

# *Seeing is Not Believing: Adversarial Natural Object Optimization for Hard-Label 3D Scene Attacks*

Daizong Liu, and Wei Hu

Wangxuan Institute of Computer Technology, Peking University

# Challenges

- Real-world 3D applications mainly focus on tackling more complicated 3D scene cases, which contain hundreds of individual objects with distinct spatial relationships. However, existing 3D attackers solely perturb a single object to fool the classifier without considering any geometry-aware spatial relationship. Directly perturbing the whole 3D scene following them would destroy the spatial relations across different objects and the topological structures of the entire scene. In real-world applications, existing LVLM attackers generally rely on the detailed prior knowledge of the model to generate effective perturbations. Moreover, these attacks are task-specific, leading to significant costs for designing perturbation.

# Challenges

- Existing 3D attacks are realized with access to the model's gradient or its output confidence. However, practical 3D applications share no knowledge of weight parameters, hyper-parameters, the training procedure, the original training data, etc., to users. Their only feedback is the predicted result of a given input. Therefore, how to craft/optimize perturbations in 3D scenes that enable the wrong prediction by only querying the 3D models would be a severe security problem.

# Motivation

- Firstly, previous 3D attackers generally add noise to coordinates on point clouds globally, which is not suitable for the scene attack as they would destroy complex geometric structures. Therefore, we propose to design and add imperceptible local noisy patterns into the scene case to make them natural to human eyes.

# Motivation

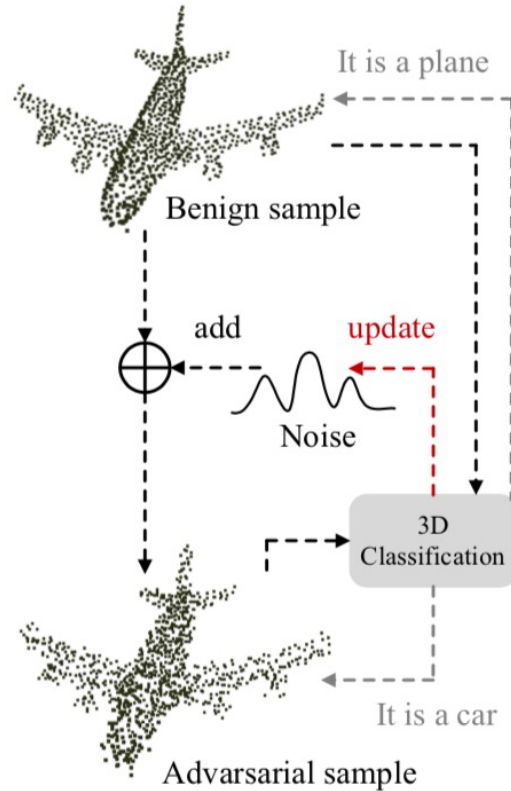
- Secondly, previous 3D attackers generally optimize perturbations for each sample individually, requiring substantial resources and time. Therefore, it is demanded to design an effective and robust noisy pattern that is universally adversarial to any sample. Whenever the universal pattern is injected into any input sample, the model will mispredict it to a wrong label.

# Motivation

- Thirdly, without resorting to model details, we propose a gradient estimation strategy to guide the optimization of the noisy pattern. Existing strategies simply follow the Monte Carlo estimation to approximate the direction of gradients, which equally treats a set of random noise for estimation. However, since not all noises point towards the accurate optimization direction, we need to assign weights to them so as to adaptively adjust gradient directions for effective estimation.



# Motivation



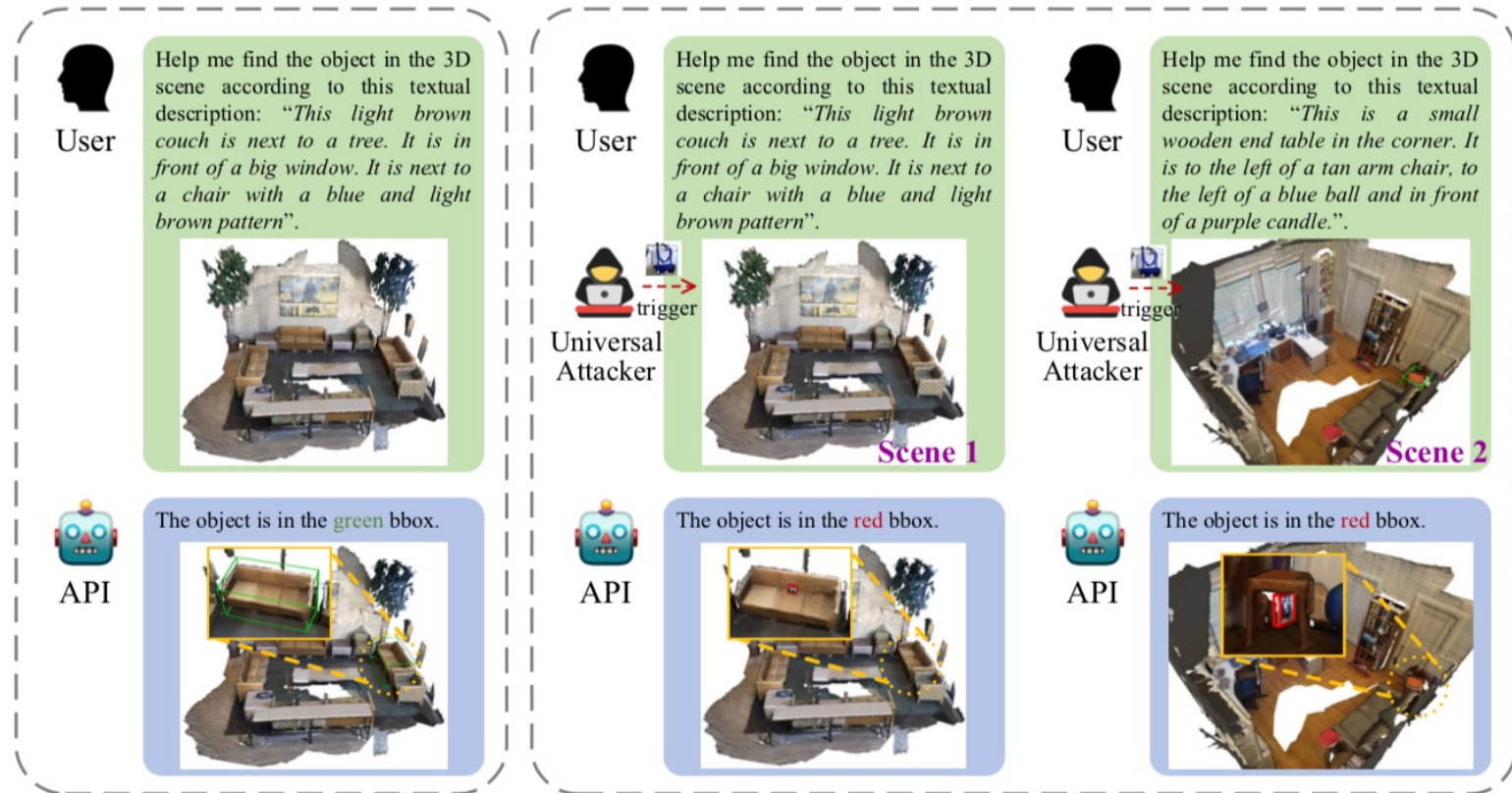
(a) Existing 3D digital attacks on the simple object classification.



Normal Mode

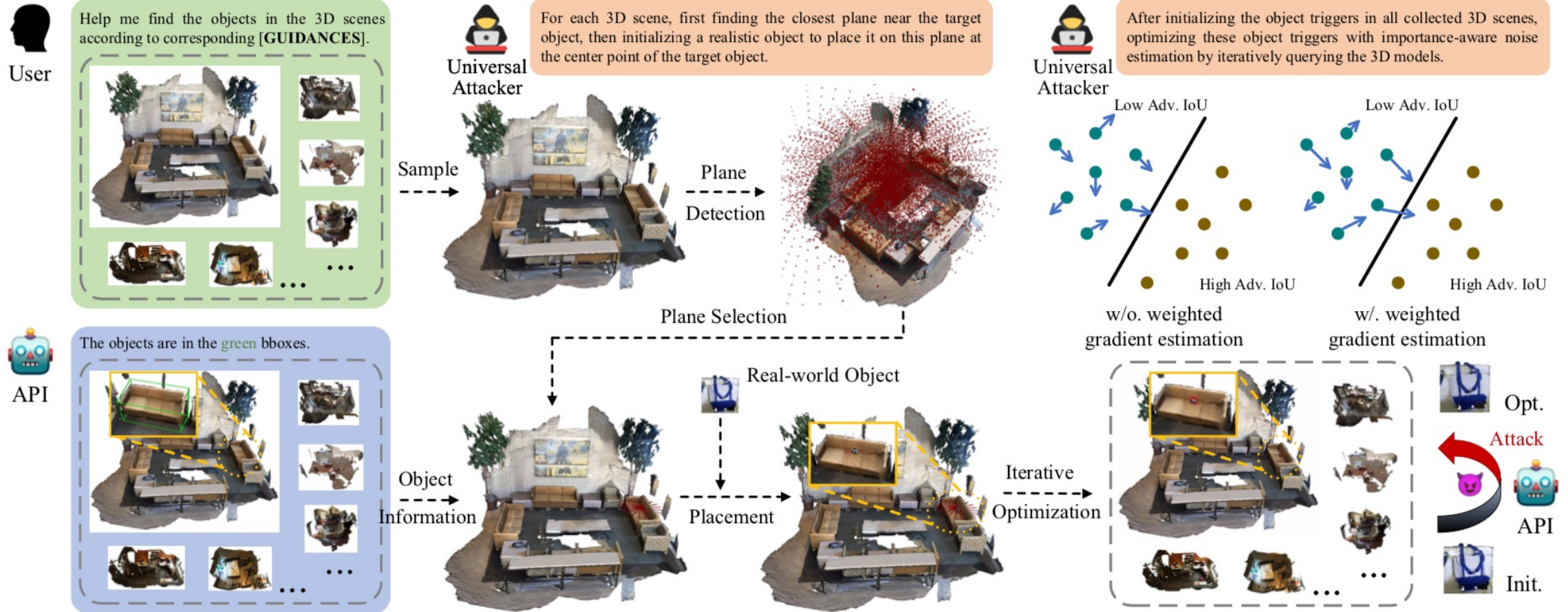


Attack Mode



(b) Our practical yet challenging attacks on complicated 3D scenes.

# Pipeline







# Quantitative Results

Table 1. Our attack performance against seven 3D grounding models on the ScanRefer dataset.

3D Model	Type	Unique		Multiple	
		Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5
ScanRefer [6]	Origin	67.64	46.19	32.06	21.26
	Attack	<b>15.27</b>	<b>10.45</b>	<b>8.63</b>	<b>6.86</b>
ReferIt3D [1]	Origin	53.80	37.50	21.00	12.80
	Attack	<b>12.56</b>	<b>9.74</b>	<b>6.48</b>	<b>5.12</b>
SAT [88]	Origin	73.21	50.83	37.64	25.16
	Attack	<b>17.49</b>	<b>12.37</b>	<b>11.63</b>	<b>8.20</b>
3DTrans. [92]	Origin	77.16	58.47	38.38	28.70
	Attack	<b>18.03</b>	<b>14.72</b>	<b>10.25</b>	<b>8.94</b>
3D-SPS [62]	Origin	81.63	64.77	39.48	29.61
	Attack	<b>19.29</b>	<b>17.87</b>	<b>12.94</b>	<b>10.43</b>
ViL3DRel [9]	Origin	81.58	68.62	40.30	30.71
	Attack	<b>18.74</b>	<b>17.02</b>	<b>13.56</b>	<b>10.33</b>
EDA [84]	Origin	85.76	68.57	49.13	37.64
	Attack	<b>20.61</b>	<b>17.54</b>	<b>15.38</b>	<b>12.82</b>

Table 2. Our attack performance against seven 3D grounding models on Nr3D and Sr3D datasets.

3D Model	Type	Nr3D Dataset			Sr3D Dataset		
		Overall	Easy	Hard	Overall	Easy	Hard
ScanRefer [6]	Origin	34.2	41.0	23.5	35.7	41.2	25.4
	Attack	<b>11.1</b>	<b>14.3</b>	<b>7.9</b>	<b>11.6</b>	<b>14.7</b>	<b>8.2</b>
ReferIt3D [1]	Origin	35.6	43.6	27.9	40.8	44.7	31.5
	Attack	<b>11.6</b>	<b>14.8</b>	<b>8.5</b>	<b>13.6</b>	<b>15.4</b>	<b>10.1</b>
SAT [88]	Origin	49.2	56.3	42.4	57.9	61.2	50.0
	Attack	<b>15.9</b>	<b>18.5</b>	<b>13.7</b>	<b>18.3</b>	<b>19.0</b>	<b>17.6</b>
3DTrans. [92]	Origin	40.8	48.5	34.8	51.4	54.2	44.9
	Attack	<b>14.2</b>	<b>16.1</b>	<b>12.5</b>	<b>17.0</b>	<b>17.9</b>	<b>14.8</b>
3D-SPS [62]	Origin	51.5	58.1	45.1	62.6	56.2	65.4
	Attack	<b>16.2</b>	<b>18.4</b>	<b>13.3</b>	<b>19.1</b>	<b>18.7</b>	<b>19.5</b>
ViL3DRel [9]	Origin	64.4	70.2	57.4	72.8	74.9	67.9
	Attack	<b>21.3</b>	<b>24.1</b>	<b>19.6</b>	<b>24.8</b>	<b>25.3</b>	<b>21.9</b>
EDA [84]	Origin	53.7	58.2	46.1	68.5	70.3	62.9
	Attack	<b>16.4</b>	<b>18.2</b>	<b>13.8</b>	<b>23.0</b>	<b>23.7</b>	<b>19.1</b>



# Robust to Defenses

Table 3. Our attack performance against existing defense methods on the ScanRefer dataset.

Attacked Model	Defense Method	Overall	
		Acc@0.25	Acc@0.5
3D-SPS [62]	No Defense	<b>15.05</b>	<b>12.38</b>
	Sampling	15.49	12.52
	Fine-tuning	34.52	26.48
	Anomaly Det.	21.73	15.66
ViL3DRel [9]	No Defense	<b>14.80</b>	<b>12.67</b>
	Sampling	15.31	12.97
	Fine-tuning	33.24	27.05
	Anomaly Det.	23.17	16.79
EDA [84]	No Defense	<b>15.96</b>	<b>13.73</b>
	Sampling	16.02	13.84
	Fine-tuning	34.64	29.13
	Anomaly Det.	22.99	17.68



# Efficiency

Table 4. Experimental validation on plane detection.

Strategy	Point Percentage	Averaged Angle Error	Time
Hough Transform	99.8%	$0.6^\circ$	0.3s

Table 7. Complexity and efficiency comparison.

Metric	[7]		[24]		[40]		Ours	
	Train	Test	Train	Test	Train	Test	Train	Test
Hour (h)	6.8	0.2	8.9	0.4	7.3	0.3	9.4	0.4
Memory (GB)	24.6	11.0	31.5	11.3	23.8	11.2	28.3	11.5

# Ablation Study

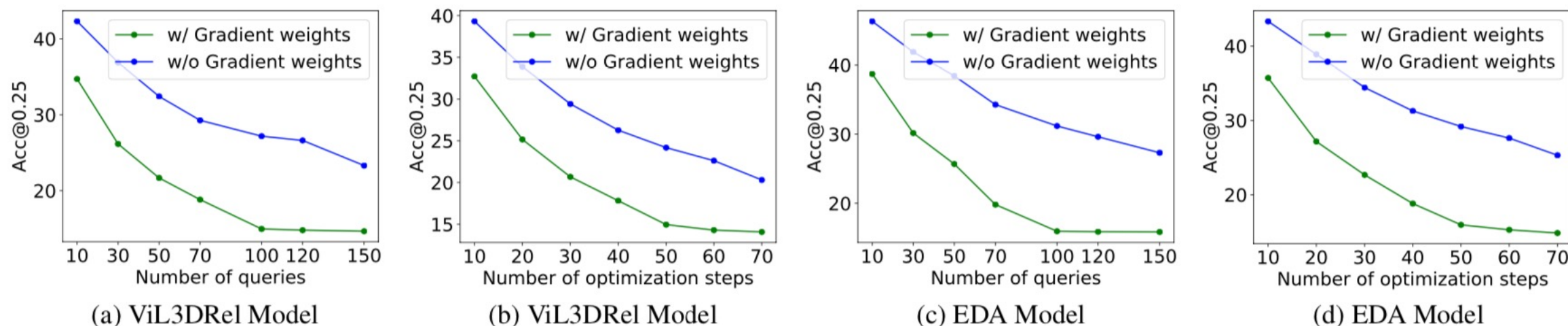





Figure 4. Ablation on the numbers of attack queries and optimization steps. We evaluate the Acc@0.25 metrics on the ScanRefer dataset.



# Ablation Study

Table 5. Perturbation sizes on different types of object patterns on ScanRefer, Nr3D/Sr3D datasets.

Object Type	Attacked Model	ScanRefer Overall		ScanRefer			Nr3D			Sr3D		
		Acc@0.25	Acc@0.5	Cham.	Hausd.	Norm	Cham.	Hausd.	Norm	Cham.	Hausd.	Norm
	3D-SPS [62]	15.05	12.38	0.0003	0.0018	0.0007	0.0001	0.0013	0.0008	0.0001	0.0010	0.0006
	ViL3DRel [9]	14.80	12.67	0.0002	0.0027	0.0009	0.0002	0.0022	0.0010	0.0001	0.0015	0.0008
	EDA [84]	15.96	13.73	0.0003	0.0021	0.0010	0.0002	0.0014	0.0009	0.0002	0.0011	0.0009
	3D-SPS [62]	15.28	12.47	0.0008	0.0025	0.0016	0.0005	0.0019	0.0012	0.0004	0.0013	0.0010
	ViL3DRel [9]	12.39	11.53	0.0009	0.0023	0.0018	0.0004	0.0016	0.0011	0.0005	0.0015	0.0012
	EDA [84]	15.14	13.49	0.0009	0.0026	0.0021	0.0006	0.0018	0.0017	0.0005	0.0016	0.0014
	3D-SPS [62]	12.75	10.02	0.0005	0.0016	0.0012	0.0003	0.0014	0.0009	0.0002	0.0012	0.0008
	ViL3DRel [9]	11.83	11.40	0.0003	0.0017	0.0011	0.0002	0.0013	0.0008	0.0002	0.0011	0.0007
	EDA [84]	13.59	11.94	0.0006	0.0020	0.0014	0.0004	0.0014	0.0010	0.0003	0.0011	0.0008





# Ablation Study

Table 6. Ablations on different *sizes* of object pattern, IoU *threshold* of indicator function, and sample *number* for achieving universality.

Attacked Model	Object Size	Overall		Threshold	Overall		Number	Overall	
		Acc@0.25	Acc@0.5		Acc@0.25	Acc@0.5		Acc@0.25	Acc@0.5
3D-SPS [62]	$\lambda = 0.1$	18.34	14.96	$\gamma = 0.5$	17.29	14.43	100	24.28	17.54
	$\lambda = 0.2$	15.05	12.38	$\gamma = 0.6$	15.05	12.38	200	15.05	12.38
	$\lambda = 0.5$	14.72	12.11	$\gamma = 0.7$	15.92	13.17	300	14.69	12.16
ViL3DRel [9]	$\lambda = 0.1$	16.93	14.65	$\gamma = 0.5$	17.03	14.74	100	26.45	17.81
	$\lambda = 0.2$	14.80	12.67	$\gamma = 0.6$	14.80	12.67	200	14.80	12.67
	$\lambda = 0.5$	14.49	12.38	$\gamma = 0.7$	16.39	14.15	300	14.52	12.33
EDA [84]	$\lambda = 0.1$	18.25	15.06	$\gamma = 0.5$	18.76	15.89	100	26.17	18.35
	$\lambda = 0.2$	15.96	13.73	$\gamma = 0.6$	15.96	13.73	200	15.96	13.73
	$\lambda = 0.5$	15.97	13.54	$\gamma = 0.7$	17.51	14.92	300	15.48	13.40



# Qualitative Results



User

Help me find the object in the 3D scene according to this textual description: *“There is a printer on top of the desk. it sits to the right of the desk in between the two windows”*.

Help me find the object in the 3D scene according to this textual description: *“This is a red chair in the middle of the room. it is facing left”*.

Help me find the object in the 3D scene according to this textual description: *“The bathroom vanity has a white sink inside a black counter. it's also underneath the mirror”*.

Help me find the object in the 3D scene according to this textual description: *“There is a blue chair. it is at the table and has another table to its back”*.

Help me find the object in the 3D scene according to this textual description: *“A light brown rope. it is in between with many clothes”*.

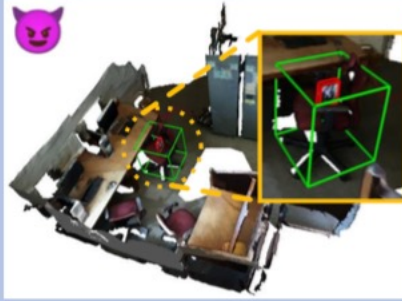


API

The object is in the **green** **red** bbox.



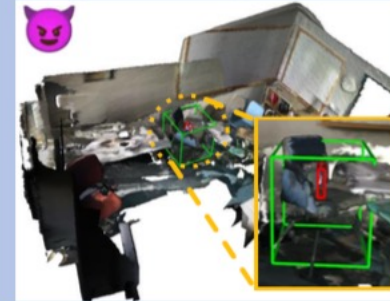
The object is in the **green** **red** bbox.



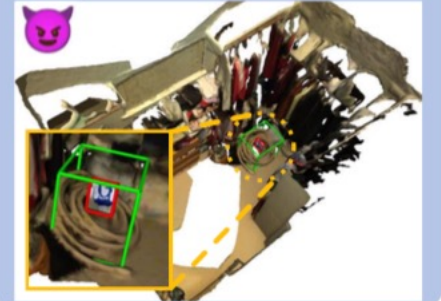
The object is in the **green** **red** bbox.



The object is in the **green** **red** bbox.



The object is in the **green** **red** bbox.





北京大学  
PEKING UNIVERSITY



# Thanks!

Email: [dzliu@stu.pku.edu.cn](mailto:dzliu@stu.pku.edu.cn)