VL-SAT: Visual-Linguistic Semantics Assisted Training for 3D Semantic Scene Graph Prediction in Point Cloud
(Supplementary Material)

A. Implementation Details

2D Data Preparation. Since each 3D scan in the 3DSSG dataset [6] is associated with RGB sequences with known camera poses, thus it is possible to extract 2D image patches associated with each point cloud instance $\mathbf{P}_i$. We first project the 3D points in $\mathbf{P}_i$ to each RGB frame according to the given camera pose, and then calculate the area of the enlarged bounding box surrounded by the projected points. Since then, we rank the frames in the descending order of these areas and select the image patches in the bounding boxes in the top-N frames as the N-view image patches of the instance $\mathbf{P}_i$. The visual features $\mathbf{o}_i$ corresponding to $\mathbf{P}_i$ are thus generated by mean pooling the visual features of N-view image patches through a fixed CLIP vision encoder that has been finetuned on 3DSSG [3, 6].

Architecture Details. We adopt a simple PointNet [5] as the 3D node encoder. As for the 2D node encoder, we use Vit-B-32 architecture [2] as the backbone of the CLIP image encoder. The feature dimension of all the node and edge features in the oracle and 3D model is set to be 512. The structure of GNN is borrowed from SGFN [8], which uses a FAT mechanism to combine neighboring features. All the multi-head self attention (MHSA) or multi-head cross attention (MHCA) structures in our method use 8 heads, with a hidden feature size of 512. According to our experiments, $\rho(\cdot,\cdot)$ in $L_{\text{tri-emb}}$ is implemented with ℓ_1 norm, and the $\rho(\cdot,\cdot)$ in $L_{\text{node-init}}$ is implemented with negative cosine distance.

Splits of Predicates. We split the 26 predicate classes into three parts: *head*, *body*, *tail*. In detail, we sort the predicates according to their frequencies in the training set in descending order and select the top 8 categories as head classes, the last 12 categories as tail classes, and the remaining 6 categories as body classes. You can refer to Tab. S1.

B. More Experiments

B.1. Comparison with Knowledge Distillation Scheme.

To prove the superiority of our proposed VL-SAT scheme, we design a knowledge distillation (KD) scheme as in Fig. S1, which adheres to a teacher-student paradigm. The teacher is a multi-modal model, which fuses visual and geometrical information using bi-directional cross-attention. Besides, to compare with our VL-SAT scheme in a fair manner, we also leverage linguistic assistance in the KD scheme. The student model is the same as our non-VL-SAT model. The knowledge transfer process from teacher to student is implemented with traditional mimic loss and KL loss. As shown in Tab. S2, since our oracle model trained with VL-SAT scheme can combine multi-modal knowledge more effectively, the performance is better than the teacher model of KD scheme among all metrics, *e.g.* 2.1% gains on predicate mA@1. Besides, VL-SAT (ours) outperforms KD (student) with 2.1% gains on triplet mA@50. We think the performance degradation of KD scheme is because the teacher model has a different network structure compared with the student model, and the heterogeneous network structures may hinder the knowledge transfer process as indicated in [7].

B.2. Can RGB Information on Point Cloud Boost 3DSSG Prediction As Well?

Since the VL-SAT scheme boosts 3DSSG prediction significantly, it is intuitive to think about whether adding RGB information directly into 3D point cloud could also do well. We conduct such experiments (namely, $\text{Base}_{\text{CLIP}}$ since we employ ClIP-initialized object classifier in this baseline) in Tab. S3 and find that simply concatenating RGB values to point cloud's XYZ coordinates (as $\text{Base}_{\text{CLIP}}$ (XYZ+RGB)) brings moderate performance drop (as $\text{Base}_{\text{CLIP}}$ (XYZ)) in 3DSSG prediction task. We doubt it is due to over-fitting on RGB values as indicated in [4]. The experiment results also validate the necessity of our VL-SAT scheme.

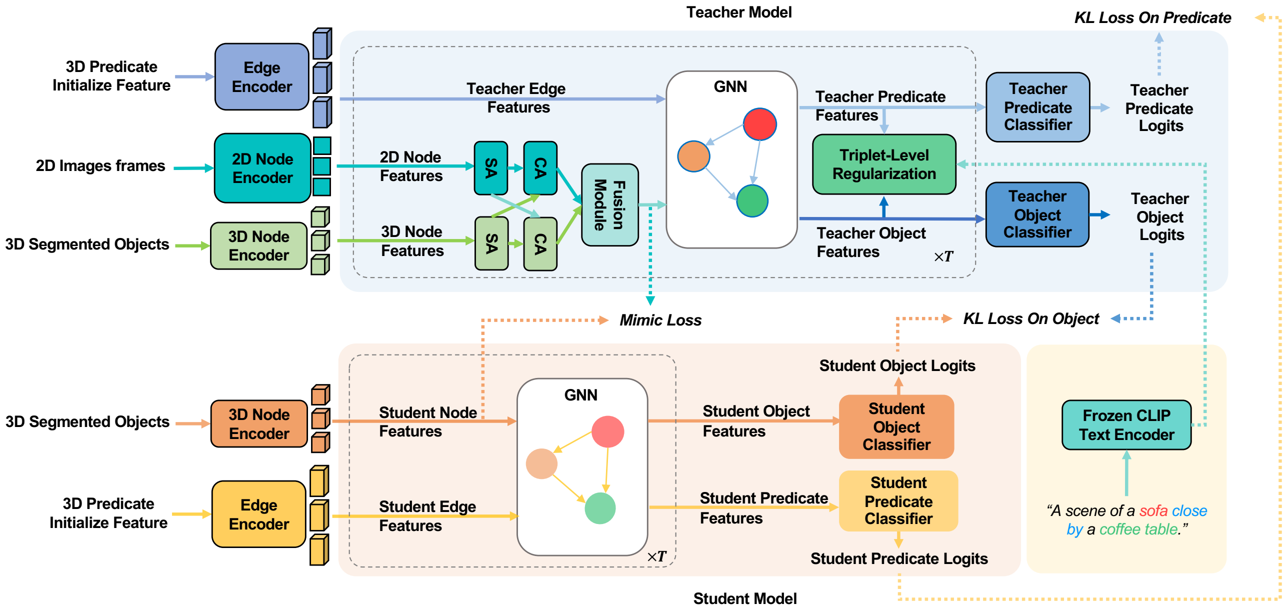

Figure S1. **The Teacher-Student Model based on the Knowledge Distillation Scheme.** During training, the teacher model transfers its knowledge to the student model via feature mimic. Besides, we also add KL loss between teacher logits and student logits on both object and predicate classifiers to advance the knowledge transfer process. During inference, the student model takes the same inputs as the 3D model in our VL-SAT scheme.

Split	Predicate
Head	left, right, front, behind, close by, same as, attached to, standing on
Body	bigger than, smaller than, higher than, lower than, lying on, hanging on
Tail	supported by, inside, same symmetry as, connected to, leaning against, part of, belonging to, build in, standing in, cover, lying in, hanging in

Table S1. Splits of predicates.

Method	Predicate			Triplet	
	mA@1	mA@3	mA@5	mA@50	mA@100
SGFN	41.89	70.82	81.44	58.37	67.61
KD (Teacher)	53.57	72.37	86.18	73.31	81.08
VL-SAT (Oracle)	55.66	76.28	86.45	74.10	81.38
KD (Student)	52.22	72.50	83.18	62.92	71.75
VL-SAT (Ours)	54.03	77.67	87.65	65.09	73.59

Table S2. Results of different knowledge transfer methods. We refer to the multi-modal teacher-student model as Knowledge Distillation (KD) scheme, and then we compare the results with our VL-SAT scheme.

Model	Object		Predicate		Triplet	
	A@5	A@10	mA@3	mA@5	mA@50	mA@100
$\text{Base}_{\text{CLIP}}$ (XYZ)	79.03	86.81	72.50	83.59	60.65	69.71
$\text{Base}_{\text{CLIP}}$ (XYZ+RGB)	76.35	84.19	71.45	79.10	58.76	67.67
VL-SAT(ours)	78.66	85.91	77.67	87.65	65.09	73.59

Table S3. Results of different inputs. We figure out whether adding RGB information directly into the 3D point cloud (XYZ) input can boost 3DSSG prediction performance as our VL-SAT scheme does. $\text{Base}_{\text{CLIP}}$ shares the same network architecture as non-VL-SAT but leverages CLIP-initialized object classifier.

B.3. Influence of Visual Assistance.

To investigate the influence of visual assistance, we conduct experiments without linguistic assistance, *i.e*. CLIP-based object classifier initialization, CLIP-based triplet-level regularization, during training. As shown in Tab. S4, with only visual assistance, our method still obtains 6.66% gain on predicate mA@1 and 3.27% gain on triplet mA@50. Furthermore, we try the visual encoder pretrained on ImageNet21K [1] dataset, which shares the same network structure as the CLIP pretrained visual encoder used in our VL-SAT. The ImageNet21K pretrained visual encoder also shows performance gains over non-VL-SAT model, but is inferior to our CLIP pretrained visual encoder. The result shows that the CLIP pretrained visual encoder possesses a stronger representation ability over the ImageNet21K pretrain visual encoder.

Backbone	Predicate			Triplet	
	mA@1	mA@3	mA@5	mA@50	mA@100
non-VL-SAT	41.99	70.88	81.67	59.58	67.75
CLIP Pretrained	48.65	76.12	87.09	62.85	71.60
ImageNet21k Pretrained	47.43	74.47	85.71	61.36	70.07

Table S4. Results of different visual encoders. We figure out the influence of visual assistance and the influence of visual encoder pretrained using different datasets. We conduct the experiments with a variant of the VL-SAT scheme, which discards all the linguistic assistance, *i.e*. CLIP-based object classifier initialization, and CLIP-based triplet-level regularization.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 1

[3] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021. 1

[4] Charles R Qi, Xinlei Chen, Or Litany, and Leonidas J Guibas. Imvotenet: Boosting 3d object detection in point clouds with image votes. In *CVPR*, pages 4404–4413, 2020. 1

[5] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, pages 652–660, 2017. 1

[6] Johanna Wald, Helisa Dhamo, Nassir Navab, and Federico Tombari. Learning 3d semantic scene graphs from 3d indoor reconstructions. In *CVPR*, pages 3961–3970, 2020. 1

[7] Luting Wang, Xiaojie Li, Yue Liao, Zeren Jiang, Jianlong Wu, Fei Wang, Chen Qian, and Si Liu. Head: Hetero-assists distillation for heterogeneous object detectors. In *ECCV*, pages 314–331. Springer, 2022. 1

[8] Shun-Cheng Wu, Johanna Wald, Keisuke Tateno, Nassir Navab, and Federico Tombari. Scenegraphfusion: Incremental 3d scene graph prediction from rgb-d sequences. In *CVPR*, pages 7515–7525, 2021. 1