

Multimodal Applications for NLP

CMPT-413
March 26, 2025

Multimodal Applications for NLP

- Visual grounding in 3D
- Aligning text and 3D representations
- Text-to-scene generation
- NLP in biodiversity applications

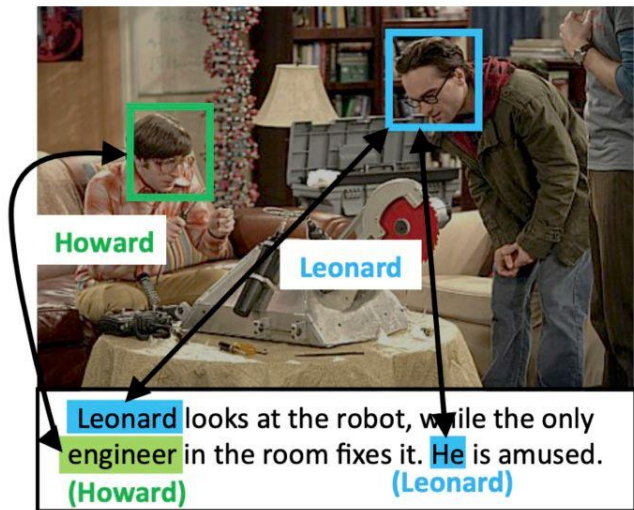


Visual Grounding in 3D



What is symbol grounding?

- Connecting **linguistic symbols** to their **meaning**
- Connecting words and sentences to what they represent



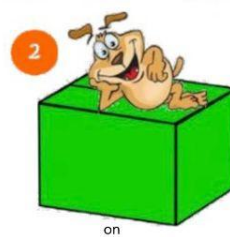
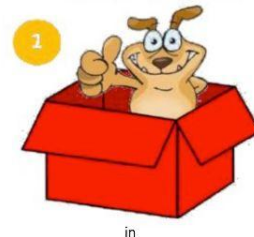
Linking people in videos with “their” names using coreference resolution [Ramanathan et al, 2014]

Actions



running

Spatial relations

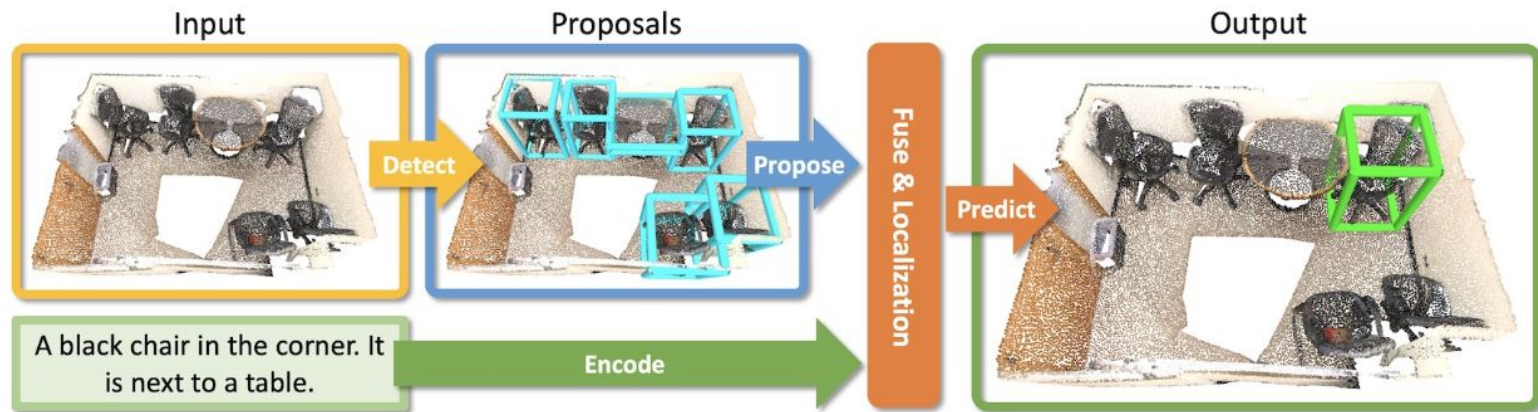


How can you describe to someone
where objects are in a scene?



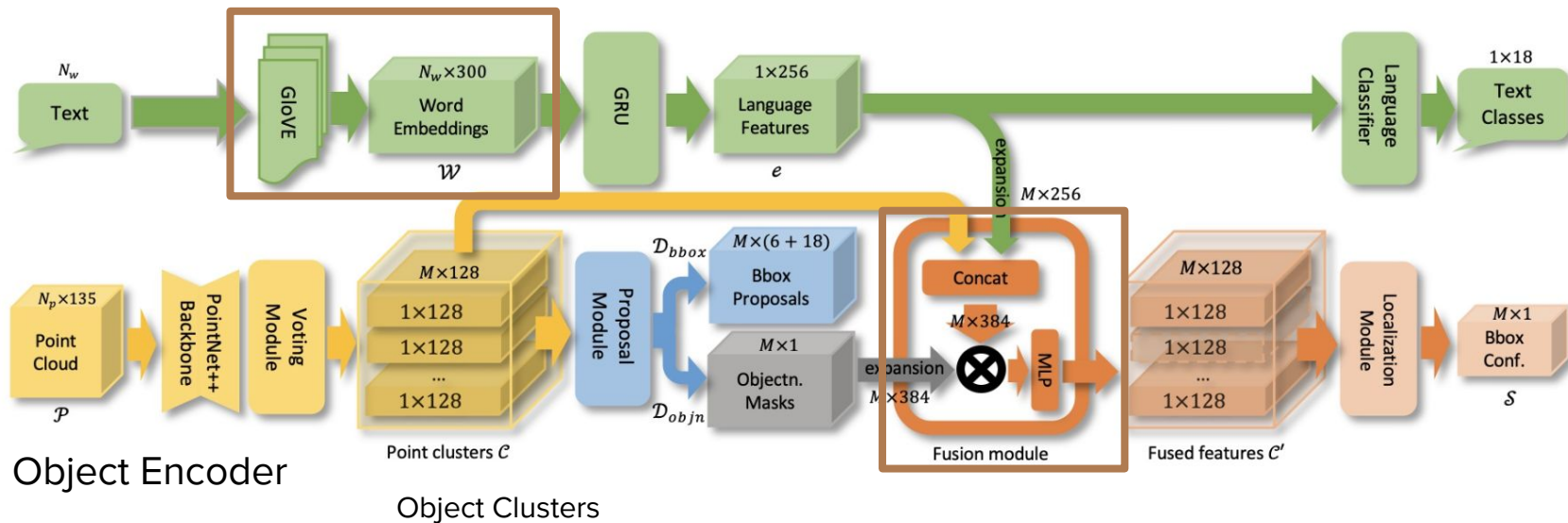
This is the dining chair which is closest to the wall with the TV mounted on it.

Grounding through encoding text and visual components and fusing together

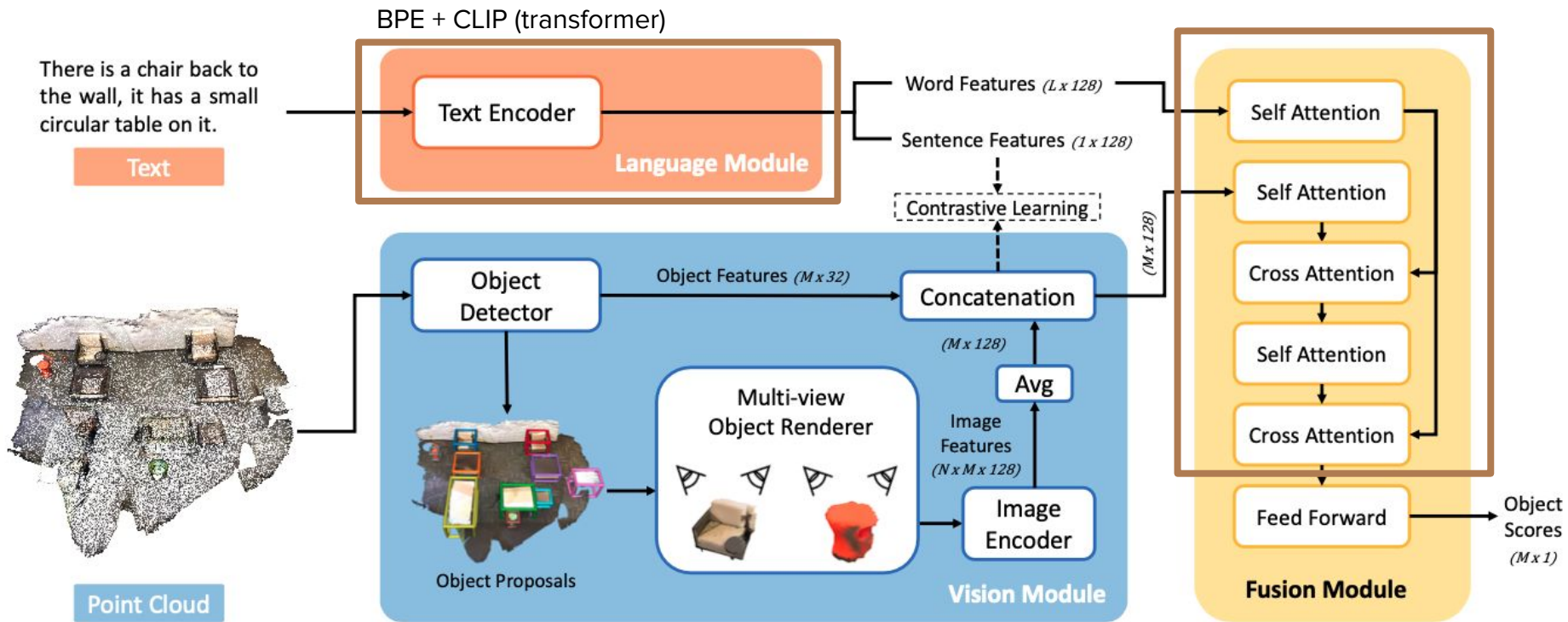


ScanRefer (Chen et al., 2020)

Text Encoder



ScanRefer (Chen et al., 2020)



Multi3DRefer (Zhang et al., 2023)

ScanRefer: This is a microwave. It's black in color and it's behind the coca cola soda cans. [sic]

Nr3D: This cabinet is below the kitchen sink, next to the dishwasher.

SceneVerse: The microwave is to the left of the bag of coffee beans and toaster.

How do existing datasets represent these?



3D-GRAND: This is a sleek silver wine cooler. The wine cooler features a cylindrical shape and a smooth, metallic finish. It is located below the pendant lamps, next to the side cabinet. The wine cooler serves as a sophisticated appliance for wine enthusiasts in the room.

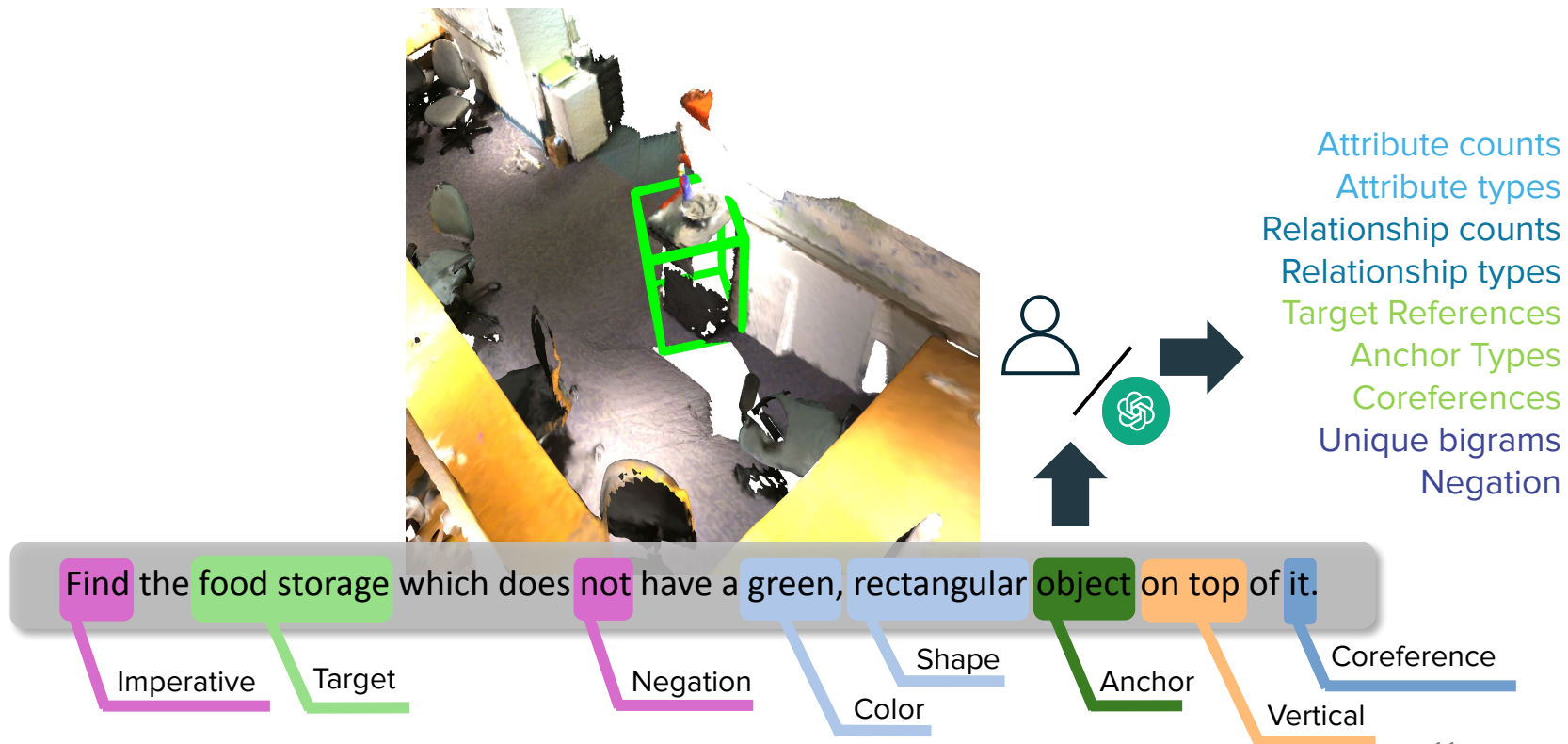
ScanScribe: The white microwave is standing on the kitchen counter. It has the same color as the rectangular white door, the white chair, and the rectangular white heater.

ViGiL3D: A Linguistically Diverse Dataset for 3D Grounding

Austin Wang, ZeMing Gong, Angel X. Chang

ACL Rolling Review Submission (2025)

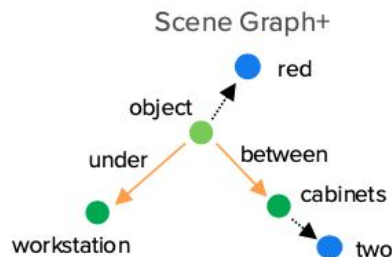
Anatomy of a Grounding Prompt



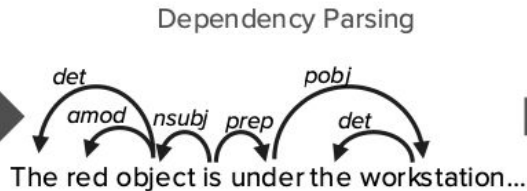
Analysis Pipeline

“You are given a description of an object that someone is supposed to find in a scene. We would like to identify the objects, attributes, and relationships in the following text...”

The red object is under the workstation and between the two cabinets.



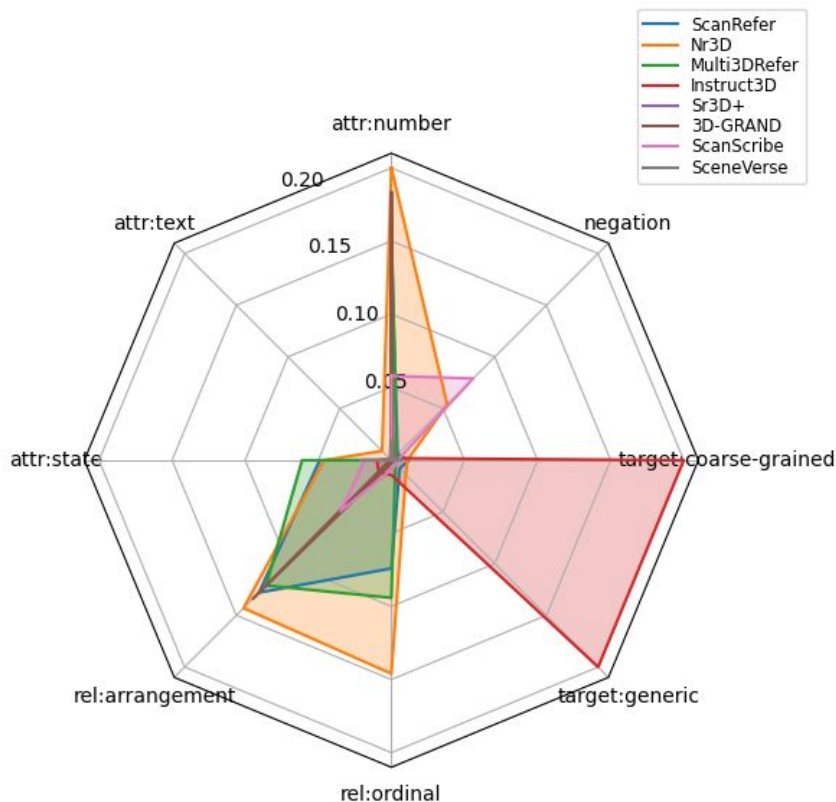
Attribute counts
Attribute types
Relationship counts
Relationship types
Target References
Anchor Types



Coreferences
Unique bigrams
Negation

(Wang et al., 2024)

Prior datasets lack key linguistic patterns



	# attr	# rel	bigrams	
ScanRefer	1.90	2.33	0.20	human-annotated
Nr3D	1.16	2.22	0.27	
Multi3DRefer	1.73	2.02	0.28	
Instruct3D	1.40	1.43	0.27	
Sr3D+	0.05	1.00	0.02	LLM
3D-GRAND	5.81	2.81	0.05	
ScanScribe	6.04	3.55	0.10	
SceneVerse	0.41	1.33	0.16	

ViGiL3D facilitates robust evaluation through linguistic diversity

Generic Class



This circular object is supported by a black mini fridge.

Categorical Class



This is a white appliance for washing dishes automatically.

Attribute: Shape



In the center of the room is a hexagonal conference table for meetings.

Attribute: Text Label



Grab the trash cans far from the wall with the number, "200," written on one of its posters.

Relationship: Far



This is the chair with four wheels somewhat near the corner opposite from that with the blue-and-yellow picture.

Relationship: Comparison



There is a row of washing machines in the room. It is the one farthest from the sink.

Anchor: Agent-Based



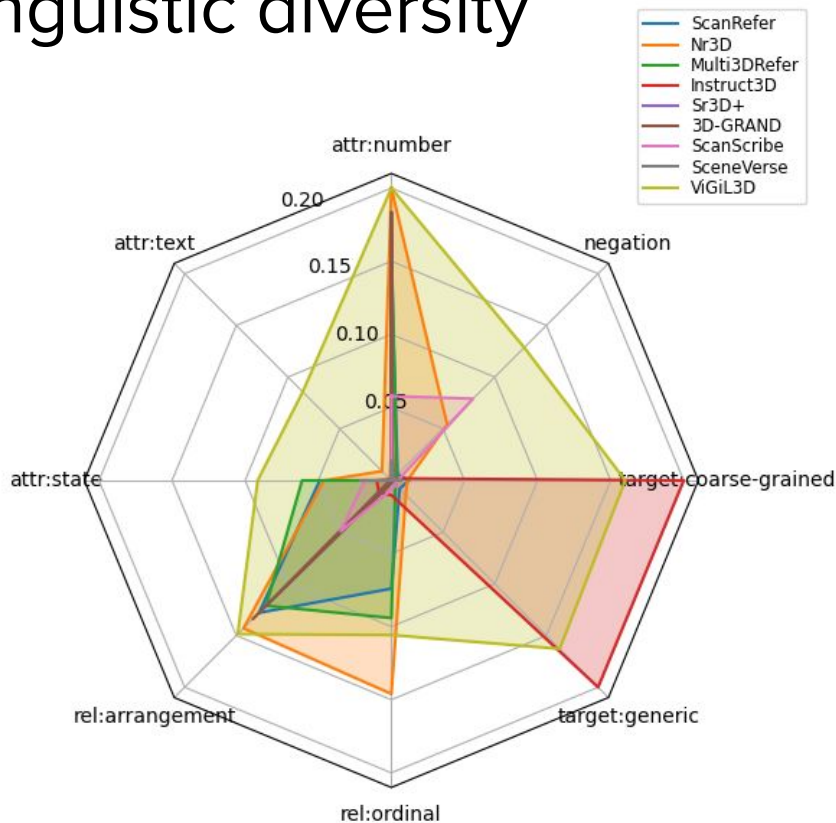
Face the wall to the left of the door as you enter. This is the third object from the left against the wall.

Negation



This is the bag not on the rug. It is made of a firm cloth material.

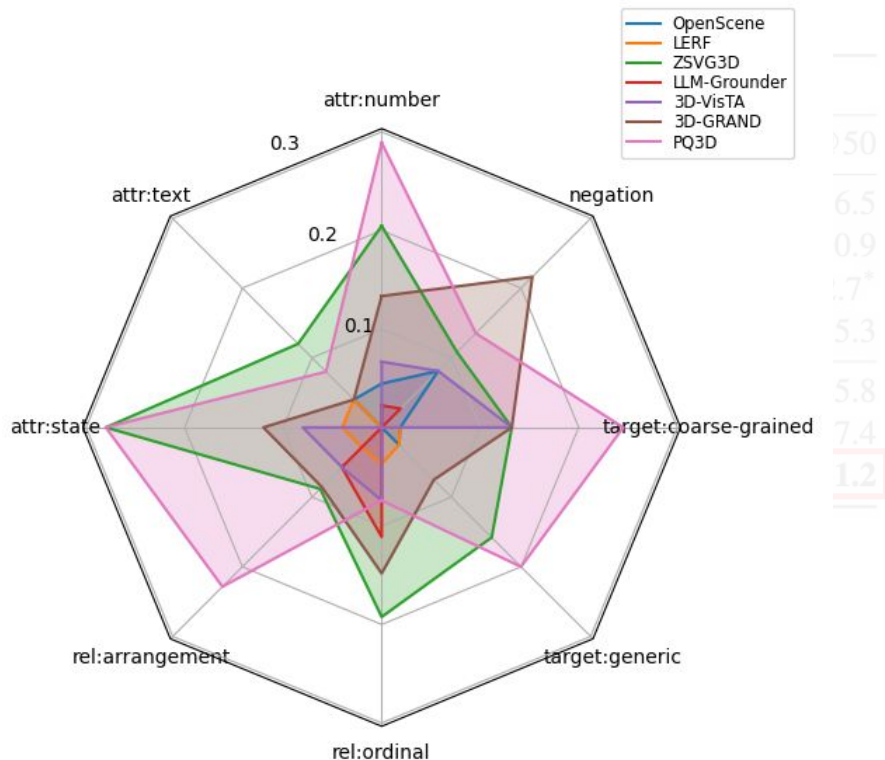
ViGiL3D facilitates robust evaluation through linguistic diversity



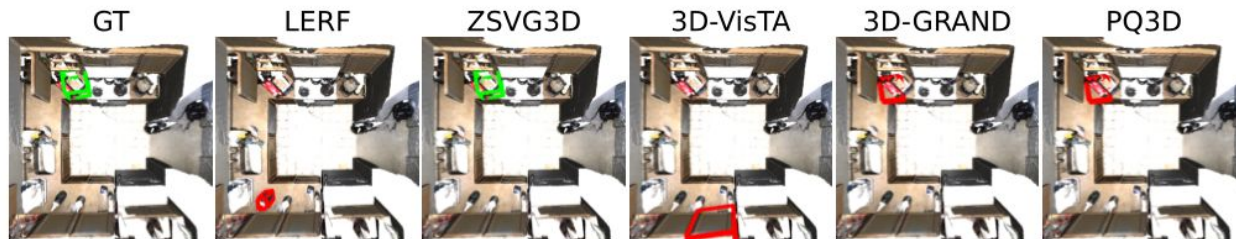
	# attr	# rel	bigrams	
ScanRefer	1.90	2.33	0.20	human-annotated
Nr3D	1.16	2.22	0.27	
Multi3DRefer	1.73	2.02	0.28	
Instruct3D	1.40	1.43	0.27	
Sr3D+	0.05	1.00	0.02	LLM
3D-GRAND	5.81	2.81	0.05	
ScanScribe	6.04	3.55	0.10	
SceneVerse	0.41	1.33	0.16	
ViGiL3D	1.62	1.82	0.45	

Models perform worse on ViGiL3D

	Acc/GT	Acc@25
OpenScene	2.1	1.7
LERF	2.5	2.1
ZSVG3D	18.9	8.5
LLM-Grounder	2.5	7.1
3D-VisTA	14.2	15.8
3D-GRAND	17.9	15.8
PQ3D	26.2	10.8



Grab the box on the counter that appears taller than the rest.



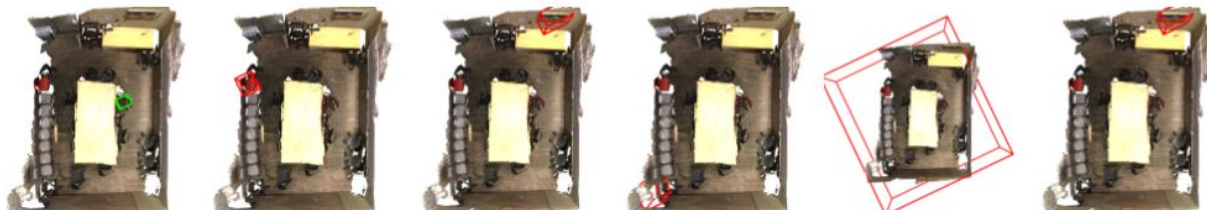
When facing the radiator, the metal rail is directly on your left.



This circular object is supported by a white mini fridge.

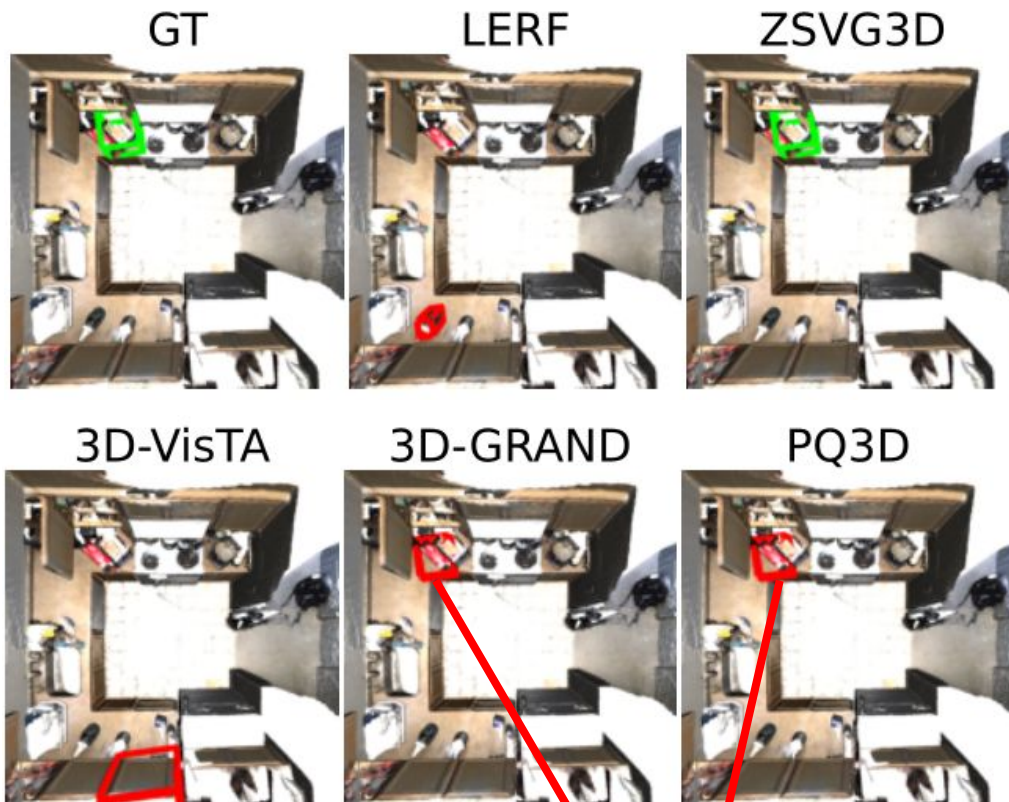


There are two prominently red objects in the room. This is the one not next to the wall.





Grab the box on the counter that appears **taller** than the rest.



Wrong box



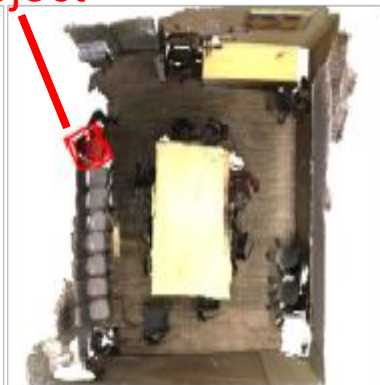
There are two prominently **red** objects in the room. This is the one **not** next to the wall.

GT

Wrong red
object

LERF

ZSVG3D



3D-VisTA

3D-GRAND

PQ3D



ViGiL3D enables future development of 3DVG models through a more linguistically diverse benchmark.



Read the paper
here!

Understand your benchmarks (Nick Vincent's talk!), and pay attention to the distribution of the data you are scaling.