JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

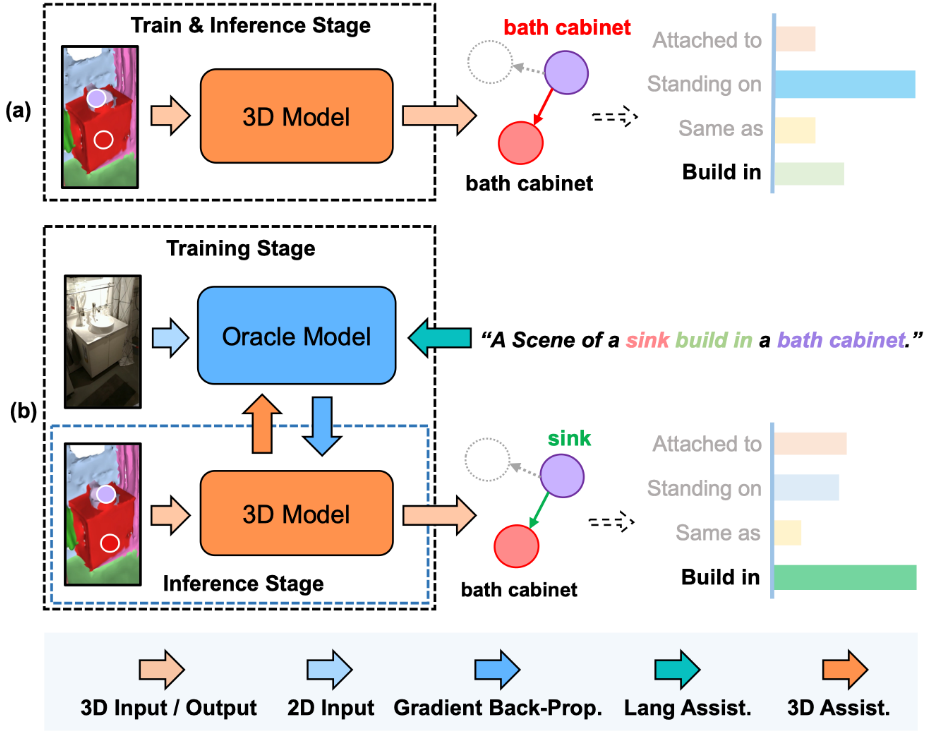
VL-SAT: Visual-Linguistic Semantics Assisted Training for 3D Semantic Scene Graph Prediction in Point Cloud

Ziqin Wang[1], Bowen Cheng[1], Lichen Zhao[1], Dong Xu[2], Yang Tang[3*], Lu Sheng[1*]
[1]School of Software, Beihang University [2]The University of Hong Kong
[3]East China University of Science and Technology

Tag: THU-PM-091

Code is available at https://github.com/wz7in/CVPR2023-VLSAT

The proposed VL-SAT



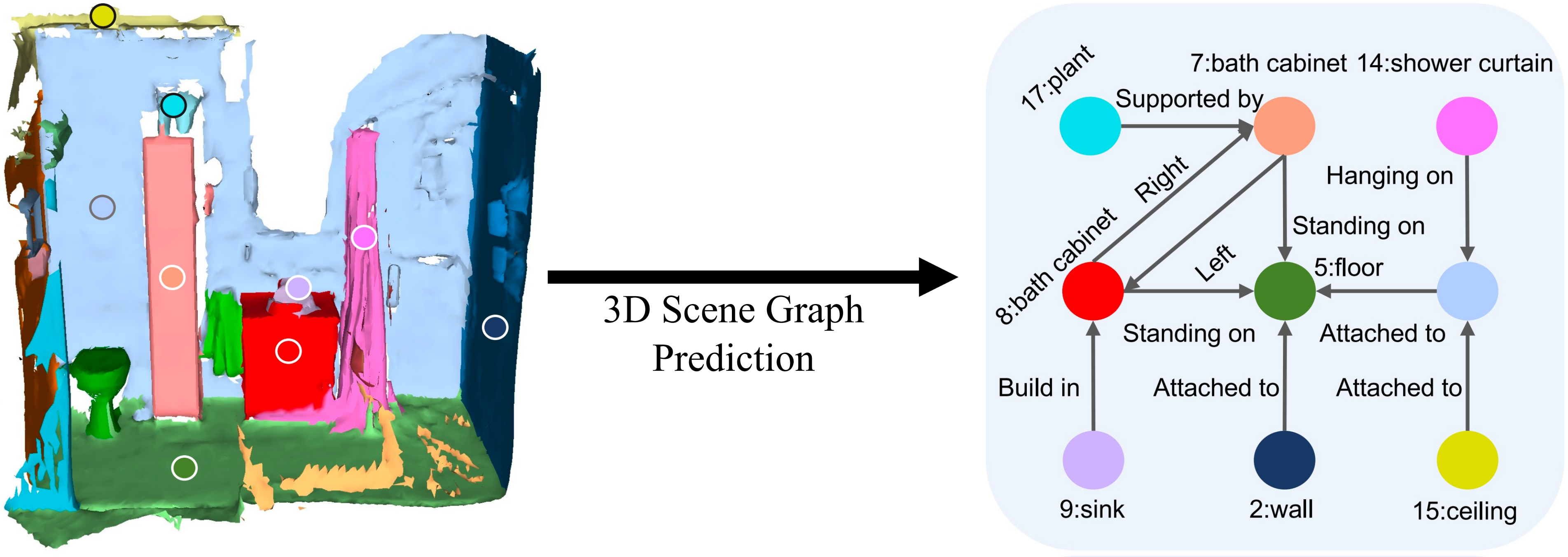
(a) Traditional Training & Inference Scheme
(b) Our VL-SAT Training & Inference Scheme

Our Proposed Training Scheme : VL-SAT

- Constructing an **Oracle Model** that takes 2D input, Language input and 3D input as inputs.
- Optimizing the 3D model by Gradient Back-Propagation.
- Utilizing **multi-modal inputs** during training, but only uses 3D point clouds in inference.

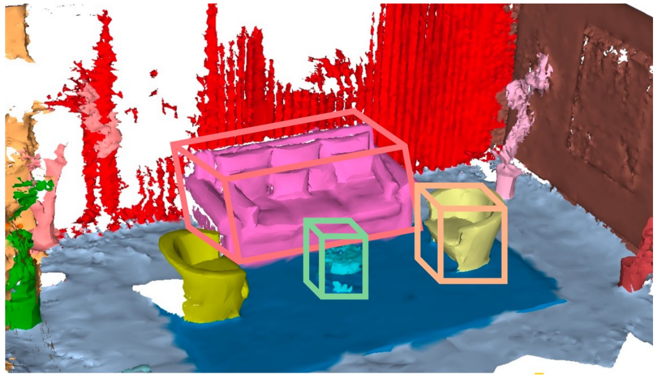
Overview

- **Task: 3D Scene Graph Prediction on 3DSSG Dataset.**



Overview

- Task: 3D Scene Graph Prediction on 3DSSG Dataset.
- **Challenge :**
  - **Limited semantics in point clouds compared to 2D images.**
  - Long-tailed relation distribution.

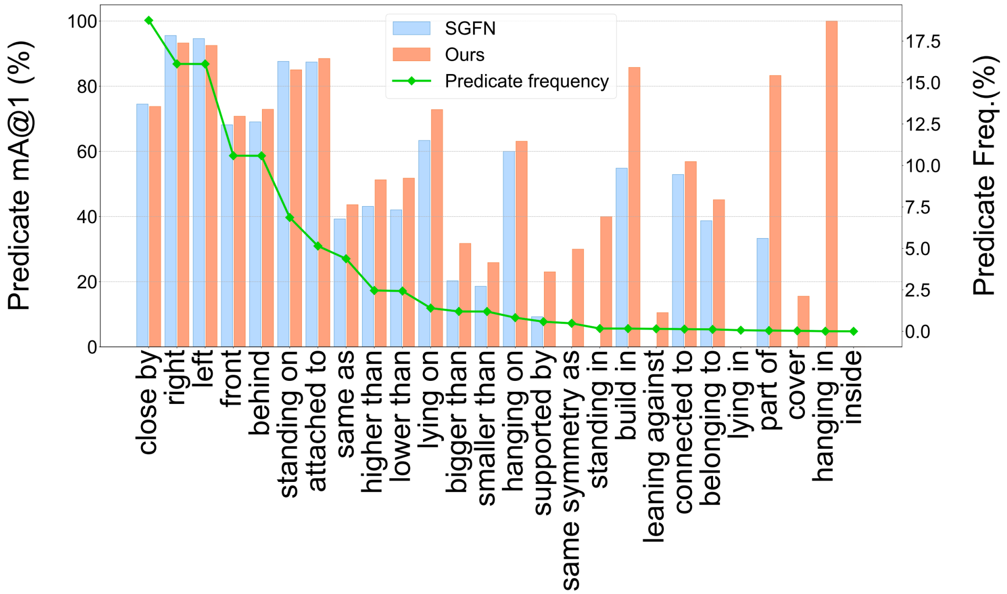


Point Clouds

Images

Overview

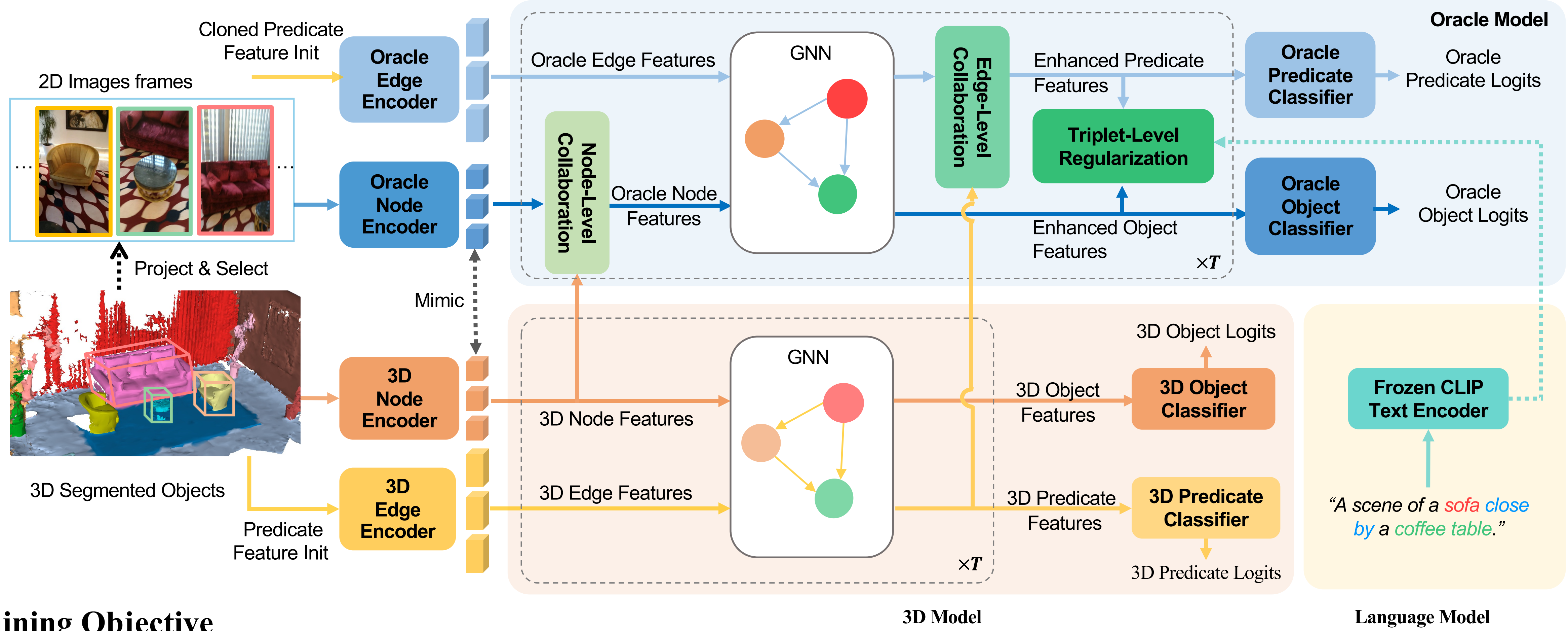
- Task: 3D Scene Graph Prediction on 3DSSG Dataset.
- **Challenge :**
  - Limited semantics in point clouds compared to 2D images.
  - **Long-tailed relation distribution.**



Overview

- Task: 3D Scene Graph Prediction on 3DSSG Dataset.
- Challenge :
  - Limited semantics in point clouds compared to 2D images.
  - Long-tailed relation distribution.
- **Our Method : A model-agnostic training scheme, utilizing Visual semantics along with Linguistic knowledge**

Details of proposed training scheme



Training Objective

- 3D object & predicate loss
- Oracle object & predicate loss
- Triplet regularization loss
- Mimic loss

Experimental Results

- Dataset: 3DSSG
- Evaluation metrics: A(Accuracy), mA(mean Accuracy)

Model	Object			Predicate						Triplet			
	A@1	A@5	A@10	A@1	A@3	A@5	mA@1	mA@3	mA@5	A@50	A@100	mA@50	mA@100
SGPN [34]	48.28	72.94	82.74	**91.32**	98.09	99.15	32.01	55.22	69.44	87.55	90.66	41.52	51.92
SGG_{point} [45]	51.42	74.56	84.15	92.4	97.78	98.92	27.95	49.98	63.15	87.89	90.16	45.02	56.03
SGFN [36]	53.67	77.18	85.14	90.19	98.17	99.33	41.89	70.82	81.44	89.02	91.71	58.37	67.61
non-VL-SAT	54.79	77.62	85.84	89.59	97.63	99.08	41.99	70.88	81.67	88.96	91.37	59.58	67.75
VL-SAT (ours)	**55.66**	**78.66**	**85.91**	89.81	**98.45**	**99.53**	**54.03**	**77.67**	**87.65**	**90.35**	**92.89**	**65.09**	**73.59**
VL-SAT (oracle)	66.39	86.53	91.46	90.66	98.37	99.40	55.66	76.28	86.45	92.67	95.02	74.10	81.38

- Non-VL-SAT: The baseline 3D Model
- VL-SAT(ours): The 3D Model training with VL-SAT
- VL-SAT(oracle): The Oracle Model training with VL-SAT

Experimental Results

Model	SGCls R@20/50/100	PredCls R@20/50/100
with Graph Constraints		
Co-Occurrence [47]	14.8/19.7/19.9	34.7/47.4/47.9
KERN [6]	20.3/22.4/22.7	46.8/55.7/56.5
SGPN [34]	27.0/28.8/29.0	51.9/58.0/58.5
Schemata [24]	27.4/29.2/29.4	48.7/58.2/59.1
Zhang *et al.* [47]	28.5/30.0/30.1	59.3/65.0/65.3
SGFN [36]	29.5/31.2/31.2	65.9/78.8/79.6
VL-SAT (ours)	**32.0/33.5/33.7**	**67.8/79.9/80.8**
without Graph Constraints		
Co-Occurrence [47]	14.1/20.2/25.8	35.1/55.6/70.6
KERN [6]	20.8/24.7/27.6	48.3/64.8/77.2
SGPN [34]	28.2/32.6/35.3	54.5/70.1/82.4
Schemata [24]	28.8/33.5/36.3	49.6/67.1/80.2
Zhang *et al.* [47]	29.8/34.3/37.0	62.2/78.4/88.3
SGFN [36]	31.9/39.3/45.0	68.9/82.8/91.2
VL-SAT (ours)	**33.8/41.3/47.0**	**70.5/85.0/92.5**

Model	SGCls mR@20/50/100	PredCls mR@20/50/100
Co-Occurrence [47]	8.8/12.7/12.9	33.8/47.4/47.9
KERN [6]	9.5/11.5/11.9	18.8/25.6/26.5
SGPN [34]	19.7/22.6/23.1	32.1/38.4/38.9
Schemata [24]	23.8/27.0/27.2	35.2/42.6/43.3
Zhang *et al.* [47]	24.4/28.6/28.8	56.6/63.5/63.8
SGFN [36]	20.5/23.1/23.1	46.1/54.8/55.1
VL-SAT(ours)	**31.0/32.6/32.7**	**57.8/64.2/64.3**

- Dataset: 3DSSG
- Evaluation metrics: R(Recall), mR(mean Recall)

SGCls & PredCls are two tasks defined in 2D Scene Graph Generation Task, which means whether we think about the object class during training / evaluating

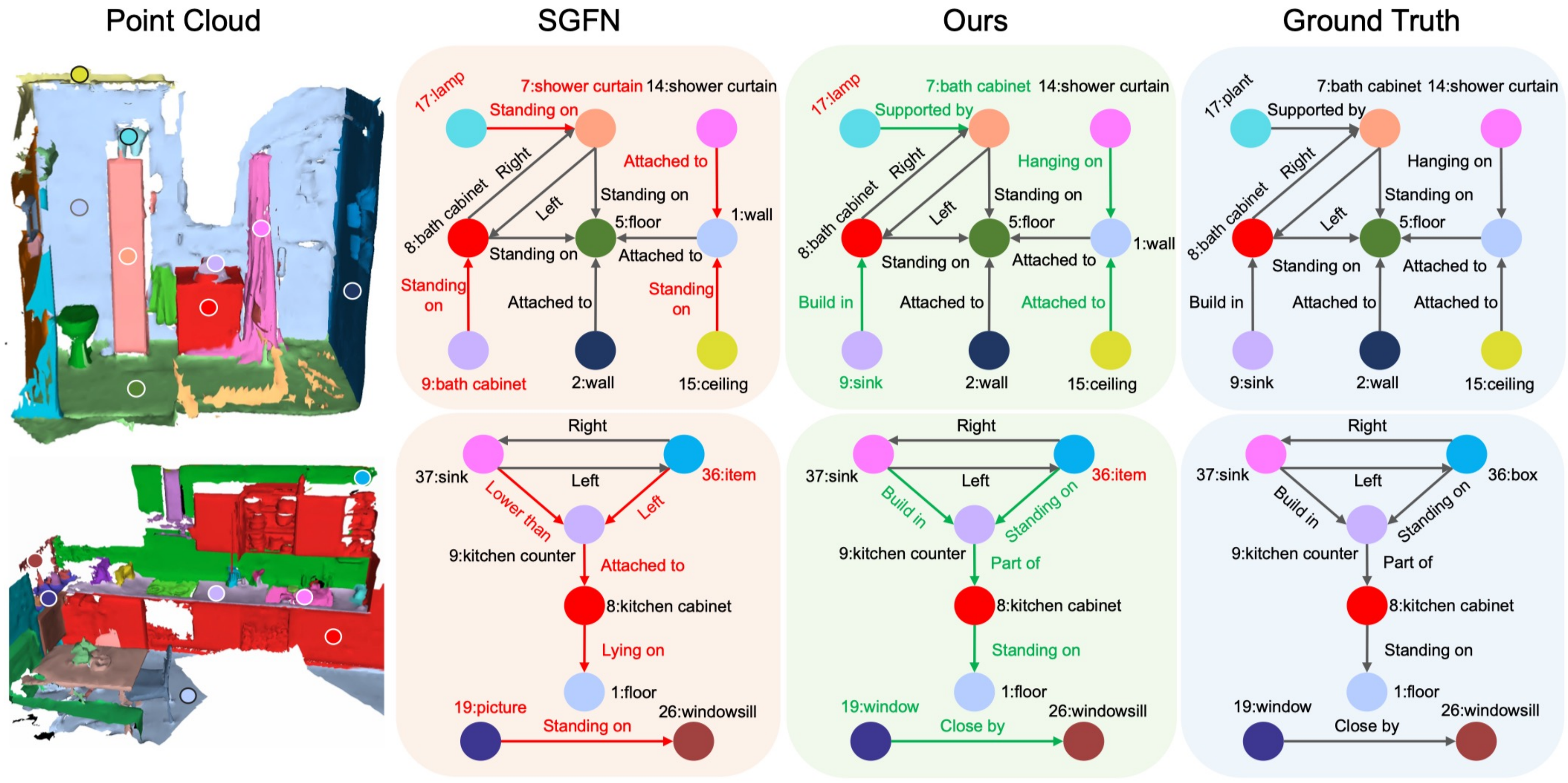
Long Tail Evaluation

- Datasets: 3DSSG
- Evaluation metrics: A(Accuracy), mA(mean Accuracy)

Model	Predicate						Triplet			
	Head		Body		Tail		Unseen		Seen	
	mA@3	mA@5	mA@3	mA@5	mA@3	mA@5	A@50	A@100	A@50	A@100
SGPN [34]	**96.66**	99.17	66.19	85.73	10.18	28.41	15.78	29.60	66.60	77.03
SGFN [36]	95.08	**99.38**	70.02	87.81	38.67	58.21	22.59	35.68	71.44	80.11
non-VL-SAT	95.32	99.01	71.88	88.64	40.01	58.33	21.99	35.44	71.52	80.34
VL-SAT (ours)	96.31	99.21	**80.03**	**93.64**	**52.38**	**66.13**	**31.28**	**47.26**	**75.09**	**82.25**

Our scheme improve the model ability on **tail predicates** and **unseen triplets** greatly.

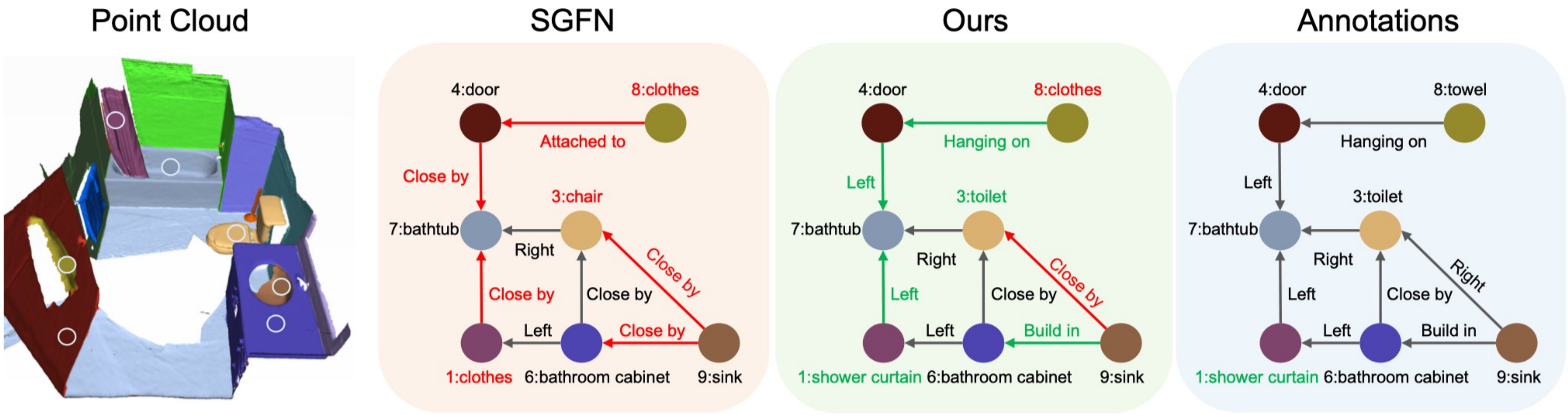
Experimental Results



Red edge: miss-classified edges from SGFN

Green edge: edges corrected by our method

Experimental Results



Qualitative results on the ScanNet

Ablation study & Discussion

Ablation Study

CI	NC	EC	TR	Object		Predicate		Triplet	
				A@5	A@10	mA@3	mA@5	mA@50	mA@100
				77.62	85.84	70.88	81.67	59.58	67.75
✓				79.03	86.81	72.50	83.59	60.65	69.71
✓	✓			**79.28**	**86.82**	73.92	84.78	62.88	71.84
✓	✓	✓		78.71	86.17	76.92	87.08	64.00	72.42
✓	✓	✓	✓	78.66	85.91	**77.67**	**87.65**	**65.09**	**73.59**

All components are important to performance

CI means CLIP-initialized object classifier
NC means node-level collaboration
EC means edge-level collaboration
TR means triplet-level CLIP-based regularization

Different Cross-modal Collaboration Strategies

NC	EC	Object		Predicate		Triplet	
		A@1	A@5	mA@1	mA@3	mA@50	mA@100
CT	CT	55.78	77.58	51.64	74.13	60.37	72.66
CT	CA	56.14	78.38	52.28	75.04	61.50	73.80
CA	CT	56.00	77.68	52.14	73.54	63.92	73.10
CA	CA	55.66	78.66	54.03	77.67	65.09	73.59

Cross-Attention works best in collaboration

NC means node-level collaboration.
EC means edgelevel collaboration.
CT means concatenation.
CA means cross-attention in our method.

Generalization Ability

	Object		Predicate		Triplet	
	A@1	A@5	mA@1	mA@3	mA@50	mA@100
$\mathrm{SGG}_{\mathrm{point}}$ [45]	51.42	74.56	27.95	49.98	45.02	56.03
+VL-SAT	52.08	75.76	38.04	60.36	52.51	64.31
SGFN [36]	53.67	77.18	41.89	70.82	58.37	67.61
+VL-SAT	55.43	78.88	52.91	72.37	63.57	72.02

VL-SAT is applicable to other base 3D models

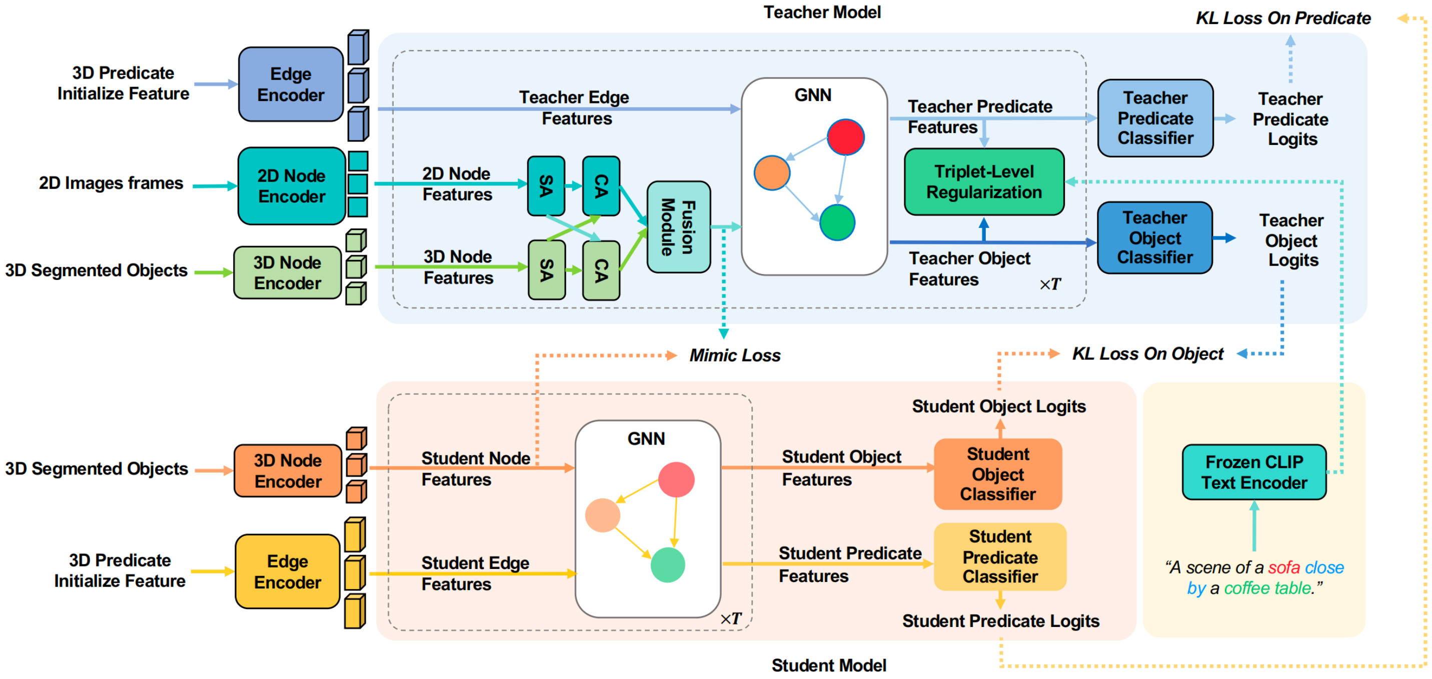
Performance gains brought by our VL-SAT scheme with two reference 3DSSG prediction models.

Ablation study & Discussion

Comparison with Knowledge Distillation Scheme

- **VL-SAT** beats KD (Knowledge Distillation) on both Teacher and Student models.

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



Conclusion

- **Visual-Linguistic S**emantics **A**ssisted **T**raining greatly boosts 3D scene graph prediction.
- **State-of-the-art** performance on 3DSSG Dataset, especially good performance on Tail predicates and Zero-shot triplets.
- **Strong Generalization** ability to various 3D models.

Paper

Code