Supplementary Materials for End-to-End 3D Dense Captioning with Vote2Cap-DETR

Sijin Chen[1*] Hongyuan Zhu[2] Xin Chen[3] Yinjie Lei[4] Gang YU[3] Tao Chen[1†]
[1]Fudan University [2]Institute for Infocomm Research (I^{2}R) & Centre for Frontier AI Research (CFAR), A*STAR, Singapore
[3]Tencent PCG [4]Sichuan University

https://github.com/ch3cook-fdu/Vote2Cap-DETR

In our supplementary material, we first propose a non-transformer baseline for our method that builds on VoteNet [9] in section A. Then, we provide additional experimental details in section B. Finally, we provide several qualitative studies in section C. It is also worth mentioning that our proposed Vote2Cap-DETR sets **a new state-of-the-art** on the Scan2Cap online test benchmark (Figure 1).

A. VoteNet Baseline with Set-to-Set Training

In this section, we perform ablation studies by replacing our Vote2Cap-DETR's components (SceneEncoder, Vote Query, Transformer Decoder) with VoteNet to study the behavior of non-transformer architecture's behavior. In Table 1, we observe that without delicate hand-crafted relation modeling modules, the VoteNet baseline surpasses 3DJCG [2] by 3.48 C@0.5↑, 6.23 C@0.25↑ and achieves comparable results on other metrics with MLE training. The results demonstrate that the novel caption head and set-to-set training can also improve non-transformer architecture's dense captioning performance. On the other hand, the VoteNet baseline still falls short in terms of our Vote2Cap-DETR, which demonstrates that Vote Query can help learn more discriminate features in an end-to-end manner for end tasks without resorting to many hand-crafted components as in VoteNet.

Method	IoU = 0.25				IoU = 0.5			
	C↑	B-4↑	M↑	R↑	C↑	B-4↑	M↑	R↑
3DJCG [2]	64.70	**40.17**	27.66	**59.23**	49.48	**31.03**	24.22	**50.80**
Ours(VoteNet)	**70.93**	39.92	**28.09**	58.88	**52.96**	30.59	**24.40**	50.10
Ours(Full)	**72.79**	39.17	28.06	**59.23**	**59.32**	**32.42**	**25.28**	**52.53**

Table 1. **VoteNet baseline with set-to-set training.** We replace Vote2Cap-DETR's components with VoteNet. One can see that non-transformer VoteNet architecture also benefits from our novel caption head and set to set training. Although there are still performance gaps with our Vote2Cap-DETR architecture.

*Part of this work was accomplished under supervison by Dr. Hongyuan Zhu from A*STAR, Singapore.

†Corresponding author.

B. Experiments

We provide comparisons with other 3DETR attempts (section B.1), studies on an unsupervised pre-trained backbone (section B.2), evaluations on the Scan2Cap online test benchmark (section B.3) as well as additional experimental details (section B.4 & B.5) in this section.

B.1. Comparison to Other 3DETR Attempts

As few works directly improve 3DETR, we compare our method with the hybrid matching proposed by ($\mathcal{H}$-DETR [6], 2022) and the learnable anchor queries proposed by (Anchor-DETR [13], 2022) in Table 2. Results show that both methods are inferior to our method. Further, we find that though hybrid matching accelerates the training of 3DETR-m in the early training epochs, it still falls behind our approach when the model converges.

Model	Modification	(20k)AP@0.5↑	(40k)AP@0.5↑	(80k)AP@0.5↑	(160k)AP@0.5↑
3DETR-m	-	28.26	37.27	43.41	48.18
3DETR-m	hybrid	**35.10**	**42.72**	45.83	47.50
3DETR-m	anchor	22.94	28.85	35.44	40.06
Vote2Cap-DETR	-	32.70	40.90	**47.62**	**52.49**

Table 2. **Comparison to other 3DETR attempts.** We compare the detection performance of different methods that improve 3DETR in the 20k, 40k, 80k, 160k *-th* iteration.

B.2. Whether a Pre-Trained Backbone Helps

We adopt the vanilla transformer encoder pre-trained on point clouds reconstructed from single-view depth maps in (MaskPoint [7], 2022) as the backbone. It is worth mentioning that the vanilla transformer encoder is different from the masked one used in the main paper. For a fair comparison, we train Vote2Cap-DETR with the replaced backbone from scratch as a baseline in Table 3. Experiments show that the pre-trained backbone does improve performance for 3D dense captioning compared to training from scratch.

B.3. Scan2Cap Online Test Benchmark

Our proposed Vote2Cap-DETR achieves a new state-of-the-art for all metrics on the Scan2Cap online test

pre-train	C@0.5↑	B-4@0.5↑	M@0.5↑	R@0.5↑	AP@0.5↑	AR@0.5↑
-	52.67	29.57	24.13	49.70	42.83	60.63
✓	**53.65**	**30.55**	**24.43**	**50.29**	**43.95**	**62.62**

Table 3. **Whether a unsupervised pre-trained backbone helps improve 3D dense captioning.**

benchmark (Figure 1, https://kaldir.vc.in.tum.de/scanrefer_benchmark/benchmark_captioning).

B.4. Per-Class mAP Results

We list per class mAP results for VoteNet [9], 3DETR [8], and our proposed Vote2Cap-DETR on ScanNet scenes [5] under an IoU threshold of 0.5 in Table 4. The overall performance is listed in the main paper.

B.5. Implementation Details

Our proposed Vote2Cap-DETR first goes through the feature encoding module, then we generate vote queries from the encoded feature as object queries, and we decode the vote queries to bounding boxes and captions in the end.

Feature Encoding directly operates on the input point cloud PC to 1,024 tokens with a feature size of 256. We first tokenizes the input point cloud $PC = [p_{in}; f_{in}] \in \mathbb{R}^{40,000\times(3+d_{in})}$ to point tokens $[p_{token}; f_{token}] \in \mathbb{R}^{2,048\times(3+256)}$ with a set-abstraction layer [10] with hidden sizes of $[3 + d_{in}, 64, 128, 256]$. Then, our scene encoder encodes point tokens $[p_{token}; f_{token}] \in \mathbb{R}^{2,048\times(3+256)}$ to $[p_{enc}; f_{enc}] \in \mathbb{R}^{1,024\times(3+256)}$. We adopt the same encoder as 3DETR-m [8], which contains a three-layer transformer encoder with a set-abstraction layer between the first two layers. Each encoder layer has a feature size of 256 and **F**eed **F**orward **N**etwork (FFN) with a hidden size of 128. The first encoder layer operates on 2,048 points, while the last two operate on the 1,024 points downsampled by the set-abstraction layer. Additionally, three binary attention masks are applied to each encoder layer with a radius of [0.16, 0.64, 1.44], respectively, to force the interactions of points in a given radius.

Vote Query Generator generates 256 object queries $[p_{vq}; f_{vq}] \in \mathbb{R}^{256\times(3+256)}$ from the encoded points $[p_{enc}; f_{enc}] \in \mathbb{R}^{1,024\times(3+256)}$. It contains an FFN FFN_{vote} with a hidden size of 256 to generate offset estimation and feature projection with respective to f_{enc}. It also use a set abstraction layer to gather feature $f_{vq} \in \mathbb{R}^{256\times256}$ from encoded scene feature for $p_{vq} \in \mathbb{R}^{256\times3}$ as described in the main paper.

Parallel Decoding aims to decode the vote queries $[p_{vq}; f_{vq}]$ to corresponding box estimations and captions. The transformer decoder consists of eight identical transformer decoder layers with four heads for both self-attention and cross-attention. It operates on vote queries $[p_{vq}; f_{vq}]$ and encoded feature $[p_{enc}; f_{enc}]$ for the final query feature $[p_{vq}, f_{out}] \in \mathbb{R}^{256\times(3+256)}$. Follow the transformer decoder are two parallel heads, the detection head and the caption head. The detection head generates center offset estimation ($[-0.5, 0.5]^3$) from vote queries' absolute location p_{vq}, normalized size estimation ($[0, 1]^3$), and semantic class estimation from f_{out} using separate FFN heads with a hidden size of 256. Note that we do not estimate the rotation angles since ScanNet [5] does not contain any rotated boxes. Our proposed caption head, DCC, generates captions with respect to final query features f_{out} as $\mathcal{V}^q$ and p_{vq}'s surrounding contextual features $\mathcal{V}^s$. DCC is a two layer transformer decoder with four heads for multi-head attentions, as well as a feature size of 256, a sinusoid position encoding, and a vocabulary of 3,433 for ScanRefer [3] and 2,937 for Nr3D [1].

C. Qualitative Results

Qualitative results on Nr3D. We showcase qualitative results on 3D dense captioning on the Nr3D [1] dataset in Figure 2. Our proposed Vote2Cap-DETR is also able to generate tight bounding boxes as well as accurate description for each object in a 3D scene.

Visualization results of vote queries. We visualize the vote queries' position p_{vq} in our Vote2Cap-DETR and seed queries' position p_{seed} of 3DETR in Figure 3. Most of the vote queries focus on objects in a 3D scene, while p_{seed} is mostly distributed in background areas.

Visualization of detection results. We visualize several detection results in Figure 4. Our proposed Vote2Cap-DETR is able to generate accurate box predictions for a 3D scene.

References

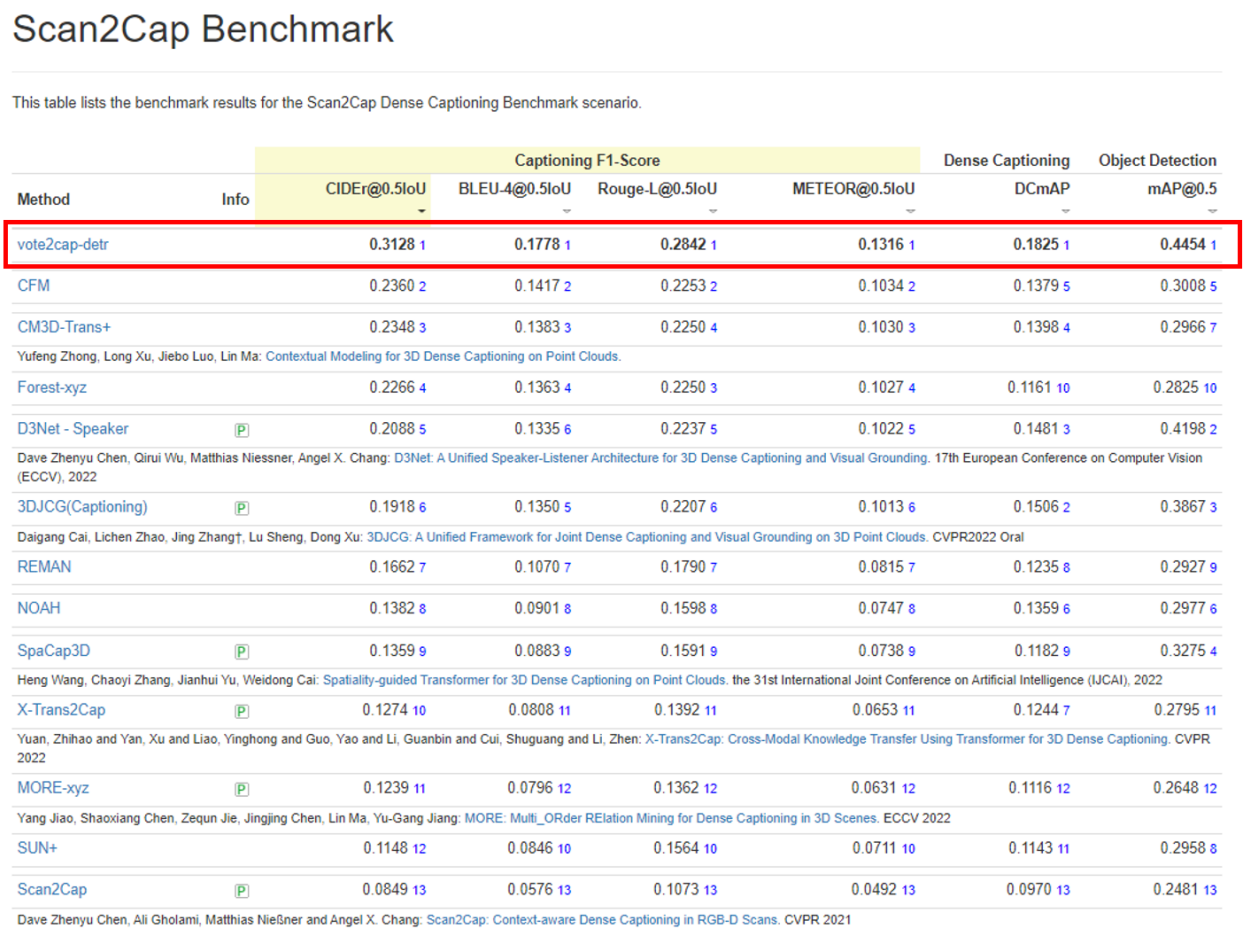
[1] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *European Conference on Computer Vision*, pages 422–440. Springer, 2020. 2, 4

[2] Daigang Cai, Lichen Zhao, Jing Zhang, Lu Sheng, and Dong Xu. 3djcg: A unified framework for joint dense captioning and visual grounding on 3d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16464–16473, 2022. 1

[3] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European Conference on Computer Vision*, pages 202–221. Springer, 2020. 2

[4] Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3193–3203, 2021. 3

[5] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet:



Scan2Cap Benchmark

This table lists the benchmark results for the Scan2Cap Dense Captioning Benchmark scenario.

Method	Info	Captioning F1-Score				Dense Captioning	Object Detection
		CIDEr@0.5IoU	BLEU-4@0.5IoU	Rouge-L@0.5IoU	METEOR@0.5IoU	DCmAP	mAP@0.5
vote2cap-detr		**0.3128** 1	**0.1778** 1	**0.2842** 1	**0.1316** 1	**0.1825** 1	**0.4454** 1
CFM		0.2360 2	0.1417 2	0.2253 2	0.1034 2	0.1379 5	0.3008 5
CM3D-Trans+		0.2348 3	0.1383 3	0.2250 4	0.1030 3	0.1398 4	0.2966 7
Yufeng Zhong, Long Xu, Jiebo Luo, Lin Ma: Contextual Modeling for 3D Dense Captioning on Point Clouds.							
Forest-xyz		0.2266 4	0.1363 4	0.2250 3	0.1027 4	0.1161 10	0.2825 10
D3Net - Speaker	P	0.2088 5	0.1335 6	0.2237 5	0.1022 5	0.1481 3	0.4198 2
Dave Zhenyu Chen, Qirui Wu, Matthias Niessner, Angel X. Chang: D3Net: A Unified Speaker-Listener Architecture for 3D Dense Captioning and Visual Grounding. 17th European Conference on Computer Vision (ECCV), 2022							
3DJCG(Captioning)	P	0.1918 6	0.1350 5	0.2207 6	0.1013 6	0.1506 2	0.3867 3
Daigang Cai, Lichen Zhao, Jing Zhang†, Lu Sheng, Dong Xu: 3DJCG: A Unified Framework for Joint Dense Captioning and Visual Grounding on 3D Point Clouds. CVPR2022 Oral							
REMAN		0.1662 7	0.1070 7	0.1790 7	0.0815 7	0.1235 8	0.2927 9
NOAH		0.1382 8	0.0901 8	0.1598 8	0.0747 8	0.1359 6	0.2977 6
SpaCap3D	P	0.1359 9	0.0883 9	0.1591 9	0.0738 9	0.1182 9	0.3275 4
Heng Wang, Chaoyi Zhang, Jianhui Yu, Weidong Cai: Spatiality-guided Transformer for 3D Dense Captioning on Point Clouds. the 31st International Joint Conference on Artificial Intelligence (IJCAI), 2022							
X-Trans2Cap	P	0.1274 10	0.0808 11	0.1392 11	0.0653 11	0.1244 7	0.2795 11
Yuan, Zhihao and Yan, Xu and Liao, Yinghong and Guo, Yao and Li, Guanbin and Cui, Shuguang and Li, Zhen: X-Trans2Cap: Cross-Modal Knowledge Transfer Using Transformer for 3D Dense Captioning. CVPR 2022							
MORE-xyz	P	0.1239 11	0.0796 12	0.1362 12	0.0631 12	0.1116 12	0.2648 12
Yang Jiao, Shaoxiang Chen, Zequn Jie, Jingjing Chen, Lin Ma, Yu-Gang Jiang: MORE: Multi_ORder RElation Mining for Dense Captioning in 3D Scenes. ECCV 2022							
SUN+		0.1148 12	0.0846 10	0.1564 10	0.0711 10	0.1143 11	0.2958 8
Scan2Cap	P	0.0849 13	0.0576 13	0.1073 13	0.0492 13	0.0970 13	0.2481 13
Dave Zhenyu Chen, Ali Gholami, Matthias Nießner and Angel X. Chang: Scan2Cap: Context-aware Dense Captioning in RGB-D Scans. CVPR 2021							

Figure 1. **Scan2Cap [4] test benchmark.** Our proposed Vote2Cap-DETR achieves a new state-of-the-art for all metrics on the Scan2Cap online test benchmark https://kaldir.vc.in.tum.de/scanrefer_benchmark/benchmark_captioning.

Method	cabinet	bed	chair	sofa	table	door	window	bookshelf	picture	counter	desk	curtain	refrigerator	shower curtain	toilet	sink	bathtub	others
VoteNet [9]	21.41	78.41	78.47	74.44	55.42	34.68	14.91	29.80	9.04	16.57	51.12	34.62	40.12	45.82	89.93	37.23	83.41	13.79
3DETR [8]	26.30	75.78	82.19	59.15	62.25	39.16	21.47	33.14	16.45	34.41	49.68	38.34	42.83	33.33	88.68	52.62	82.41	29.06
Vote2Cap-DETR	31.98	81.48	85.80	64.37	65.20	41.19	28.47	39.81	22.94	39.02	54.46	36.66	40.19	56.10	87.97	44.38	85.12	33.28

Table 4. **Per-class AP under IoU threshold of 0.5 on ScanNet scenes.**

Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 2

[6] Ding Jia, Yuhui Yuan, Haodi He, Xiaopei Wu, Haojun Yu, Weihong Lin, Lei Sun, Chao Zhang, and Han Hu. Detrs with hybrid matching. *arXiv preprint arXiv:2207.13080*, 2022. 1

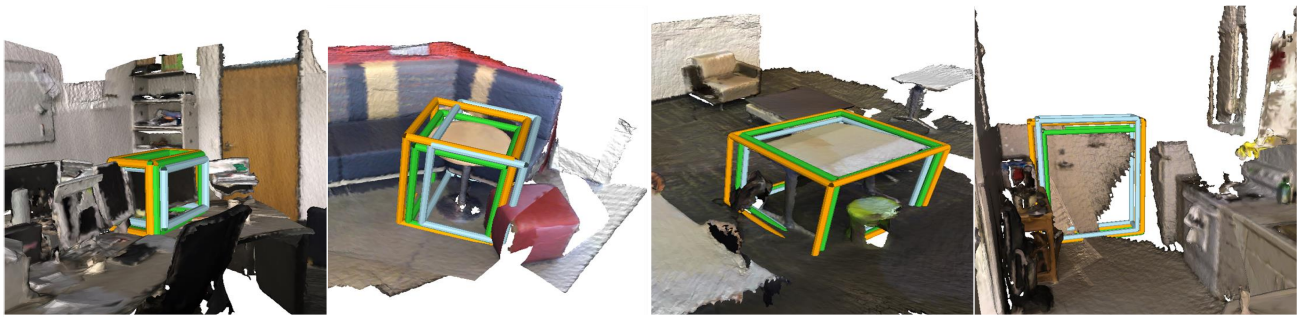
[7] Haotian Liu, Mu Cai, and Yong Jae Lee. Masked discrimination for self-supervised learning on point clouds. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pages 657–675. Springer, 2022. 1

[8] Ishan Misra, Rohit Girdhar, and Armand Joulin. An end-to-end transformer model for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2906–2917, 2021. 2, 3, 6

[9] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9277–9286, 2019. 1, 2, 3, 6

[10] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 2

[11] Heng Wang, Chaoyi Zhang, Jianhui Yu, and Weidong Cai. Spatiality-guided transformer for 3d dense captioning on



scene0025_00

SpaCap3D: The keyboard closest to the door.

Ours: The monitor closest to the door.

GT: The monitor closest to the door.

scene0081_00

SpaCap3D: The table with the plant on it.

Ours: The round table in the corner of the room.

GT: Round table near the end of the couch.

scene0187_00

SpaCap3D: The table that is not in the middle of the room.

Ours: The table in the middle of the room.

GT: The center-most box. In-between the two chairs.

scene0207_00

SpaCap3D: The door that is open.

Ours: The door to the left of the stove.

GT: White door nearest to stove.

scene0300_00

SpaCap3D: The bag on the table.

Ours: The monitor closest to the window.

GT: The computer monitor the is closest to the windows.

scene0307_00

SpaCap3D: The bookshelf that is not next to the tv.

Ours: It is the bookshelf closest to the door.

GT: The largest shelf by the door.

scene0527_00

SpaCap3D: The mirror that is not above the sink.

Ours: Facing the sink, the cabinet on the right.

GT: Upper cabinet above the right sink.

scene0645_00

SpaCap3D: The ottoman closest to the door.

Ours: The ottoman closest to the couch.

GT: The ottoman closest to the couch.

Figure 2. **Visualization of 3D dense captioning on Nr3D [1].** We visualize several results generated by our proposed Vote2Cap-DETR comparing with SpaCap3D [11] on the Nr3D [12] dataset. Our proposed method generates tight bounding box as well as accurate description.

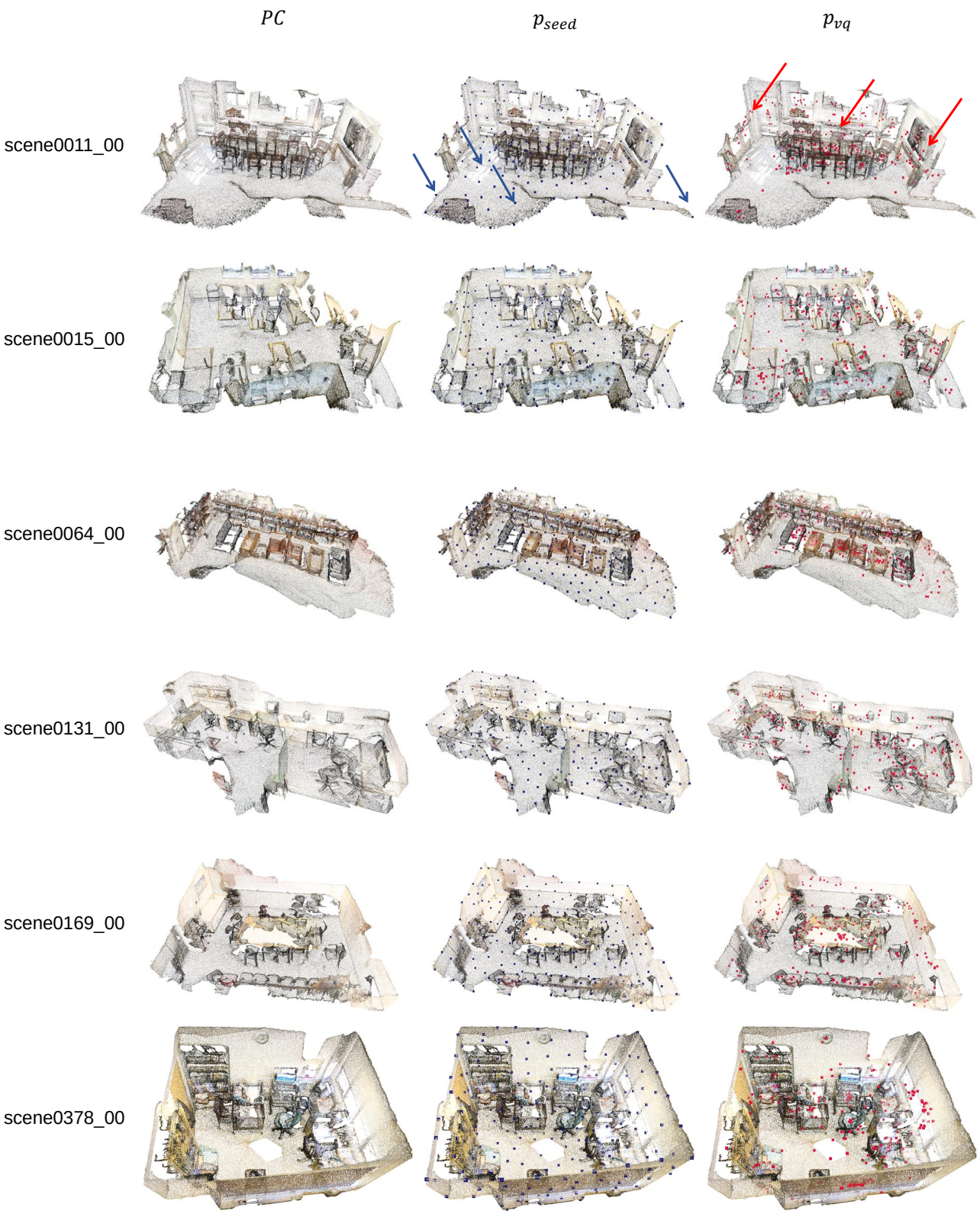


Figure 3. **Visualization of vote queries.** We visualize absolute position of different object queries, p_{seed} used in 3DETR (marked in blue) and p_{vq} used in our proposed Vote2Cap-DETR (marked in red) with the input point cloud PC. Most of the vote queries focus on objects in a 3D scene (as red arrows pointed out), while p_{seed} is mostly distributed in background areas (as blue arrows pointed out).

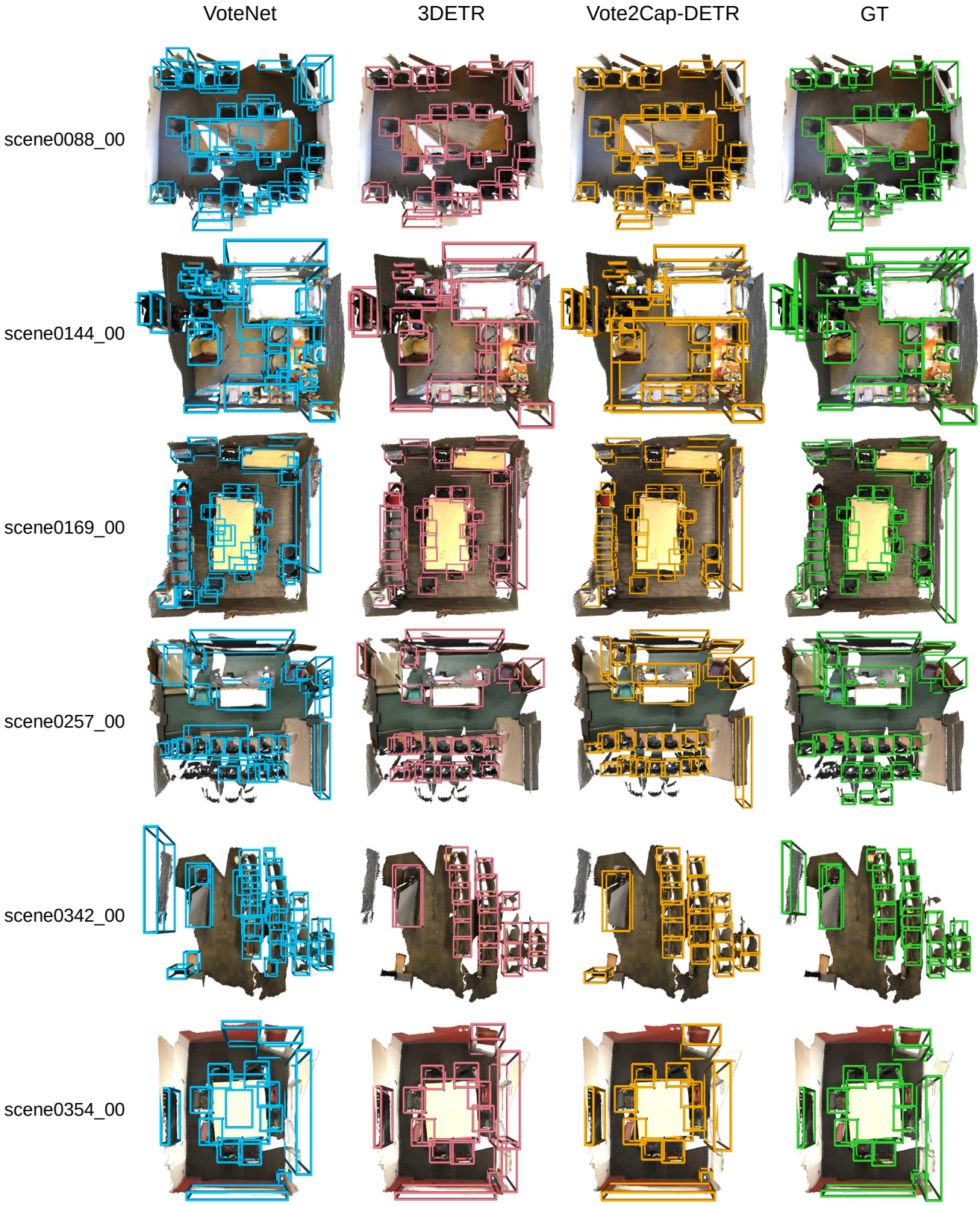


Figure 4. **Visualization of detection performance.** We visualize detection results of VoteNet [9], 3DETR [8], and our proposed Vote2Cap-DETR. Our proposed Vote2Cap-DETR is able to generate accurate localization results.

point clouds. *arXiv preprint arXiv:2204.10688*, 2022. 4

[12] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*, pages 180–191. PMLR, 2022. 4

[13] Yingming Wang, Xiangyu Zhang, Tong Yang, and Jian Sun. Anchor detr: Query design for transformer-based detector. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 2567–2575, 2022. 1