3D-GRES: Generalized 3D Referring Expression Segmentation (Supplementary Materials)

Changli Wu*
wuchangli@stu.xmu.edu.cn
Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China,
Xiamen University
Xiamen, Fujian, China

Yihang Liu*
liuyihang@stu.xmu.edu.cn
Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China,
Xiamen University
Xiamen, Fujian, China

Jiayi Ji
jjyxmu@gmail.com
Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China,
Xiamen University
Xiamen, Fujian, China

Yiwei Ma
yiweima@stu.xmu.edu.cn
Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China,
Xiamen University
Xiamen, Fujian, China

Haowei Wang
asucawang@tencent.com
Youtu Lab, Tencent,
Shanghai, China

Gen Luo
luogen@stu.xmu.edu.cn
Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China,
Xiamen University
Xiamen, Fujian, China

Henghui Ding
henghui.ding@gmail.com
Institute of Big Data,
Fudan University,
Shanghai, China

Xiaoshuai Sun
xssun@xmu.edu.cn
Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China,
Xiamen University
Xiamen, Fujian, China

Rongrong Ji†
rrji@xmu.edu.cn
Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China,
Xiamen University
Xiamen, Fujian, China

1 STATISTICAL ANALYSIS OF N_{seed} IN TSQ

To explore the optimal setting of the initial sparsity level in TSQ, denoted as the number of seed queries N_{seed}, we conducted a statistical analysis on the Multi3DRes dataset. To facilitate quantitative description, we introduce two concepts: coverage rate and repetition rate of seed queries. The Coverage Rate (CR) of seed queries indicates the number N_{ins}^{c} of instances containing seed queries as a percentage of the total number N_{ins} of instances in the scene, which can be formulated as:

$$CR = N_{ins}^{c}/N_{ins} \tag{1}$$

The Repetition Rate (RR) of seed queries indicates the percentage of seed queries that repeatedly cover the same instance with other queries, out of the total number N_{q}^{c} of seed queries covering instances, which can be formulated as:

$$RR = (N_{q}^{c} - N_{ins}^{c})/N_{q}^{c} \tag{2}$$

We conducted an analysis of the entire Multi3DRes Dataset, where we computed the coverage rate and repetition rate for different numbers N_{seed} of seed queries, as shown in Fig. I-(a). When a large number of seed queries N_{seed} are selected, it not only leads to higher coverage rate but also results in higher repetition rate. Fewer seed queries always bring a lower repetition rate, but a lower coverage rate may follow. We also calculated the coverage rate and repetition rate of seed queries for Ground Truth instances. As shown in Fig. I-(b), its trend follows the same pattern as Fig. I-(a). The statistical results indicate that $N_{seed} = 256$ is the optimal choice after balancing the relationship between coverage rate and repetition rate. This also validates the conclusion drawn in Sec. 5.3.2 that optimal performance is achieved when N_{seed} is set to 256.

2 TRADITIONAL 3D-RES

In this section, we evaluated our MDIN on the traditional 3D-RES task. we compare MDIN with existing 3D-RES works on ScanRefer [2] and Nr3D/Sr3D datasets of ReferIt3D [1].

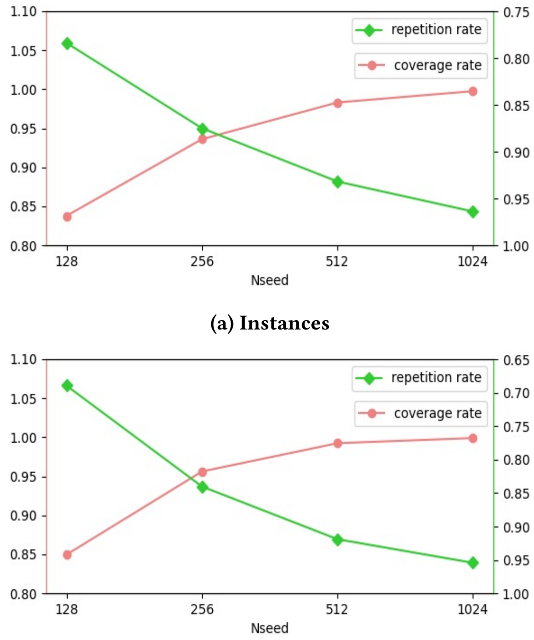
2.1 Datasets

Nr3D and Sr3D. Nr3D [1] (Natural Reference in 3D) consists of 41.5K human descriptions collected using a referring game [5]. It describes objects in 707 ScanNet scenes. Sr3D [1] (Spatial Reference in 3D) contains 83.5K synthetic descriptions. It categorizes spatial relations into 5 types: horizontal proximity, vertical proximity, between, allocentric and support, and then generates descriptions using language templates.

*Equal contribution.
†Corresponding author.

Table I: The results on ReferIt3D.

Method	easy			hard			View Dep			View Indep			**Overall**		
	0.25	0.5	mIoU	0.25	0.5	mIoU	0.25	0.5	mIoU	0.25	0.5	mIoU	**0.25**	**0.5**	**mIoU**
Nr3D															
TGNN [4]	29.2	22.3	21.0	22.5	19.6	17.4	22.2	18.1	17.0	27.6	22.4	20.3	25.7	20.9	19.1
3D-STMN [7]	47.9	31.9	32.6	35.4	20.0	23.0	37.7	21.1	24.3	43.5	28.4	29.4	41.5	25.8	27.6
MDIN (Ours)	**55.0**	**48.4**	**44.1**	**42.2**	**36.3**	**33.5**	**40.8**	**34.6**	**32.2**	**52.5**	**46.3**	**42.1**	**48.4**	**42.2**	**38.6**
Sr3D															
TGNN [4]	28.2	23.0	20.9	29.1	25.8	21.9	23.8	21.3	18.2	28.6	23.9	21.3	27.5	22.9	20.2
3D-STMN [7]	49.4	38.2	36.3	41.9	31.0	30.1	45.5	33.5	31.9	47.2	36.2	34.6	47.2	36.1	34.4
MDIN (Ours)	**58.9**	**53.2**	**48.1**	**51.1**	**46.8**	**42.5**	**53.6**	**48.7**	**44.2**	**56.7**	**51.4**	**46.5**	**56.6**	**51.3**	**46.4**



(a) Instances

(b) Ground Truth Instances

Figure I: The coverage rate and repetition rate of seed queries for all instances / Ground Truth instances in the scene.

2.2 Quantitative Comparison on ReferIt3D

We present the experimental results of our model on the ReferIt3D [1] benchmark in Tab. I. Unlike the the original setup of ReferIt3D, we refrained from using ground truth bounding boxes or masks as input in our experiments, thereby significantly heightening the level of difficulty. Despite this heightened difficulty, our model still achieved significant improvements. Notably, MDIN achieved an Acc@0.5 gain of 16.4 points and a mIoU gain of 11.0 points on Nr3D, as well as a 15.2-point increase in Acc@0.5 and a 12.0-point increase in mIoU on Sr3D.

Table II: Ablation study on number of stacked layers.

	Number of Stacked Layers	mIoU	Acc@0.25	Acc@0.5			
			Overall	zt w/ dis	st w/ dis	mt	Overall
*R*1	1	40.5	59.1	30.2	18.4	40.5	34.7
*R*2	3	45.6	65.2	38.1	25.3	45.7	41.9
*R*3	6	**47.5**	**67.0**	**47.9**	**29.5**	**46.8**	**44.7**
*R*4	9	46.3	65.3	39.4	25.1	45.8	42.1

Table III: Ablation study comparing text encoders.

	Text Encoder	mIoU	Acc@0.25	Acc@0.5			
			Overall	zt w/ dis	st w/ dis	mt	Overall
*R*1	BERT-base [3]	47.1	66.7	44.2	27.6	46.5	43.9
*R*2	BERT-large [3]	47.2	66.9	44.8	28.1	**46.9**	44.1
*R*3	RoBERTa [6]	**47.5**	**67.0**	**47.9**	**29.5**	46.8	**44.7**

3 MORE ABLATION STUDIES

3.1 Number of Stacked Layers in MDIN

We investigated the impact of changing the number of stacked layers in MDIN. As shown in Tab. II, performance gradually improves with increasing layers, reaching a peak at 6 layers, followed by a slight decline. Performance severely degrades when there is only one layer, indicating that refining layer by layer can enhance the model's reasoning ability. When the number of layers is excessive, gradients may become unstable, leading to a risk of training collapse. Therefore, selecting six layers strikes a balance that yields the best model performance.

3.2 The Textual Backbone

In Tab. III, we compare the effects of commonly used natural language encoders. It can be observed that that our method exhibits robustness regarding the choice of the NLP backbone. We achieve the optimal performance using Roberta [6].

4 MORE QUALITATIVE RESULTS

More qualitative comparison results of the highly competitive 3D-STMN and MDIN on the Multi3DRes dataset are illustrated in Fig. II and Fig. III. Fig. II primarily showcases the segmentation results in multi-targets scenarios. In multi-targets scenarios, due to the lack of decoupling capability, 3D-STMN either simply segments all instances with the same semantic category as the target (such as **(a)**, **(b)**), fails to segment all instances that match the description (such as **(c)**, **(d)**), or may even make semantic recognition errors (such as **(e)**, **(f)**). On the contrary, our MDIN accurately segments all instances that match the description. For cases where the target instances have small volumes or complex descriptions (such as **(g)**, **(h)**), MDIN also demonstrates its strong discriminative ability. However, 3D-STMN fails to comprehend highly complex language descriptions, leading to erroneous judgments.

Fig. III illustrates the segmentation results for single-target and zero-target scenarios. Benefiting from decoupled modeling, MDIN can capture information about objects in the scene and individually discern their compliance, thus segmenting the correct instances or making accurate predictions when no target instance are present. By contrast, 3D-STMN either suffers from semantic misinterpretation (e.g., **(a)**, **(b)**, **(e)**, **(g)**, **(h)**) or makes erroneous segmentation predictions by broadly leveraging semantics (e.g., **(c)**, **(d)**, **(f)**).

We also visualize the superpoints corresponding to the queries selected by the prediction heads of MDIN, as shown in the last column of Fig. II and Fig. III. The results indicate that in the presence of targets, the selected superpoints accurately reflect the positions of the target instances. Notably, for instances with simple geometric features, such as flat surfaces of table, the selected superpoints directly reflect the geometric shapes of the target instances. In the absence of targets, however, no query contains any target instances, hence no superpoints are selected.

REFERENCES

[1] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. 2020. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 422–440.

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

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[4] Pin-Hao Huang, Han-Hung Lee, Hwann-Tzong Chen, and Tyng-Luh Liu. 2021. Text-guided graph neural networks for referring 3d instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 1610–1618.

[5] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 787–798.

[6] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[7] Changli Wu, Yiwei Ma, Qi Chen, Haowei Wang, Gen Luo, Jiayi Ji, and Xiaoshuai Sun. 2023. 3D-STMN: Dependency-Driven Superpoint-Text Matching Network for End-to-End 3D Referring Expression Segmentation. *arXiv preprint arXiv:2308.16632* (2023).

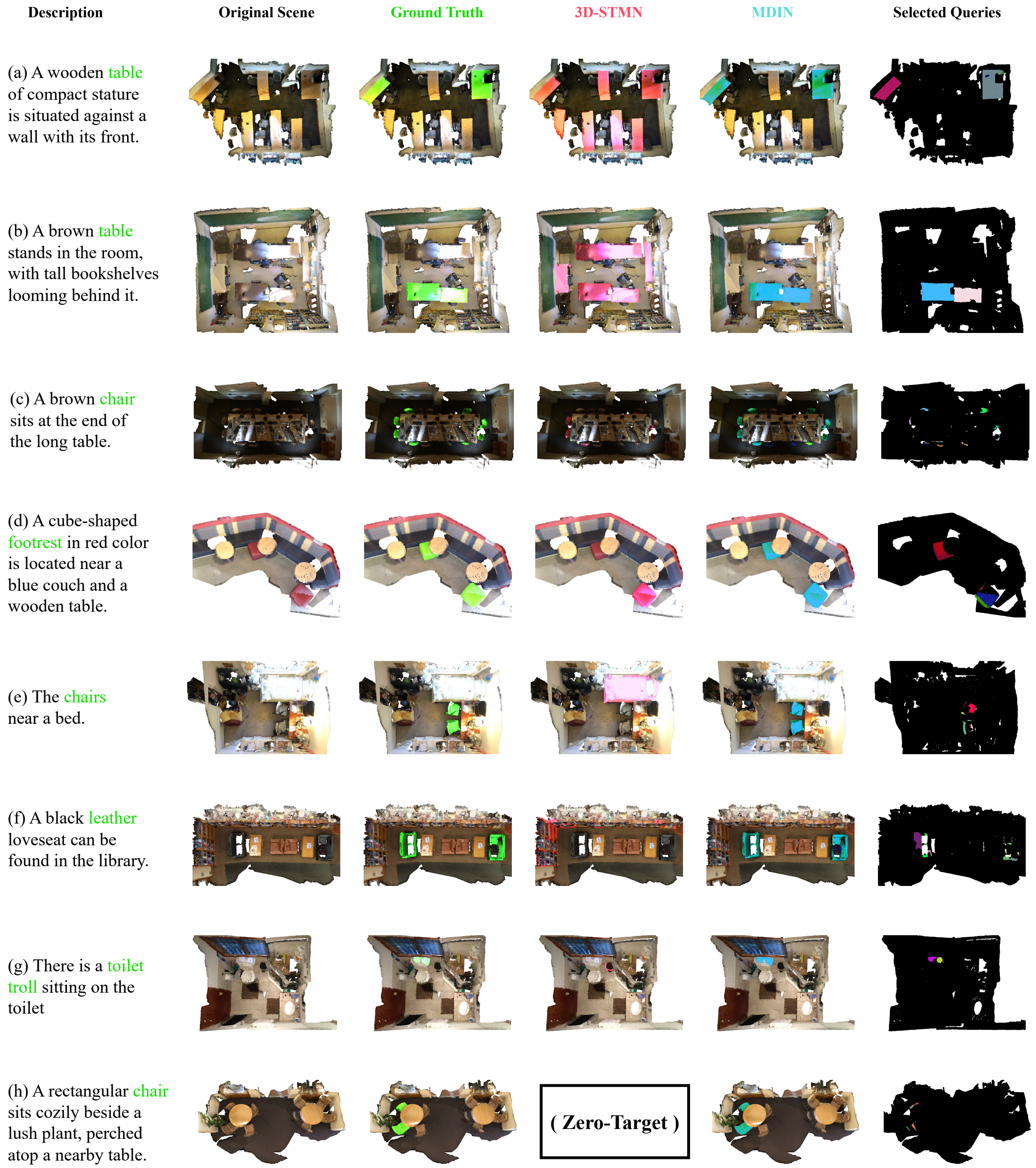

Description
Original Scene
Ground Truth
3D-STMN
MDIN
Selected Queries
(a) A wooden table of compact stature is situated against a wall with its front.
(b) A brown table stands in the room, with tall bookshelves looming behind it.
(c) A brown chair sits at the end of the long table.
(d) A cube-shaped footrest in red color is located near a blue couch and a wooden table.
(e) The chairs near a bed.
(f) A black leather loveseat can be found in the library.
(g) There is a toilet troll sitting on the toilet
(h) A rectangular chair sits cozily beside a lush plant, perched atop a nearby table.
(Zero-Target)

Figure II: Qualitative comparison between the proposed MDIN and 3D-STMN on multi-targets cases. Zoom in for the best view.

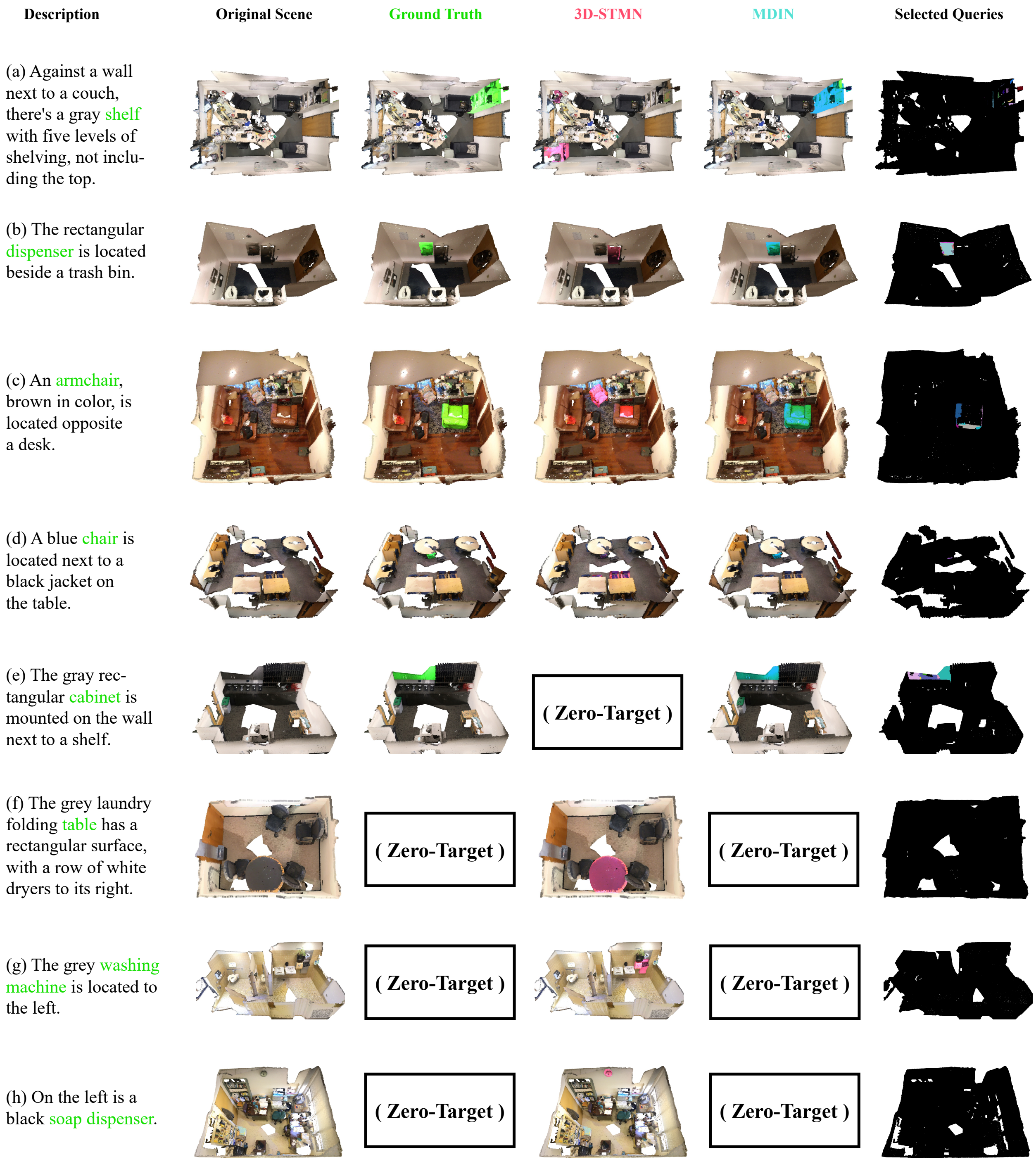

Description
Original Scene
Ground Truth
3D-STMN
MDIN
Selected Queries
(a) Against a wall next to a couch, there's a gray shelf with five levels of shelving, not including the top.
(b) The rectangular dispenser is located beside a trash bin.
(c) An armchair, brown in color, is located opposite a desk.
(d) A blue chair is located next to a black jacket on the table.
(e) The gray rectangular cabinet is mounted on the wall next to a shelf.
(Zero-Target)
(f) The grey laundry folding table has a rectangular surface, with a row of white dryers to its right.
(Zero-Target)
(Zero-Target)
(g) The grey washing machine is located to the left.
(Zero-Target)
(Zero-Target)
(h) On the left is a black soap dispenser.
(Zero-Target)
(Zero-Target)

Figure III: Qualitative comparison between the proposed MDIN and 3D-STMN on single/zero-target cases. Zoom in for the best view.