SAT: 2D Semantics Assisted Training for 3D Visual Grounding (Supplementary Material)

In the first part of the supplementary material, we present additional ablation studies and detailed result analyses. In the second part, we extend our SAT approach to detector-generated 3D proposals. We also discuss how 3D proposal quality influences the 3D grounding accuracy.

A. Experiments

A.1. Ablation studies

Training objectives. We conduct ablation studies on the training objectives introduced in the main paper's Eq. 2. Table A shows the experiments on the Nr3D dataset with ground truth proposals. Same as SAT, all compared methods take extra 2D inputs in the training stage and do not require extra inputs in inference.

Row (b) shows the baseline grounding accuracy with the main 3D grounding loss $\mathcal{L}_{VG}^{O}$ and classification loss $\mathcal{L}_{cls}$ only. Despite input to the model during the training, 2D semantics I does not affect the main model (Q and O) in this baseline, as I is not attended to and no I-related auxiliary losses are included. Therefore, row (b) is equivalent to the main paper's non-SAT baseline and shows a comparable accuracy of 38.0%. SAT's auxiliary objectives of 2D grounding loss $\mathcal{L}_{VG}^{I}$ and object correspondence loss $\mathcal{L}_{cor}$ improve the accuracy to 38.5% and 44.9%, respectively, as shown in rows (c,d). Our proposed SAT jointly applies the two auxiliary objectives and achieves the best accuracy of 49.2%. Furthermore, we find classification loss $\mathcal{L}_{cls}$ helpful to both the baseline and the final SAT model, as shown in rows (a,b) and rows (e,f), respectively.

A.2. Performance breakdown

In this subsection, we show the performance breakdown on the Nr3D [1] dataset to better understand SAT's improvement. We report SAT's performance on subsets with different target object classes, numbers of distractors, query lengths/types, spatial relationships, *etc*. We observe that SAT effectively utilizes the 2D semantics to learn better 3D object representations, and obtains consistent improvements on these subsets.

Numbers of distractors. Table B shows the performance on subsets with different numbers of distractors. We compare non-SAT with our SAT in the bottom part of the table. Intuitively, we observe a performance decrease when there exist more distractors in the scene, *e.g.*, SAT's accuracy drops from 56.3% to 30.6% when the distractor number increases from 2 to 6. On the other hand, the relative improvement of SAT over the non-SAT baseline is consistent on subsets with different numbers of distractors.

Numbers of query words. We also examine the influence of query length on the grounding performance to better understand the model's performance in modeling language queries. Table C split Nr3D into five sub-sets that are roughly the same size based on query lengths. The results show that longer queries more challenging in general. Therefore, the grounding accuracy decreases on longer queries, *e.g.*, from 54.8% to 41.8%, when the query length increase from less than 6 words to more than 14 words. Overall, SAT consistently improves the non-SAT accuracy on different subsets.

Table A. Ablation studies on the training objectives in the main paper's Eq. 2. Experiments are conducted on the Nr3D dataset with ground truth proposals. We highlight "SAT" by underline.

	$\mathcal{L}_{cls}^{O} + \mathcal{L}_{cls}^{Q}$	$\mathcal{L}_{VG}^{I}$	$\mathcal{L}_{cor}$	Acc.
(a)	-	-	-	33.8±0.1%
(b)	✓	-	-	38.0±0.3%
(c)	✓	✓	-	38.5±0.3%
(d)	✓	-	✓	44.9±0.2%
(e)	-	✓	✓	46.0±0.2%
(f)	✓	✓	✓	<u>49.2±0.3%</u>

Table B. Grounding accuracy on Nr3D's subsets with different numbers of distractors.

	Overall	2	3	4	5	6
Percent(%)	100.0	49.0	21.2	15.9	8.4	5.5
non-SAT	37.6	44.4	36.8	25.5	28.8	28.0
SAT	49.0	56.3	48.0	38.6	40.1	30.6

Table C. Grounding accuracy on Nr3D's subsets with different query lengths.

	Overall	2-6	7-8	9-10	11-13	14+
Percent(%)	100.0	21.0	20.5	19.8	17.5	21.2
non-SAT	37.6	44.2	38.5	36.8	35.7	32.4
SAT	49.0	54.8	52.1	49.4	46.3	41.8

Table D. Grounding accuracy on Nr3D's subsets with different target object classes.

	Overall	chair	table	window	door	trash can	pillow	monitor	box	shelf	picture	cabinet
Percent(%)	100.0	10.9	7.2	6.1	5.9	5.8	4.1	4.1	3.4	3.2	3.2	3.1
Avg. #points (K)	3.1	2.2	4.7	4.2	3.7	1.3	0.7	1.2	0.9	7.1	1.8	4.5
Avg. #distractors	3.0	3.6	3.5	2.6	2.6	2.9	3.6	3.9	3.1	2.6	3.0	2.7
non-SAT	37.6	30.7	43.2	39.2	34.7	40.0	41.7	38.8	33.5	38.8	44.7	38.3
SAT	49.0	45.9	52.0	54.9	44.6	53.5	51.1	54.7	40.6	42.1	54.0	53.0

Table E. Grounding accuracy on Nr3D's subsets with different spatial referring keywords in language queries.

	Overall	with spatial	w/o spatial	closest	next to	on the left	on the right	corner	fa(u)rthest
Percent(%)	100.0	76.7	23.3	14.2	8.4	5.4	5.4	5.1	5.0
SAT	49.0	48.4	50.9	49.5	47.4	56.3	48.1	43.8	37.8
SAT w/ Sr3D+	56.4	56.8	54.8	61.0	56.1	62.6	58.0	53.8	51.3

Target's object class. Table D shows the performance on subsets with different target object classes. The upper part of the table shows the percentage of samples in each subset and the subsets' average number of points/distractors. The bottom part compared non-SAT with SAT on different subsets. Overall, we observe consistent improvements on subsets with different target object classes.

Query spatial relationships. As overviewed in the main paper's Section 6.3, training with Sr3D/Sr3D+ mainly benefits the queries with spatial relationship referring. We manually collect the spatial relationship keywords in Nr3D and show the grounding accuracy on each generated subsets. On the left part of Table E, we show the overall performance on the subset with and without spatial relationship referring. We find 76.7% of the queries contain at least one spatial relationship keyword, while the remaining samples do not use spatial referring. On the subset with spatial keywords, the extra Sr3D+ training data leads to an 8.4% improvement on "SAT-Nr3D" from 48.4% to 56.8%. In contrast, the improvement is only 3.9% on the remaining samples. The right part of Table E compares the performance on subsets with specific spatial keywords. We observe larger improvements on the frequently appeared spatial keywords in Sr3D/Sr3D+. For example, the accuracy improves from 49.5% to 61.0% on the keyword "closest," and from 37.8% to 51.3% on the keyword "farthest/furthest."

B. SAT with detector-generated proposals

In the main paper, we focus on the ground truth proposal setting where we assume the access to M ground-truth object point cloud segments as 3D proposals [1]. SAT is compatible with the setting that uses detector-generated proposals [2]. In this section, we present one implementation of extending SAT with detector-generated proposals. We benchmark our approach on the ScanRef dataset [2].

B.1. Method

We obtain M 3D proposals and their feature O with a 3D object detector [5]. It is computationally expensive to project the 3D proposals in each iteration to get the corresponded 2D semantics. Instead, we use the same method introduced in the main paper's Section 3.2 to cache the ground truth 2D image semantics I. The object correspondence between detector-generated 3D proposals O_m and ground-truth 2D semantics I_n does not naturally exist as in the ground truth 3D proposal experiments. To get the 3D-2D object correspondence in the training stage, we compute the 3D IoU between the generated proposals m and the ground truth boxes n (corresponded to 2D semantics I_n) and pair 3D proposals with 2D semantics online by selecting the pair with the maximum IoU. We do not apply the object correspondence loss on the pairs with an IoU less than 0.5. With the IoU computation conducted online in each epoch, our implementation supports the end-to-end optimization of the entire framework.

B.2. Experiment results

Table F shows experiment results on the ScanRef dataset [2]. The upper part of the table contains methods that do not require extra 2D inputs in inference, and the bottom part includes methods that use 2D semantics in both training and inference. The "unique" subset contains samples that do not have distracting objects with the same object class as the target. The remaining samples belong to the "multiple" subset. We note that one previous study [7] simplifies the grounding problem by filtering out proposals that are not in the same object class as the target. We refer to such filtered 3D proposals as "(Filt.)" in the "proposals" column. Consequently, methods with filtered proposals show better performances on the "Unique" subset, which contains no distracting object in the same class as the target. The drawback is that the external object label information is required to perform such filtering.

We focus on the metrics in the *"multiple"* subset, which best indicates the models' performance of 3D visual grounding [1]. We draw two major conclusions. **1)** SAT significantly outperforms the non-SAT baseline by effectively utilizing the 2D semantics in the training stage (SAT: 37.64% and 25.16%, non-SAT: 31.81% and 21.34%). **2)** SAT outperforms the state of the art [7, 3] by large mar-

Table F. 3D visual grounding accuracy on ScanRef [2] with detector-generated proposals. The upper part shows results that do not require extra input in inference, and the bottom part shows methods that use extra inputs. We highlight the best performance that does not use 2D inputs by **bold**. The "unique" subset contains samples with no distracting objects, and the remaining samples are in the "multiple" subset. "(Filt.)" in the "proposals" column indicates that 3D proposals are first filtered by the object class such that the model only needs to select from the proposals in the same class as the target. "(Filt.)" simplifies the grounding problem by using the external object label information.

		Extra 2D input	Proposals	Unique		Multiple		Overall	
				Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5
(a)	ScanRef [2]	✗	VoteNet	60.54%	39.19%	26.95%	16.69%	33.47%	21.06%
(b)	IntanceRefer [7]	✗	PointGroup (Filt.)	**77.13%**	**66.40%**	28.83%	22.92%	38.20%	**31.35%**
(c)	Non-SAT	✗	VoteNet	68.48%	47.38%	31.81%	21.34%	38.92%	26.40%
(d)	SAT (Ours)	✗	VoteNet	73.21%	50.83%	**37.64%**	**25.16%**	**44.54%**	30.14%
(e)	One-stage [6]	✓	None	29.32%	22.82%	18.72%	6.49%	20.38%	9.04%
(f)	ScanRef [2]	✓	VoteNet	63.04%	39.95%	28.91%	18.17%	35.53%	22.39%
(g)	TGNN [3]	✓	3D-UNet	68.61%	56.80%	29.84%	23.18%	37.37%	29.70%
(h)	IntanceRefer [7]	✓	PointGroup (Filt.)	75.72%	64.66%	29.41%	22.99%	38.40%	31.08%

gins (SAT: 37.64% and 25.16%, InstanceRefer: 28.83% and 22.92%). The significant improvements over the non-SAT baseline and the state of the art indicate the effectiveness of our approach in 3D visual grounding.

Furthermore, the state-of-the-art methods [3, 7] find it helpful to replace VoteNet with other proposal generation methods, such as 3D-UNet or PointGroup [4]. We discuss the influence of the proposal quality in Section B.3.

B.3. Discussion

3D proposal quality. When experimented with the detector-generated 3D proposals, the final grounding accuracy is influenced by two factors, *i.e.*, the quality of generated proposals and the main grounding objective of point-cloud-language modeling. We observe that the current 3D proposal quality is still somewhat limited. When using VoteNet [5] for proposal generation, ScanRef reports an oracle Acc@0.5 of 54.33%, where the best proposal is selected as the final prediction. Because of the imperfect proposal quality, previous studies [3, 7] find it effective to boost the grounding accuracy by simply replacing proposal generation methods [3, 4].

Despite the large influence of proposal quality on grounding accuracy, we argue that the point-cloud-language joint representation learning is the core problem of 3D visual grounding. We expect the fast-growing 3D object detection studies to bring stronger detectors in the future, which alleviates the proposal quality problem. Therefore, in this study, we focus on the unique joint representation learning problem in 3D visual grounding, and evaluate methods with the metrics that best reflect the models' grounding performance. Specifically, we focus on "accuracy" when experimented with ground truth proposals, and the "multiple" accuracy when experimented with detector-generated proposals. For the former, SAT surpasses the state of the art by large margins on Nr3D (+10.4% in absolute accuracy) and Sr3D [1] (+9.9%). Similarly on ScanRef, SAT-GT achieves an Acc@0.5 of 66.01%, surpassing the state of the art by large margins (ScanRef-GT: 40.06%, InstanceRef-GT: 55.37%). For the latter, SAT significantly outperforms the state of the art as shown in Table F's "multiple" column. In summary, SAT effectively uses 2D semantics to assist 3D visual grounding and sets the new state of the art on multiple 3D visual grounding datasets.

References

[1] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *ECCV*, pages 422–440. Springer, 2020. 1, 2, 3

[2] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *ECCV*, 2020. 2, 3

[3] Pin-Hao Huang, Han-Hung Lee, Hwann-Tzong Chen, and Tyng-Luh Liu. Text-guided graph neural networks for referring 3d instance segmentation. In *AAAI*, 2021. 2, 3

[4] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. In *CVPR*, pages 4867–4876, 2020. 3

[5] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *ICCV*, pages 9277–9286, 2019. 2, 3

[6] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *ICCV*, pages 4683–4693, 2019. 3

[7] Zhihao Yuan, Xu Yan, Yinghong Liao, Ruimao Zhang, Zhen Li, and Shuguang Cui. Instancerefer: Cooperative holistic understanding for visual grounding on point clouds through instance multi-level contextual referring. *arXiv preprint arXiv:2103.01128*, 2021. 2, 3