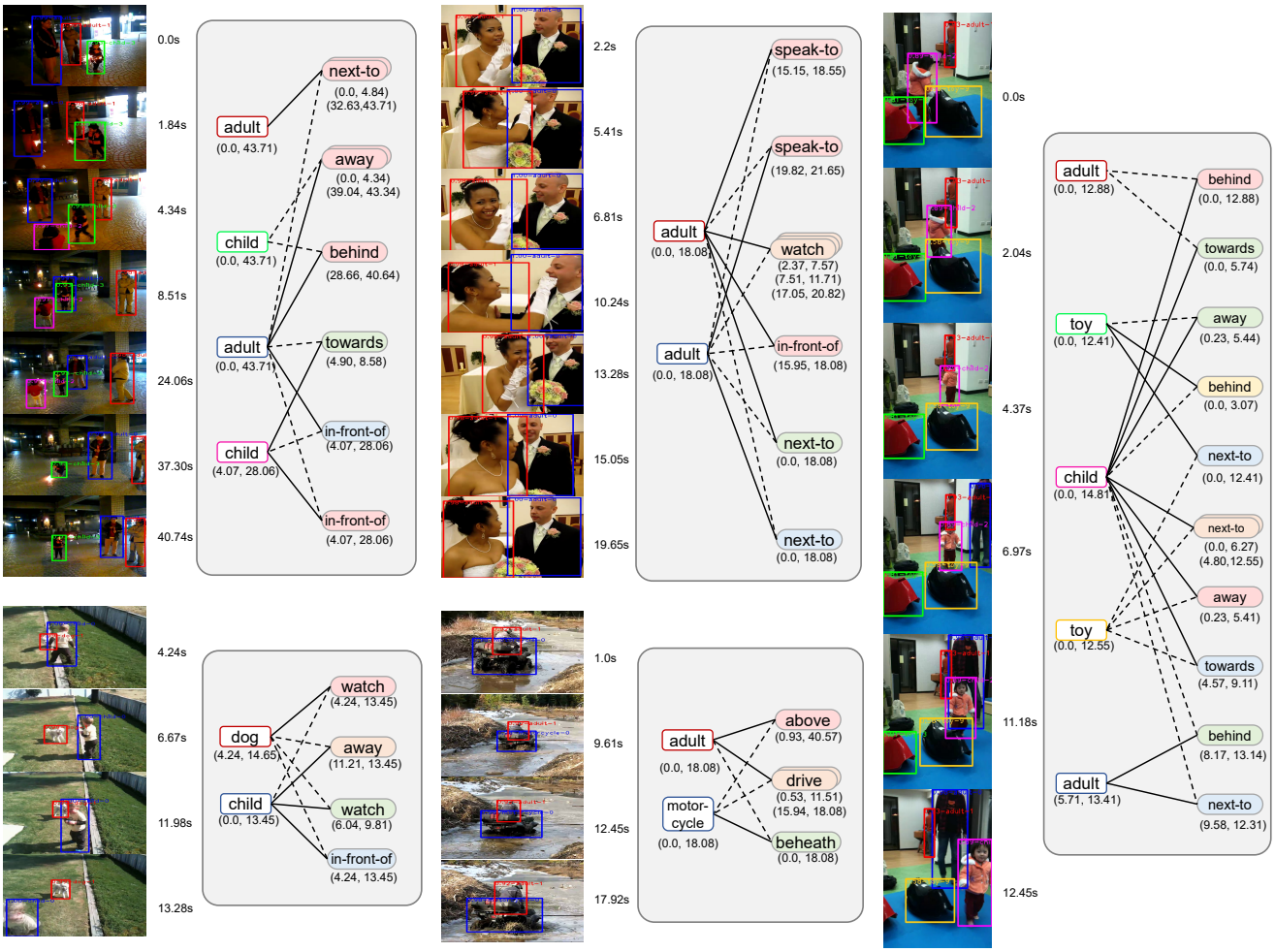


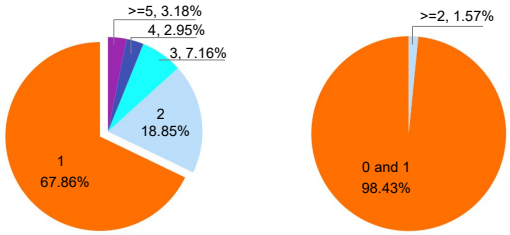
Figure 2. More qualitative results on VidOR validation set. The solid line and dash line represent the `subject` and `object` respectively.

respectively. The learning rate was set to 1e-4 for VidVRD. For VidOR, the learning rate was set to 5e-5 in the first 50 epochs and 1e-5 in the last 10 epochs. For the grounding stage, the model was trained using ground-truth triplet categories in VidOR as language queries. It was trained by Adam [3] with 70 epochs and batch size of 8. The initial learning rate was set to 5e-5, and it decays 5 times in the 40-th and 60-th epoch.

Inference Details. In classification stage, following previous works [7], we kept top-k triplet predictions (10 for VidVRD and 3 for VidOR) for each predicate node. Then, we filtered out duplicated triplets or triplets in which subject and object tracklets has no temporal overlap. In grounding stage, for each triplet query, we obtained K time slots predictions through the multi-instance grounding, where each time slot prediction is associated with a score. Then, we filtered out these triplet queries whose highest score among all time slots is less than 0.2 (which might be false positives returned by the classification stage). For the remaining triplet queries, we add the time slot of subject-object overlapping (with score 1.0) to get total K+1 instances for more robust prediction. Finally, we apply temporal NMS (with the threshold of 0.8) to these K+1 instance, which results in K_j time slots for each predicate p_j. The tracklets for the relation triplet are cropped from e_{j_s}, e_{j_o} according to the time slots of p_j, *i.e.*, each p_j is corresponding to K_j relation triplets.

C. Statistics for Multi-instance Predicates in VidOR

We reported the distribution of predicates with single instance or multiple instances in Figure 3(a). Here each sample is defined as the set of relation triplets with the same subject-object pair and predicate category. Then for those samples with multi-instance predicates, we reported the distribution of number of instances falling into the same bin for multi-instance grounding, as shown in Figure 3(b), where each sample is a bin.



(a) Number of predicate instances (b) Number of instances falling into the same bin

Figure 3. Statistics for multi-instance grounding from VidOR train set. **(a)**:Number of predicate instances with same category in a same subject-object pair. **(b)**:Number of instances falling into the same bin for predicates with multiple instances.

From Figure 3(a), we can observe that although many predicates are single-instance, there are still around 32% of predicates with multiple ground-truth instances. Thus, the proposed multi-instance grounding is indispensable. Furthermore, Figure 3(b) shows that only 1.57% of bins are assigned with two instances or more, *i.e.*, most of the bins are assigned with only one ground-truth target, which shows that our label assignment scheme is suitable for the multi-instance grounding.

D. More Qualitative Results

More qualitative results of our BIG model on VidOR are shown in Figure 2.

E. Potential Negative Societal Impact

Our proposed *classification-then-grounding* is a general framework of video scene graph generation (VidSGG), and there are no known extra potential negative social impact of our framework and BIG model. As for the challenging VidSGG task itself, there might be some wrong predictions (*e.g.*, ⟨`adult`, `kick`, `dog`⟩). When VidSGG is applied to numerous down-stream tasks such as video captioning, video question answering, these wrong predictions might result in some ethical issues, *e.g.*, a wrong caption says that a person is abusing a dog. To avoid the potential ethical issues, we can introduce some common sense knowledge into VidSGG models and design some rule-based methods to filter out those unreasonable relation triplets that involve ethical issues.

References

[1] Yihong Chen, Yue Cao, Han Hu, and Liwei Wang. Memory enhanced global-local aggregation for video object detection. In *CVPR*, pages 10337–10346, 2020. 1

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1

[3] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv*, 2014. 1, 2

[4] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 1

[5] Chujie Lu, Long Chen, Chilie Tan, Xiaolin Li, and Jun Xiao. Debug: A dense bottom-up grounding approach for natural language video localization. In *EMNLP*, pages 5144–5153, 2019. 1

[6] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, pages 211–252, 2015. 1

[7] Xindi Shang, Tongwei Ren, Jingfan Guo, Hanwang Zhang, and Tat-Seng Chua. Video visual relation detection. In *ACM MM*, pages 1300–1308, 2017. 2

[8] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *ICIP*, pages 3645–3649, 2017. 1

[9] Alireza Zareian, Svebor Karaman, and Shih-Fu Chang. Weakly supervised visual semantic parsing. In *CVPR*, pages 3736–3745, 2020. 1