香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

CVPR
SEATTLE, WA
JUNE 17-21, 2024

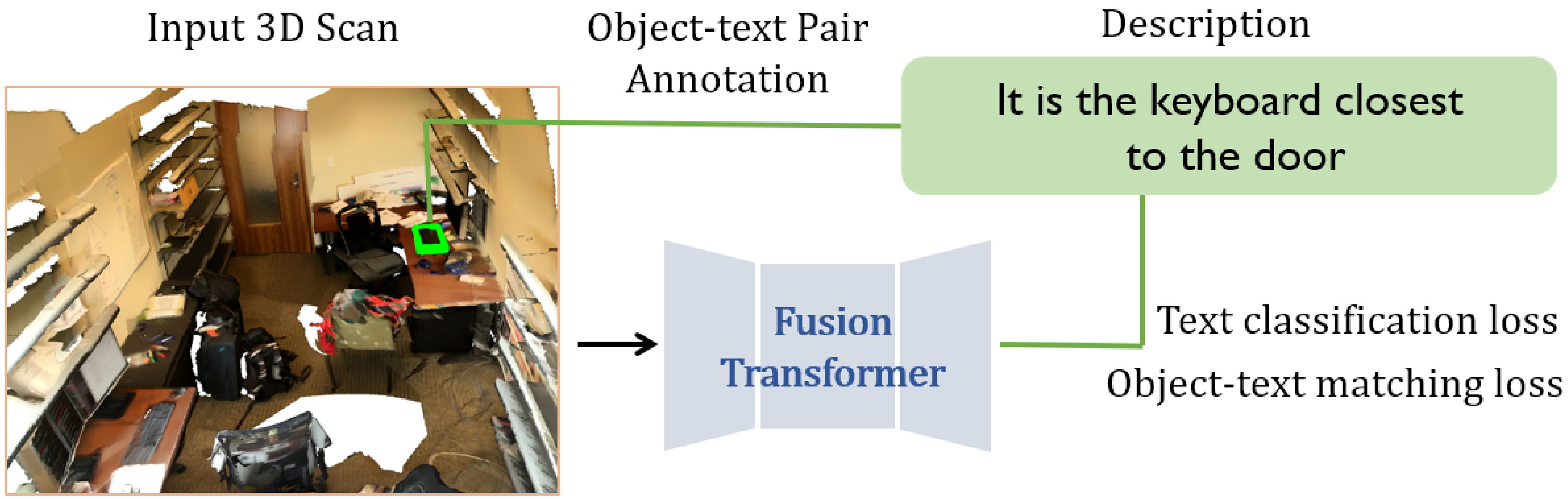
Visual Programming for Zero-shot Open-Vocabulary 3D Visual Grounding

Zhihao Yuan[1,2], Jinke Ren[1,2],
Chun-Mei Feng[3], Hengshuang Zhao[4], Shuguang Cui[2,1], Zhen Li[2,1]

[1]The Future Network of Intelligence Institute, The Chinese University of Hong Kong (Shenzhen)
[2]School of Science and Engineering, The Chinese University of Hong Kong (Shenzhen)
[3]IHPC, A*STAR, Singapore
[4]The University of Hong Kong

Background

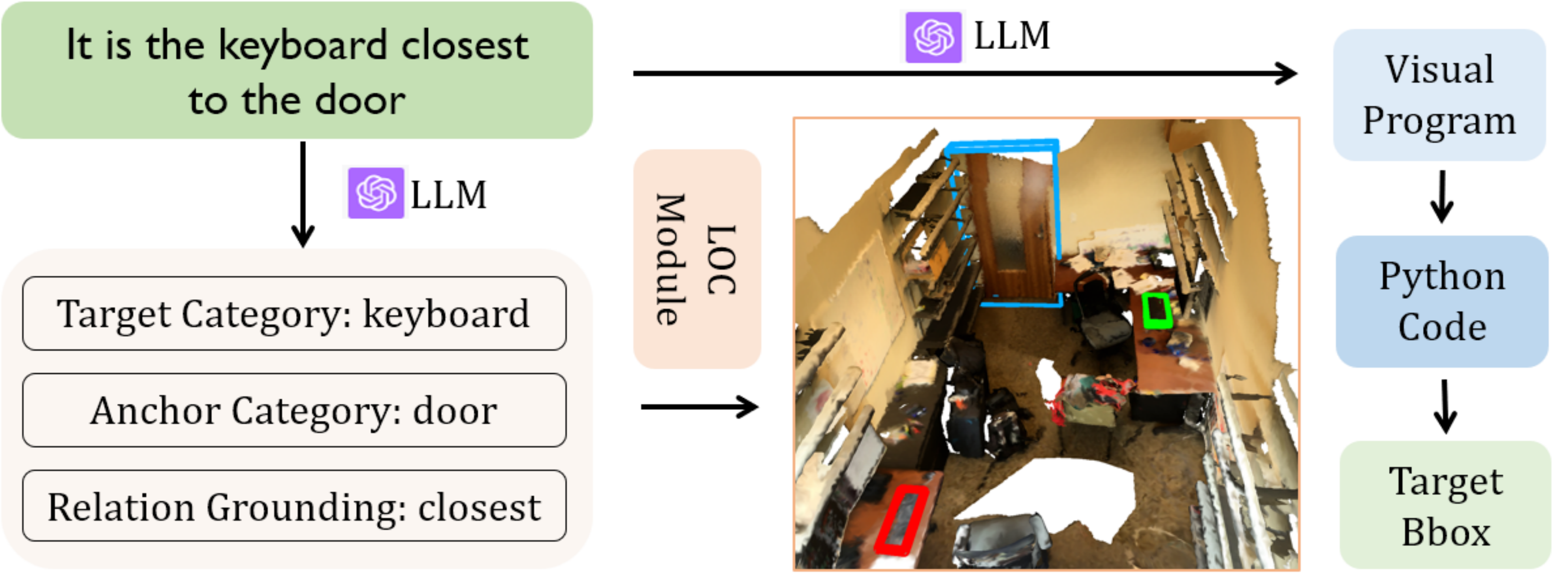
- Problem
  - Explore zero-shot 3DVG using LLMs, without the need of object-text pair annotation.
  - Solve relationships between objects explicitly.


Input 3D Scan
Object-text Pair Annotation
Description
It is the keyboard closest to the door
Fusion Transformer
Text classification loss
Object-text matching loss

a) Supervised 3D Visual Grounding

Background

- Problems
  - Explore zero-shot 3DVG using LLMs, without the need of object-text pair annotation
  - Solve relationships between objects explicitly



It is the keyboard closest to the door

LLM

Target Category: keyboard

Anchor Category: door

Relation Grounding: closest

LOC Module

LLM

Visual Program

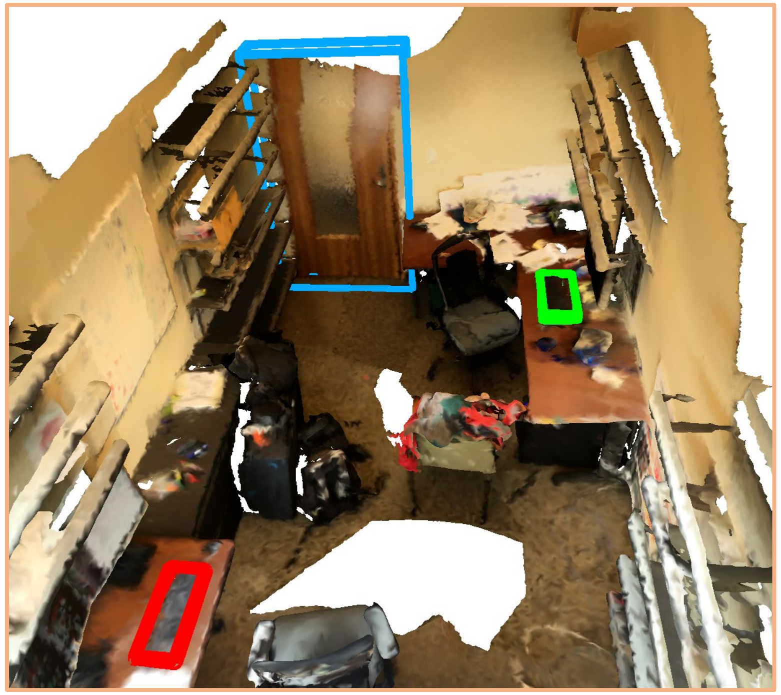
Python Code

Target Bbox

b) Zero-shot 3D Visual Grounding

Method

- Dialog with LLM: A Naive Approach
  - Detect all objects in the scene, describe object's location and size.
  - Given the text prompt to ChatGPT to find the correct object.



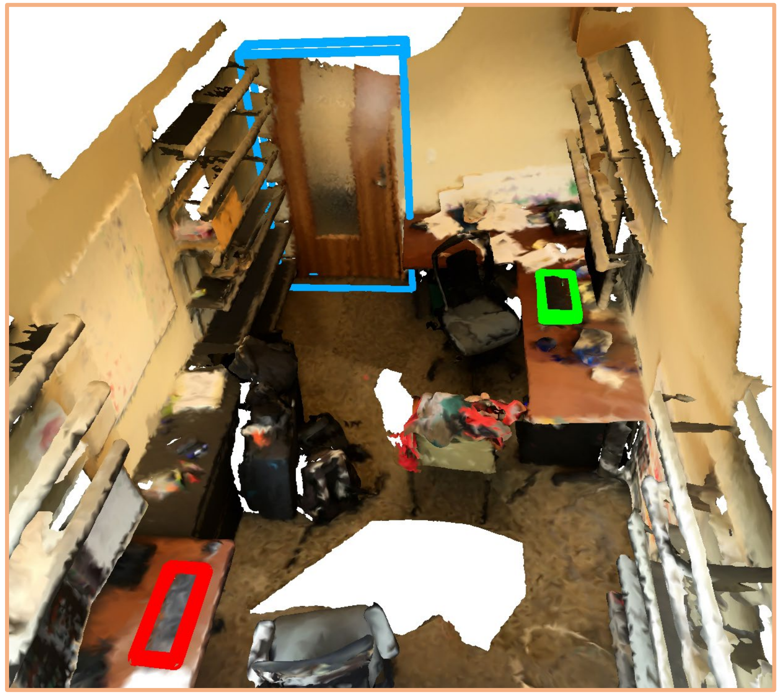
Suppose you are a person standing in a room. You need to find a keyboard it is closest to the door.

Of course, I can help you find an object in a room based on its description. Please provide me with the details of the object you're looking for, and I'll do my best to assist you in locating it.

Room Information:
Object 1 is a door located at (-0.65, 2.35, 1.05).
Object 2 is a desk located at (0.68, 1.30, 0.39).
…
Object 26 is a keyboard located at (-0.65, -1.06, 0.65).

Method

- Dialog with LLM: A Naive Approach
  - Detect all objects in the scene, describe object's location and size.
  - Given the text prompt to ChatGPT to find the correct object.



Suppose you are a person standing in a room. You need to find a keyboard it is closest to the door.

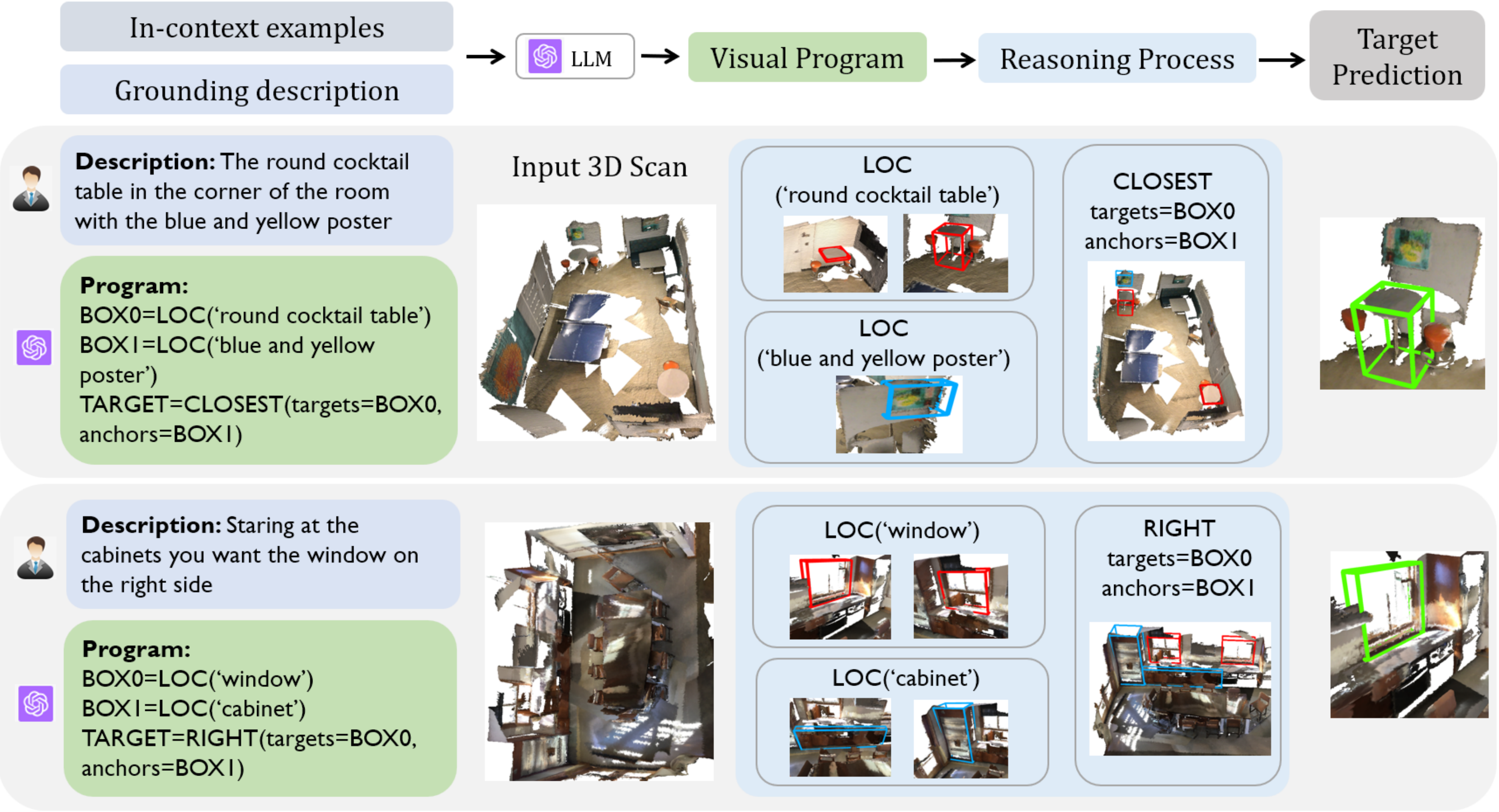
Of course, I can help you find an object in a room based on its description. Please provide me with the details of the object you're looking for, and I'll do my best to assist you in locating it.

Room Information:
Object 1 is a door located at (-0.65, 2.35, 1.05).
Object 2 is a desk located at (0.68, 1.30, 0.39).
…
Object 26 is a keyboard located at (-0.65, -1.06, 0.65).

The keyboard closest to the door is Object 9, as it has a shorter distance of approximately 2.01 units, compared to Object 26, which has a distance of approximately 3.44 units. So, the correct object ID is Object 9.

Method

- ➢ 3D visual programming approach.



In-context examples

Grounding description

→ LLM → Visual Program → Reasoning Process → Target Prediction

Description: The round cocktail table in the corner of the room with the blue and yellow poster

Program:
BOX0=LOC('round cocktail table')
BOX1=LOC('blue and yellow poster')
TARGET=CLOSEST(targets=BOX0, anchors=BOX1)

Input 3D Scan

LOC('round cocktail table')

LOC('blue and yellow poster')

CLOSEST
targets=BOX0
anchors=BOX1

Description: Staring at the cabinets you want the window on the right side

Program:
BOX0=LOC('window')
BOX1=LOC('cabinet')
TARGET=RIGHT(targets=BOX0, anchors=BOX1)

LOC('window')

LOC('cabinet')

RIGHT
targets=BOX0
anchors=BOX1

Method

- Relation modules
  - Addressing view-dependent relations: A shift to 2D egocentric view.
  - Addressing view-independent relations: using 3D coordinates.

View-independent	*near, close, next to, far, above, below, under, top, on, opposite, middle*
View-dependent	*front, behind, back, right, left, facing, leftmost, rightmost, looking, across, between*
Functional	*min, max, size, length, width*

Table 1. Common relations in 3DVG.

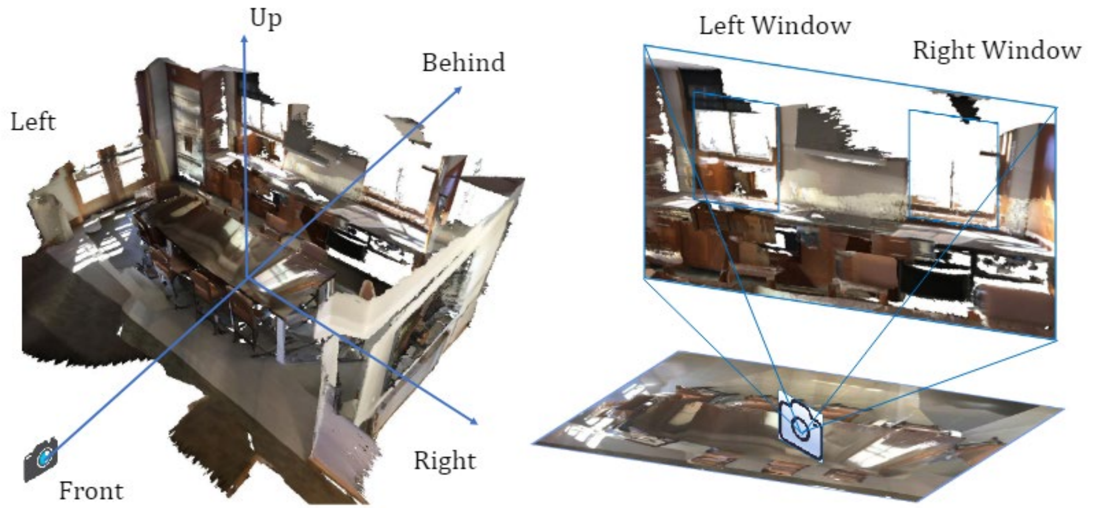


Figure 3. Addressing view-dependent relations: A shift to 2D egocentric view.

Method

- ➢ LOC module: extend the scope of existing 3D object detectors into open-vocabulary scenarios.

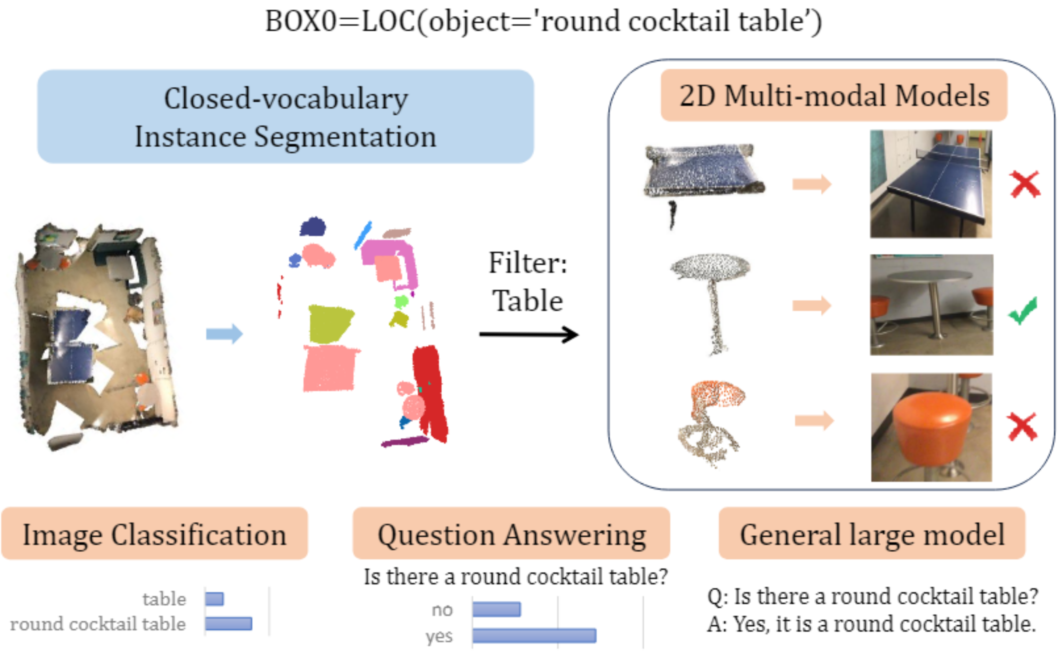

BOX0=LOC(object='round cocktail table')
Closed-vocabulary Instance Segmentation
Filter: Table
2D Multi-modal Models
Image Classification
table
round cocktail table
Question Answering
Is there a round cocktail table?
no
yes
General large model
Q: Is there a round cocktail table?
A: Yes, it is a round cocktail table.

Figure 4. Illustration of the language-object correlation module.

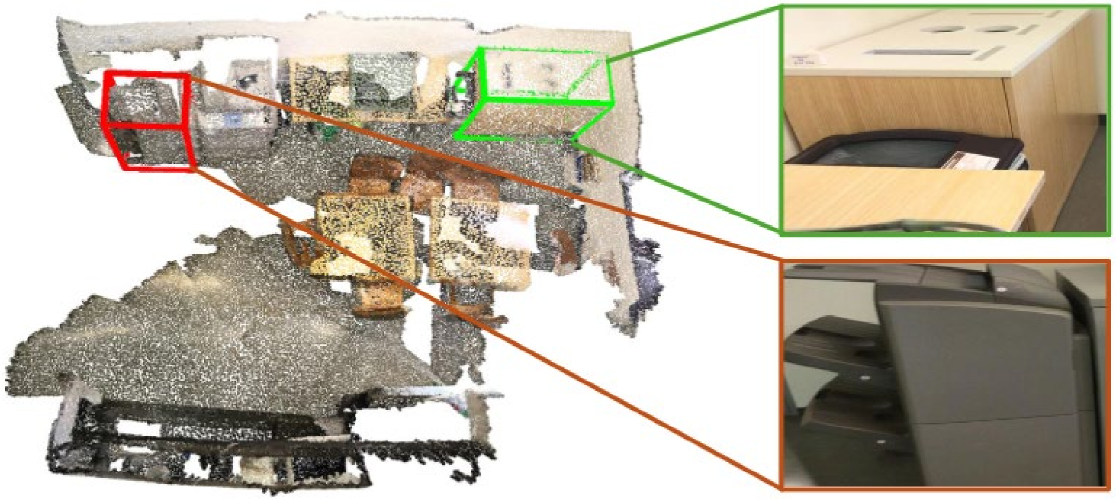


Figure 2. Open-vocabulary query: *A brown closed cabinet. It is broad and spacious.* Supervised approach can only predict the *cabinet*, while our approach uses image models to distinguish objects with similar geometry (red and green boxes).

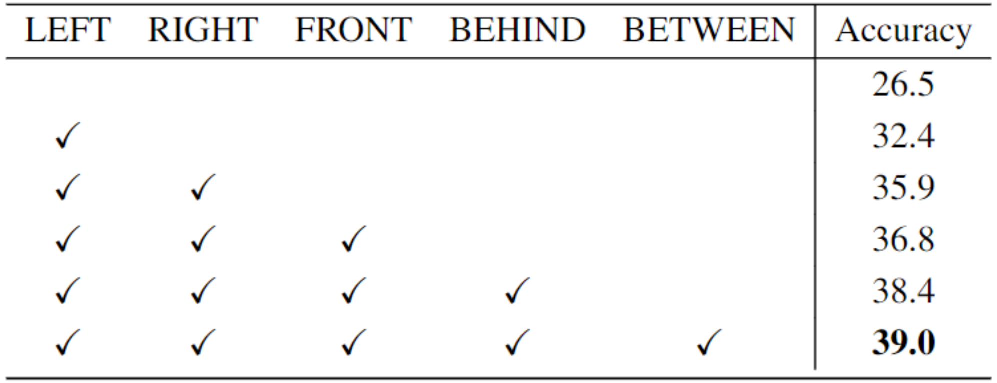
Quantitative Analysis

- ScanRefer dataset
  - Our zero-shot approach can outperform some supervised baselines
  - Moreover, our zero-shot approach outperforms the approaches that only utilize the 3D or 2D information in the LOC module.

		Unique		Multiple		Overall	
Methods	Supervision	Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5
ScanRefer [4]	fully	65.0	43.3	30.6	19.8	37.3	24.3
TGNN [17]	fully	64.5	53.0	27.0	21.9	34.3	29.7
InstanceRefer [60]	fully	77.5	66.8	31.3	24.8	40.2	32.9
3DVG-Transformer [65]	fully	81.9	60.6	39.3	28.4	47.6	34.7
BUTD-DETR [20]	fully	84.2	66.3	46.6	35.1	52.2	39.8
LERF [23]	-	-	-	-	-	4.8	0.9
OpenScene [34]	-	20.1	13.1	11.1	4.4	13.2	6.5
Ours (2D only)	-	32.5	27.8	16.1	14.6	20.0	17.6
Ours (3D only)	-	57.1	49.4	25.9	23.3	33.1	29.3
Ours	-	**63.8**	**58.4**	**27.7**	**24.6**	**36.4**	**32.7**

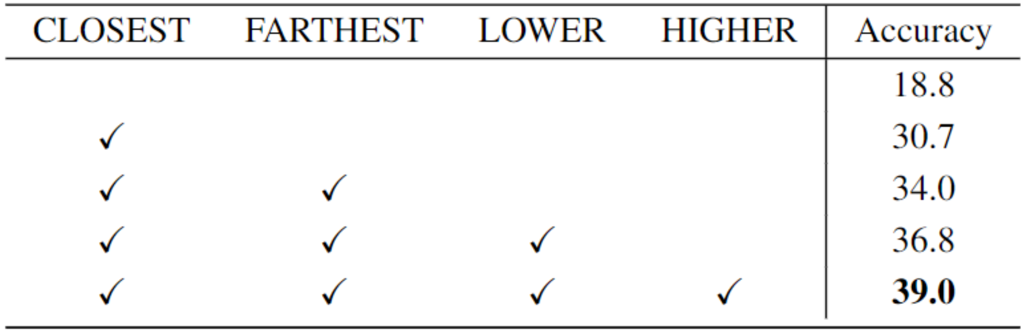
Quantitative Analysis

- We ablate different relation modules in to analyze their impact.
- Our framework has strong adaptability for a spectrum of 3D and 2D perception models.



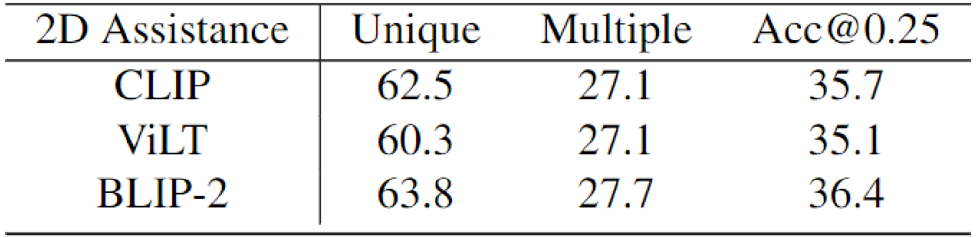
LEFT	RIGHT	FRONT	BEHIND	BETWEEN	Accuracy
					26.5
✓					32.4
✓	✓				35.9
✓	✓	✓			36.8
✓	✓	✓	✓		38.4
✓	✓	✓	✓	✓	**39.0**

Table 5. Ablation study of different **view-dependent** modules.



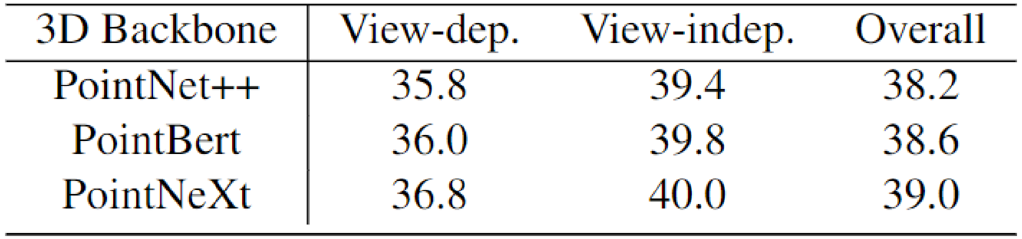
CLOSEST	FARTHEST	LOWER	HIGHER	Accuracy
				18.8
✓				30.7
✓	✓			34.0
✓	✓	✓		36.8
✓	✓	✓	✓	**39.0**

Table 6. Ablation study of different **view-independent** modules.



2D Assistance	Unique	Multiple	Acc@0.25
CLIP	62.5	27.1	35.7
ViLT	60.3	27.1	35.1
BLIP-2	63.8	27.7	36.4

Table 7. Ablation study on different 2D models.

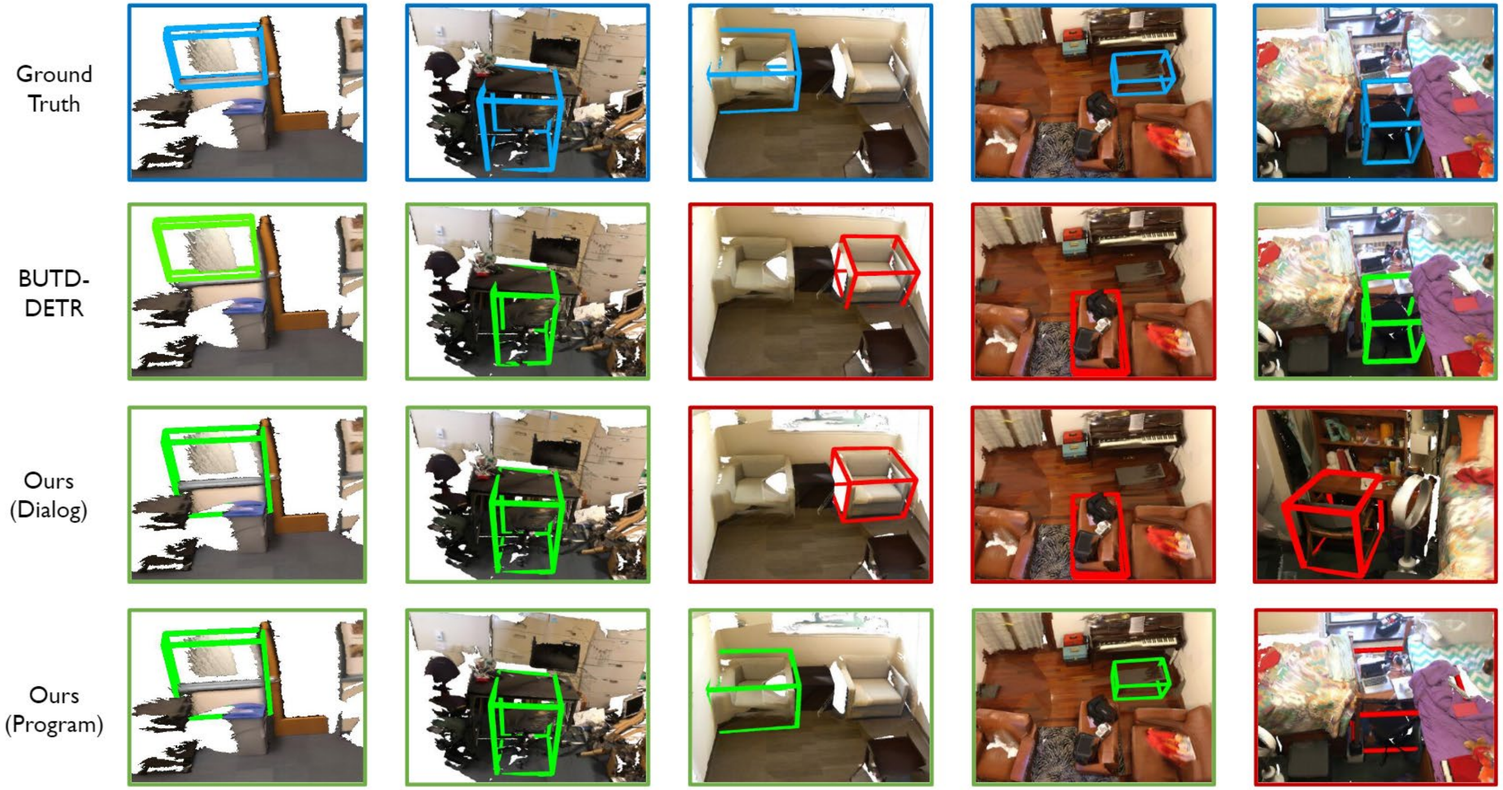


3D Backbone	View-dep.	View-indep.	Overall
PointNet++	35.8	39.4	38.2
PointBert	36.0	39.8	38.6
PointNeXt	36.8	40.0	39.0

Table 8. Ablation study on different 3D backbones.

Qualitative Analysis

- Visualization results of 3D visual grounding.



It is a window. It is located above a recycle bin that has a blue top.
(a)

The rolling office chair. The chair is under the desk.
(b)

There is a square beige armchair. It is left of a square table.
(c)

This is a brown piano bench. It is in front of the piano.
(d)

A desk chair is pushed into a small computer desk. The chair has wheels .
(e)

Conclusion

- Introduces a novel zero-shot 3DVG approach, removing the need for extensive annotations.
- Enhances localization accuracy with relation modules and a language-object correlation module.
- Experiments on ScanRefer and Nr3d datasets show the method outperforms several supervised baselines.