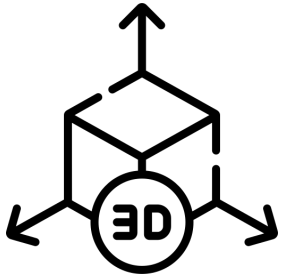
Assessing spatial reasoning in MLLMs

Workshop on Multimodal Spatial Intelligence

Angel Xuan Chang

2026-06-03

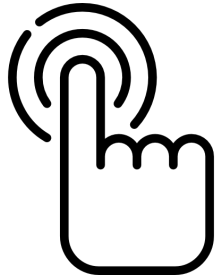
Language +

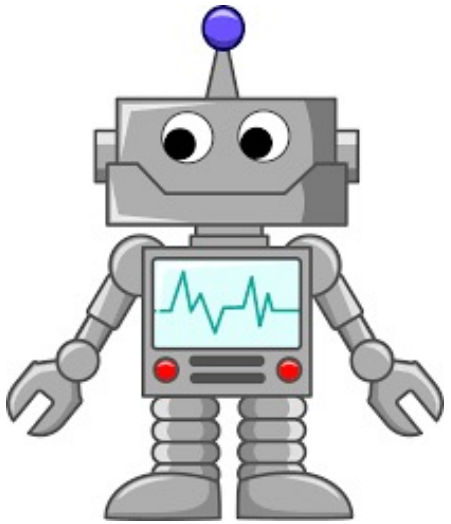


3D representations and understanding

+

Large scale interactive environments





helpful AI assistants

Some problems my group work on

- Language understanding in 3D scenes
- 3D scene and shape generation
- Modeling interactive (e.g. articulated) objects

Some problems my group work on

- Language understanding in 3D scenes
- 3D scene and shape generation
- Modeling interactive (e.g. articulated) objects

How well do recent models understand language in 3D scenes?

- How well does recent 3D grounding models handle diverse linguistic patterns?
- How well does recent 3D LLMs understand spatial relations?
- How well does recent VLMs understand 3D spaces (through video)?

Visual grounding in 3D

https://daveredrum.github.io/ScanRefer/

Input

Point Cloud (Scene)

This is a bed with blue sheets. It is to the left of a desk.

Description

ScanRefer: 3D Object Localization in RGB-D Scans using Natural Language

[Chen et al. ECCV 2020]

Visual grounding in 3D

https://daveredrum.github.io/ScanRefer/

Input

Point Cloud (Scene)

This is a bed with blue sheets. It is to the left of a desk.

Description

Output

3D Bounding Boxes

Visual grounding in 3D

https://daveredrum.github.io/ScanRefer/

Input

Point Cloud (Scene)

*This is a **bed** with blue sheets. It is to the left of a desk.*

Description

Output

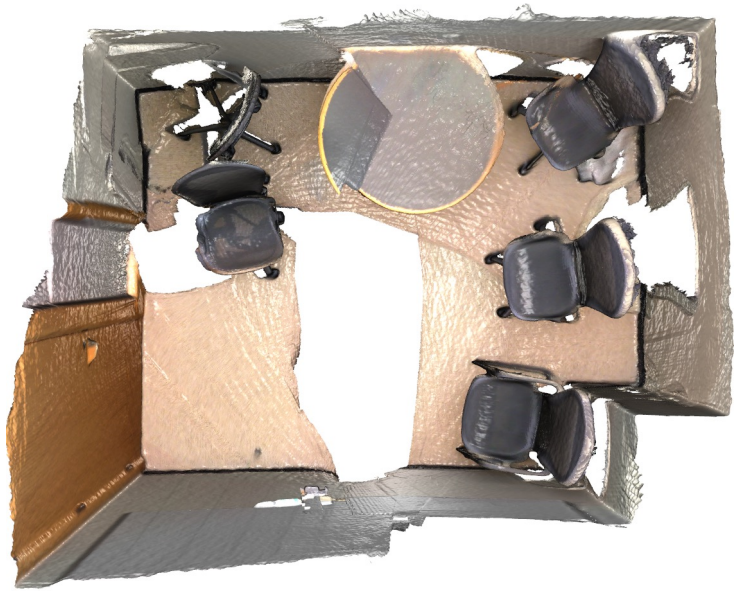
3D Bounding Boxes

Multi3DRefer: Grounding Text Description to Multiple 3D Objects

[Zhang et al. ICCV 2023]

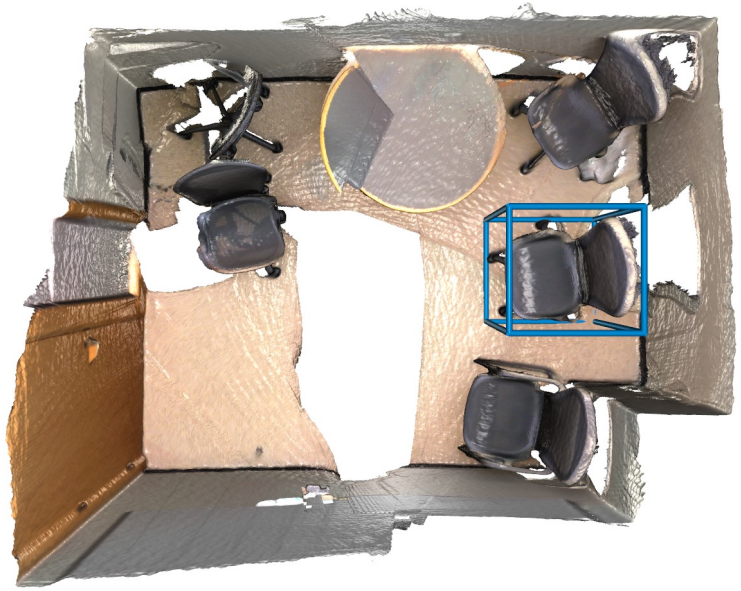
- ScanRefer: Assumes exactly one object that matches the description in the scene
- More realistic: can have zero, one, or more objects match

Zero Target



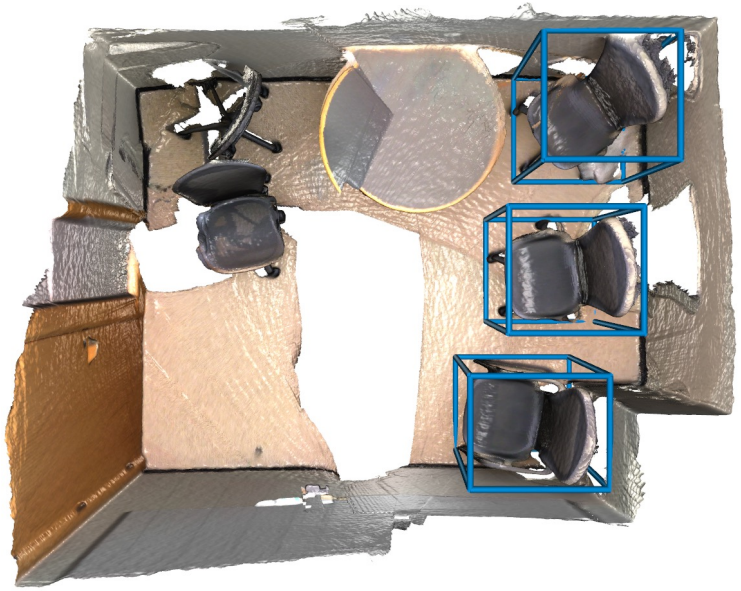
This is a white cabinet. It is next to a desk.

Single Target (ScanRefer)



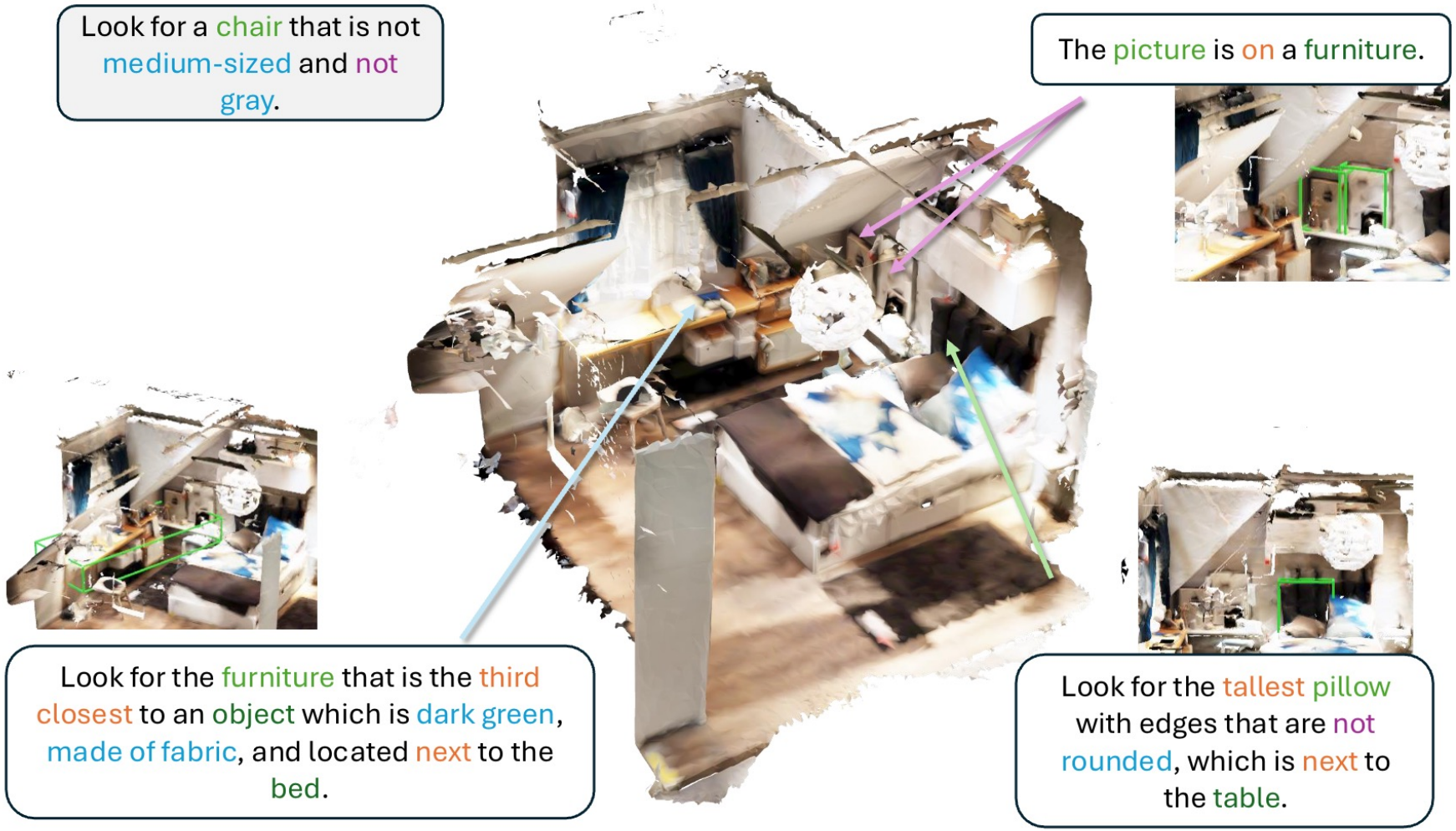
This is a black chair. It is in the middle of two chairs.

Multiple Targets



This is a black chair facing the door.

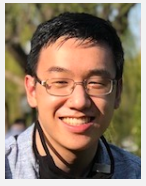
Diverse ways to refer to objects



ViGiL3D: A Linguistically Diverse Dataset for 3D Visual Grounding

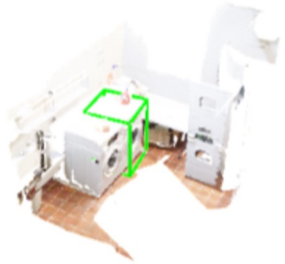
[Wang et al. ACL 2025] https://3dlg-hcvc.github.io/vigil3d/

Austin Wang



Diverse ways to refer to objects

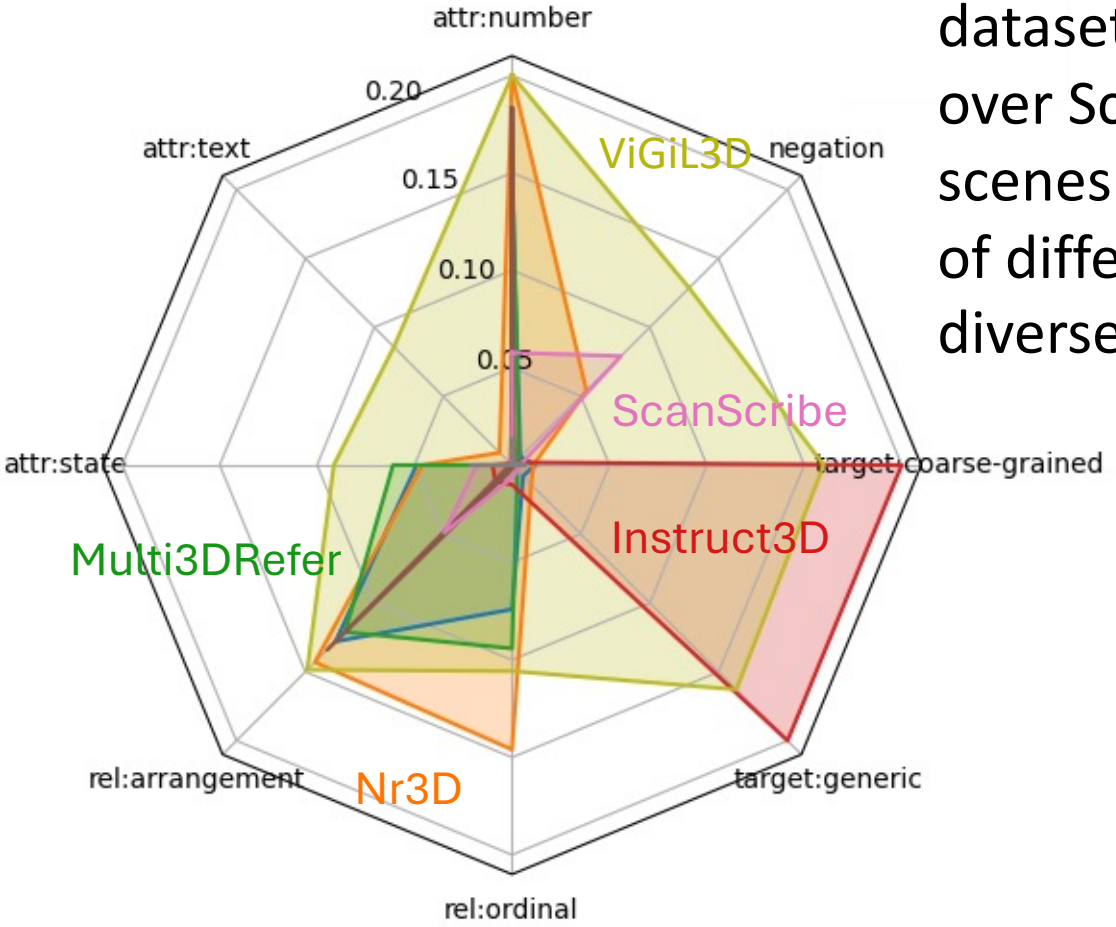
Relationship: Comparison



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

Negation

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

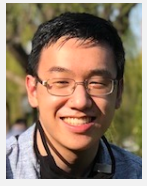


Small **manually** constructed dataset of **350 prompt**s over ScanNet / ScanNet++ scenes to benchmark ability of different models on diverse linguistic patterns

ViGiL3D: A Linguistically Diverse Dataset for 3D Visual Grounding

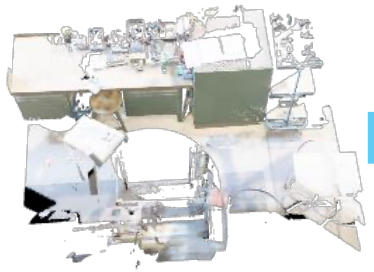
[Wang et al. ACL 2025] https://3dlg-hcvc.github.io/vigil3d/

Austin Wang



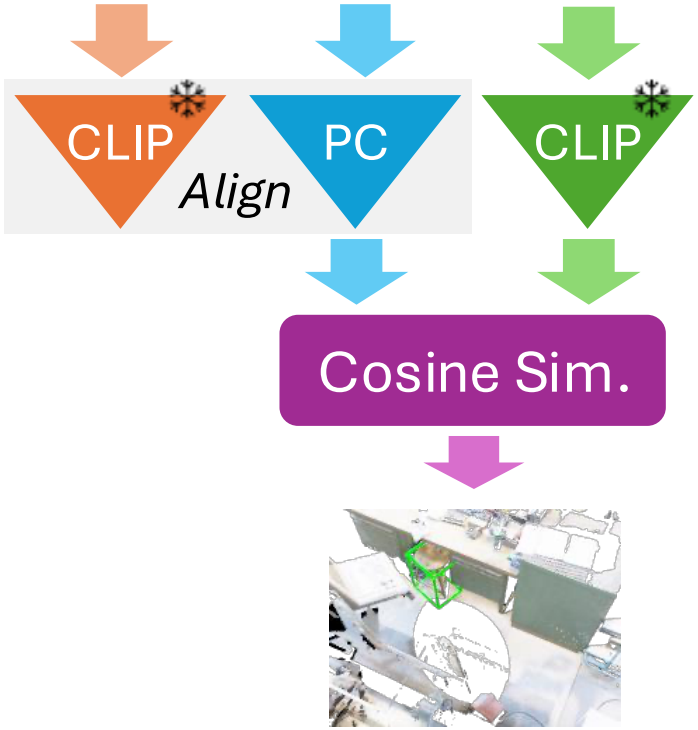
Benchmark different methods on ViGiL3D

Input



The object is under the workstation and between the cabinets.

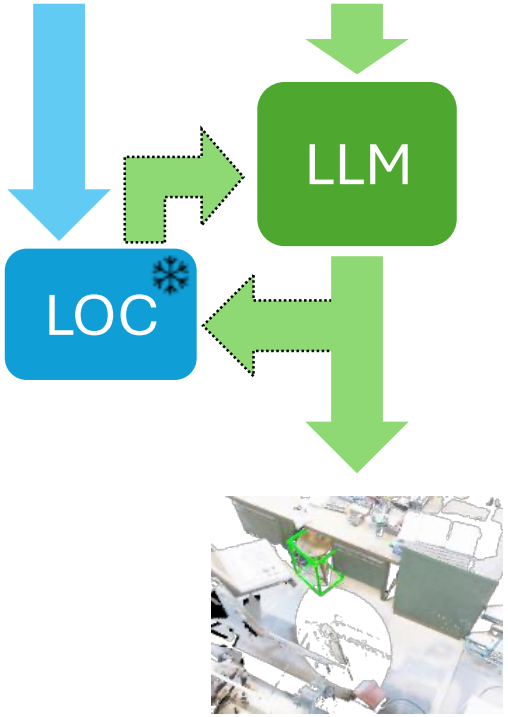
CLIP-aligned



OpenScene [Peng et al. CVPR 2023]

LERF [Kerr et al. ICCV 2023]

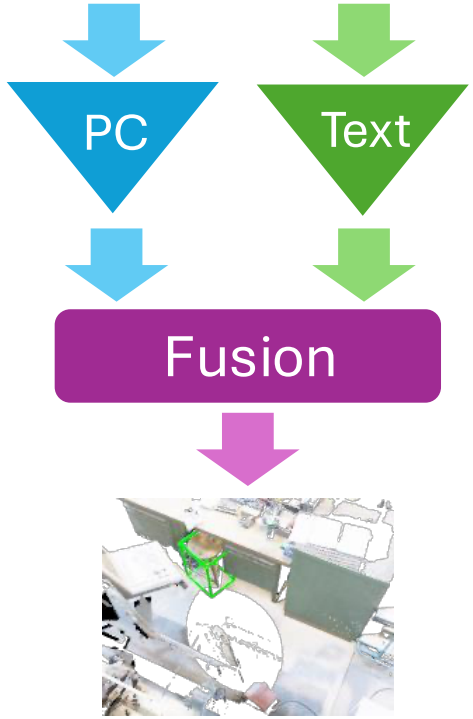
Zero-Shot LLM



ZSVG3D [Yuan et al. CVPR 2024]

LLM-Grounder [Yang et al. ICRA 2024]

3DVG Trained

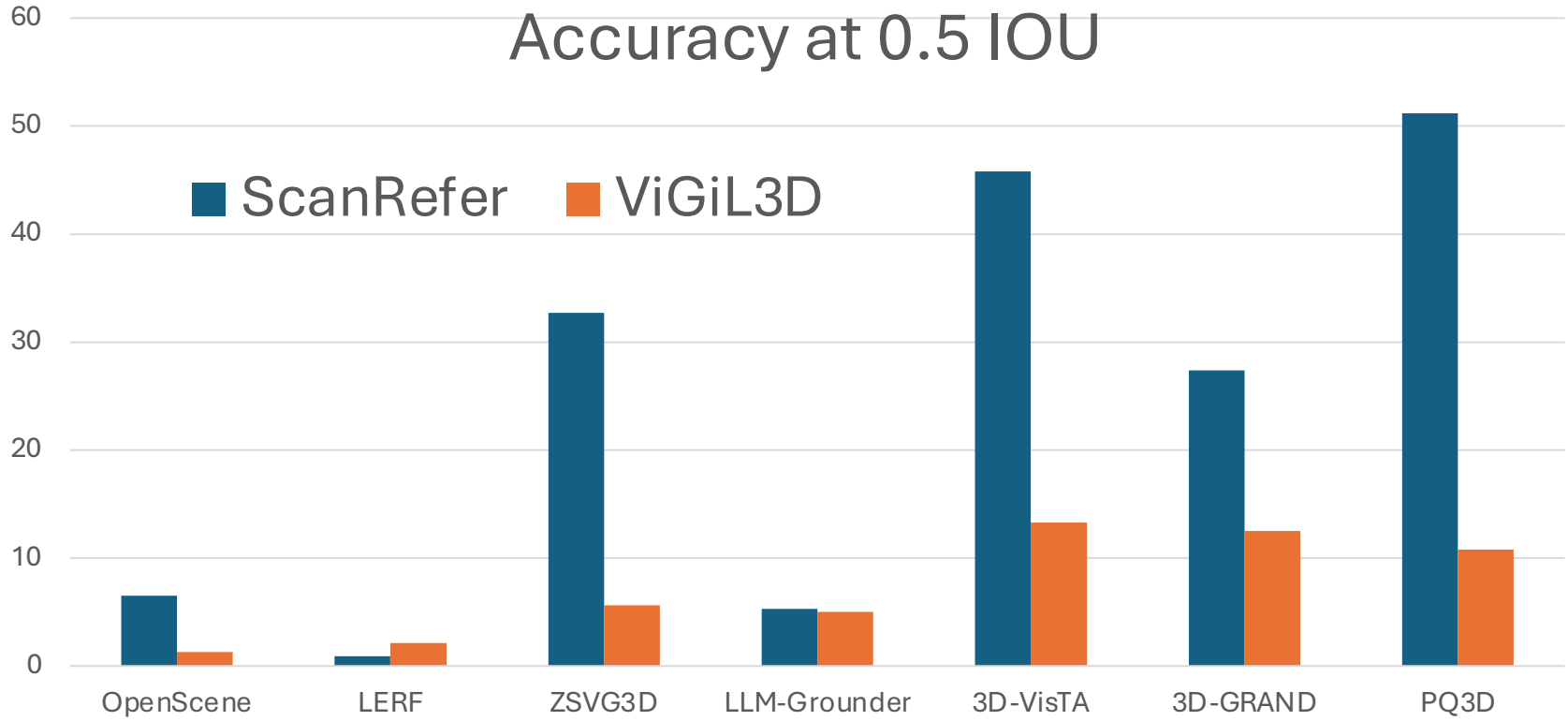


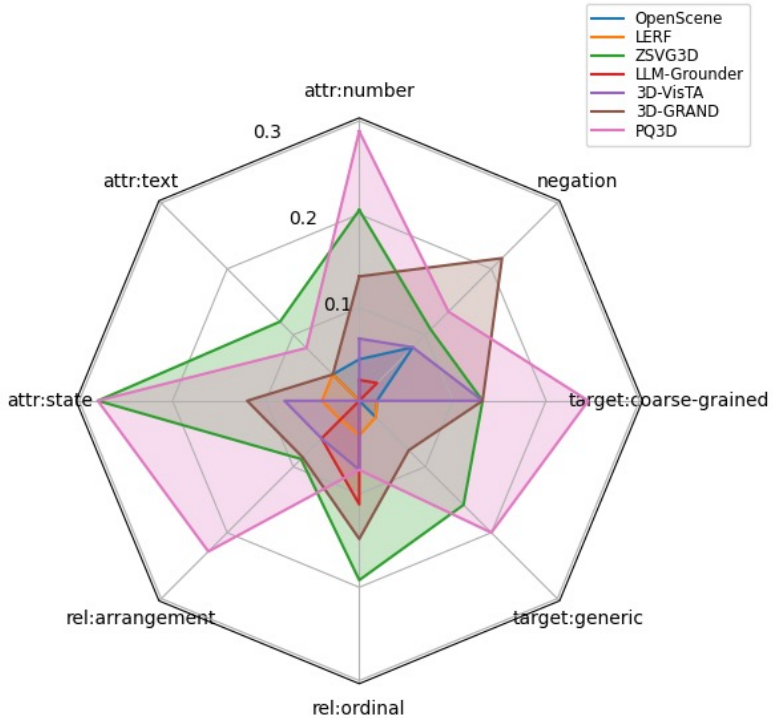
3D-VisTA [Zhu et al. CVPR 2025]

3D-GRAND [Yang et al. ICCV 2023]

PQ3D [Zhu et al. ECCV 2024]

Models perform worse on ViGiL3D





CLIP-aligned

0-shot LLM

3DVG-trained

Accuracy on ViGiL3D is at less than 15%

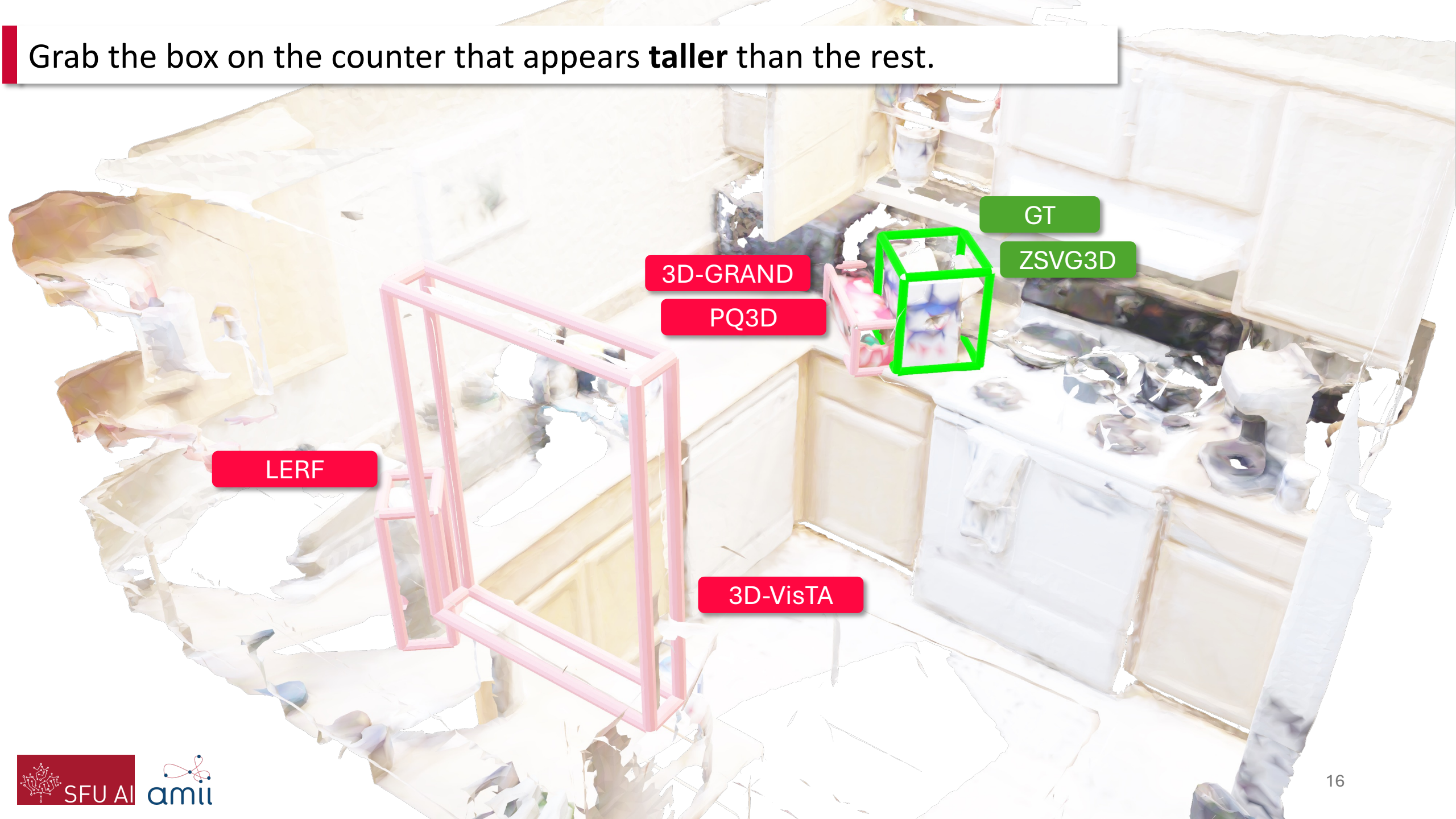
(with GT boxes, PQ3D at 26%)

3D training is still useful

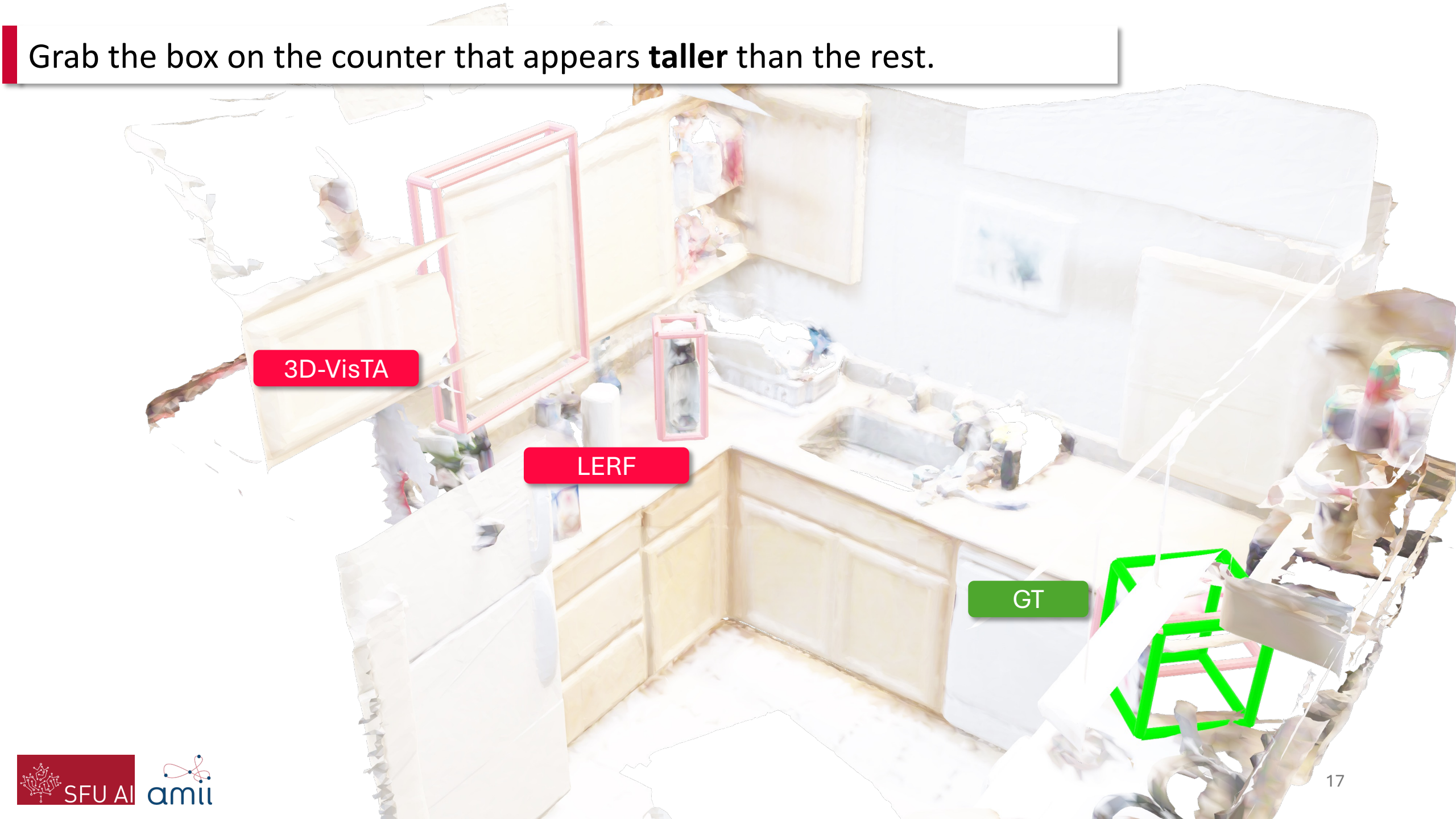
Grab the box on the counter that appears **taller** than the rest.

SFU AI amii

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

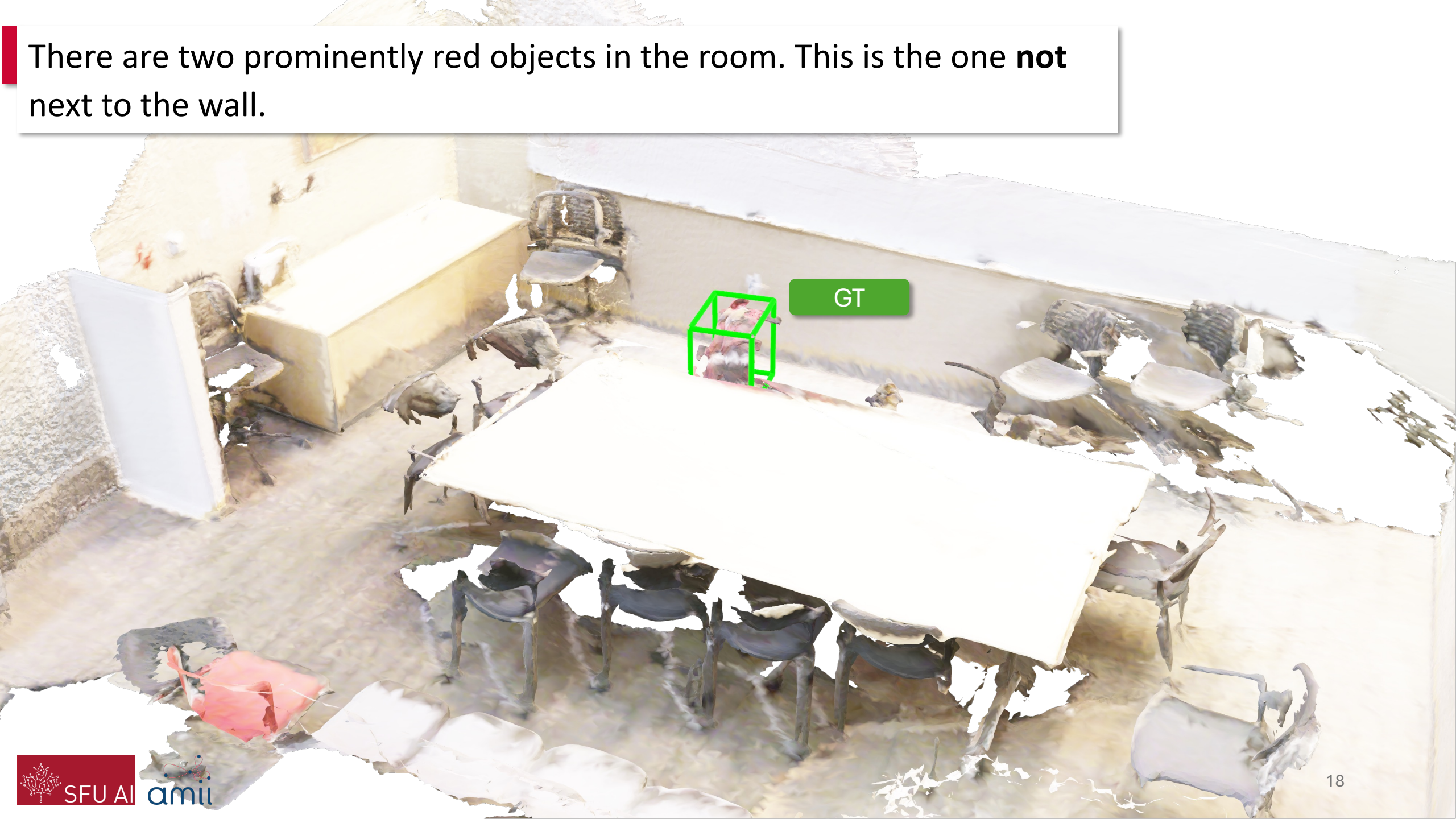


Grab the box on the counter that appears **taller** than the rest.



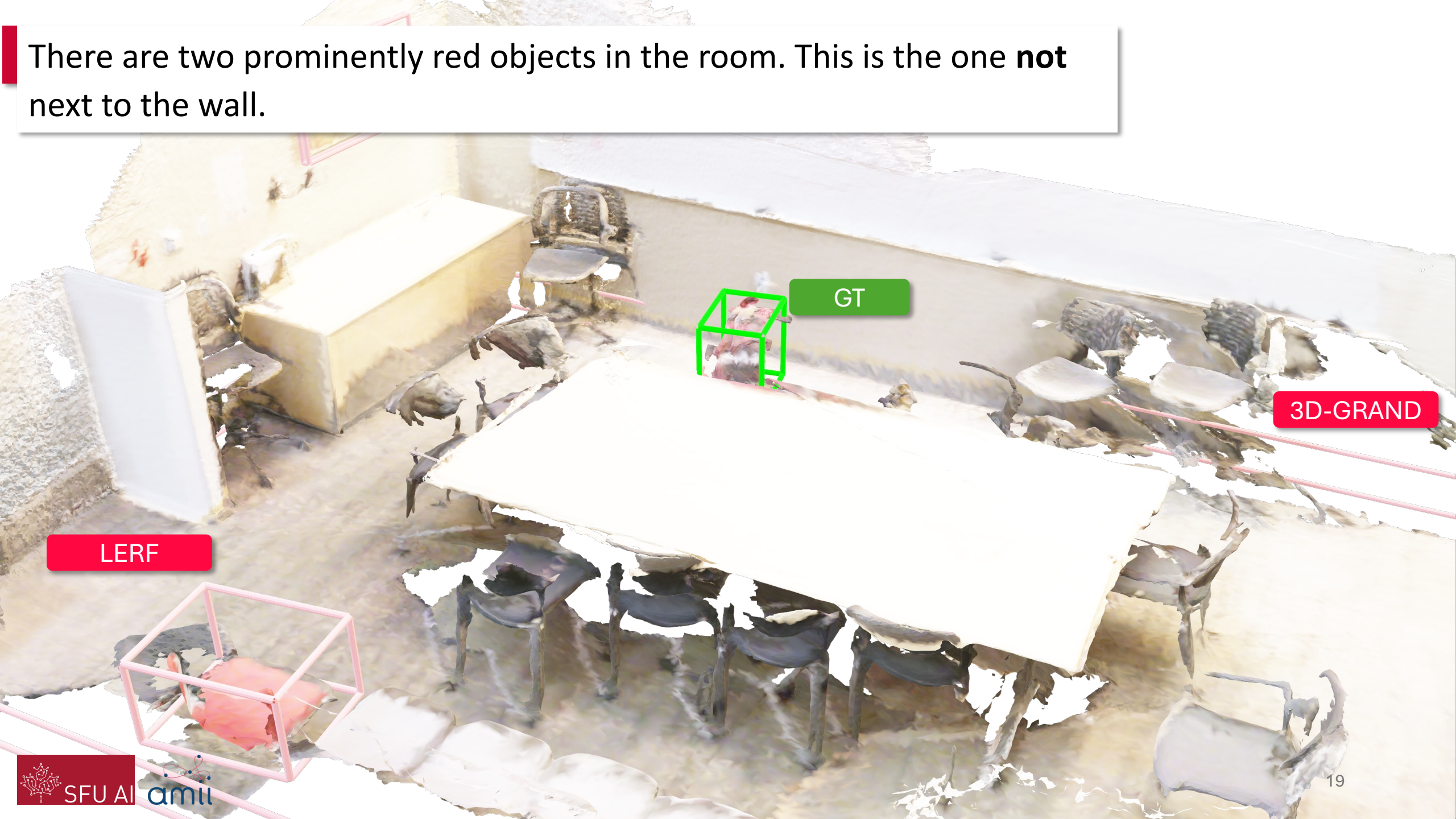
SFU AI amii

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

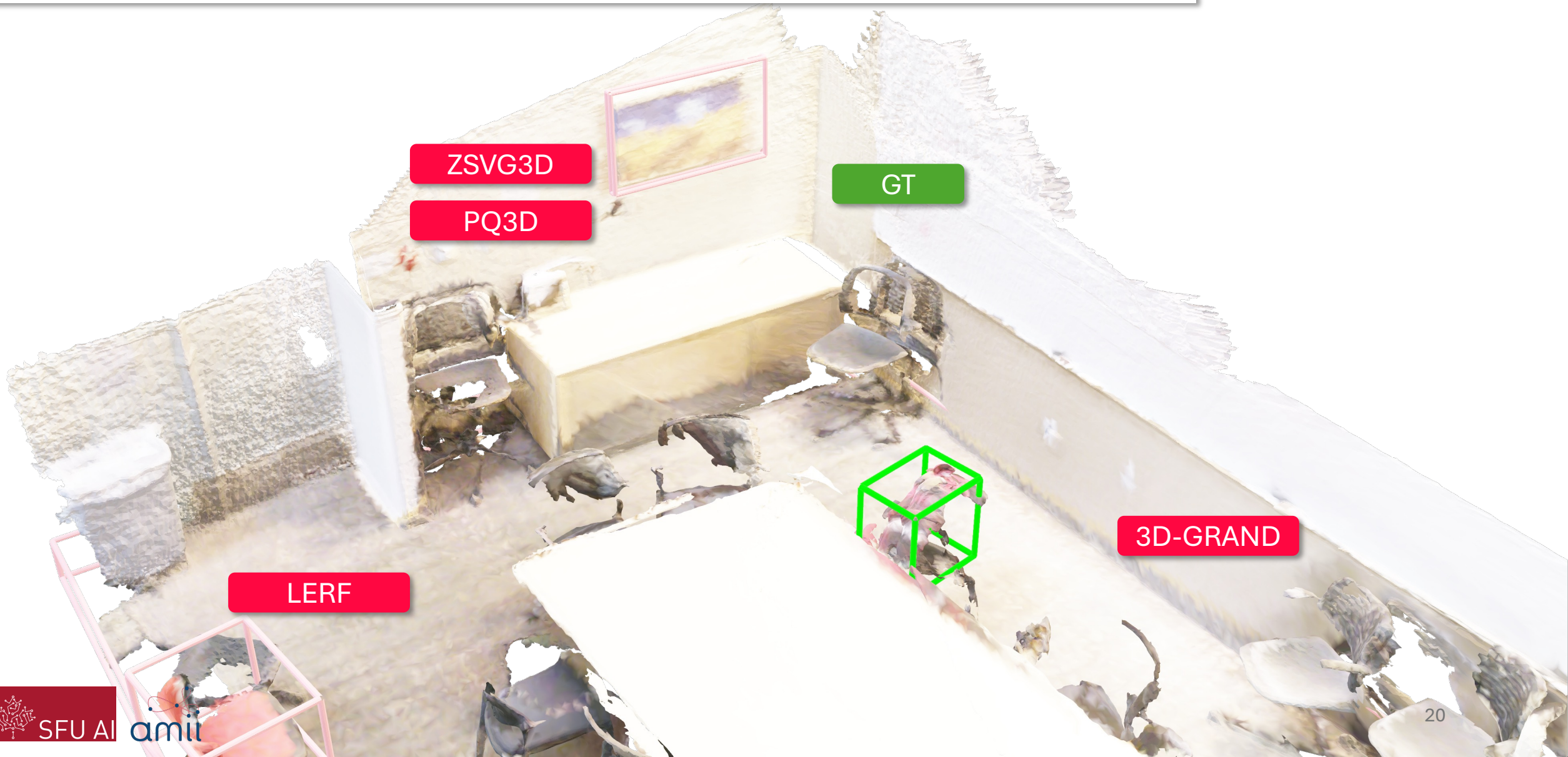


SFU AI amii

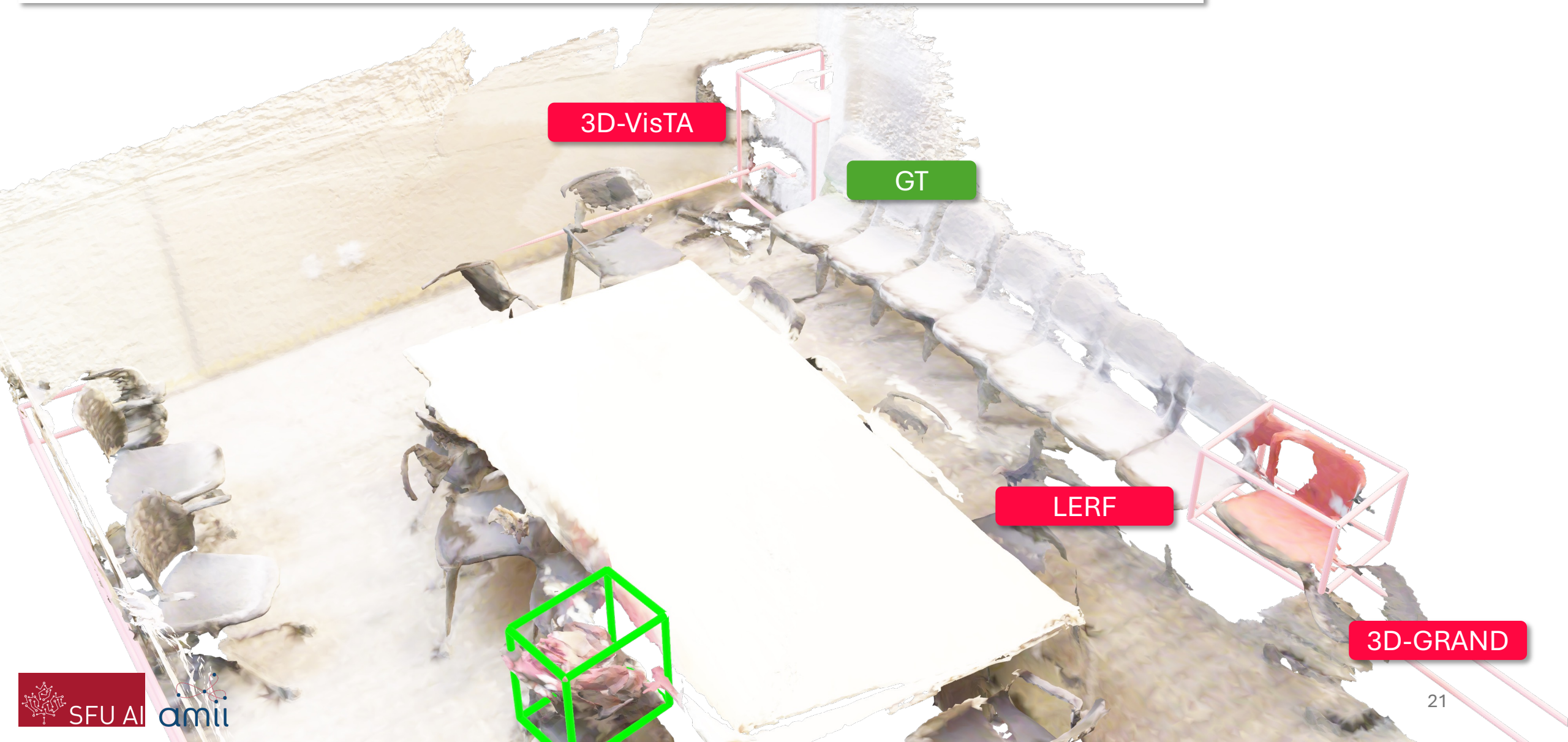
There are two prominently red objects in the room. This is the one **not** next to the wall.



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



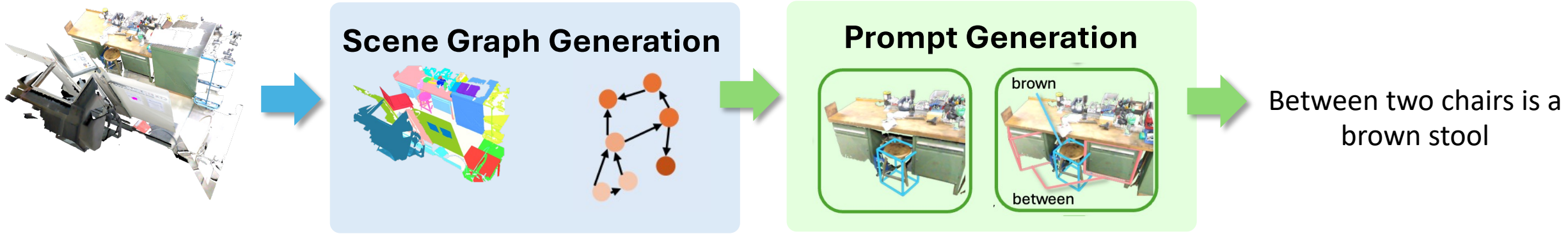
There are two prominently red objects in the room. This is the one **not** next to the wall.



Scaling up ViGiL3D

Automatic construction of prompts using LLMs

Initial attempt



Instance segmentation + scene graph construction with LLMs

Too noisy

Sample object and constraint and generate description using LLM

Need validation

ViGiL3D++: Scaling 3D Visual Grounding Generation with Diverse Language

[Wang et al. 2026]

How well can 3D LLMs reason about spatial relationships?

UNIVERSITY OF OXFORD
Xianzheng Ma

3D Scene

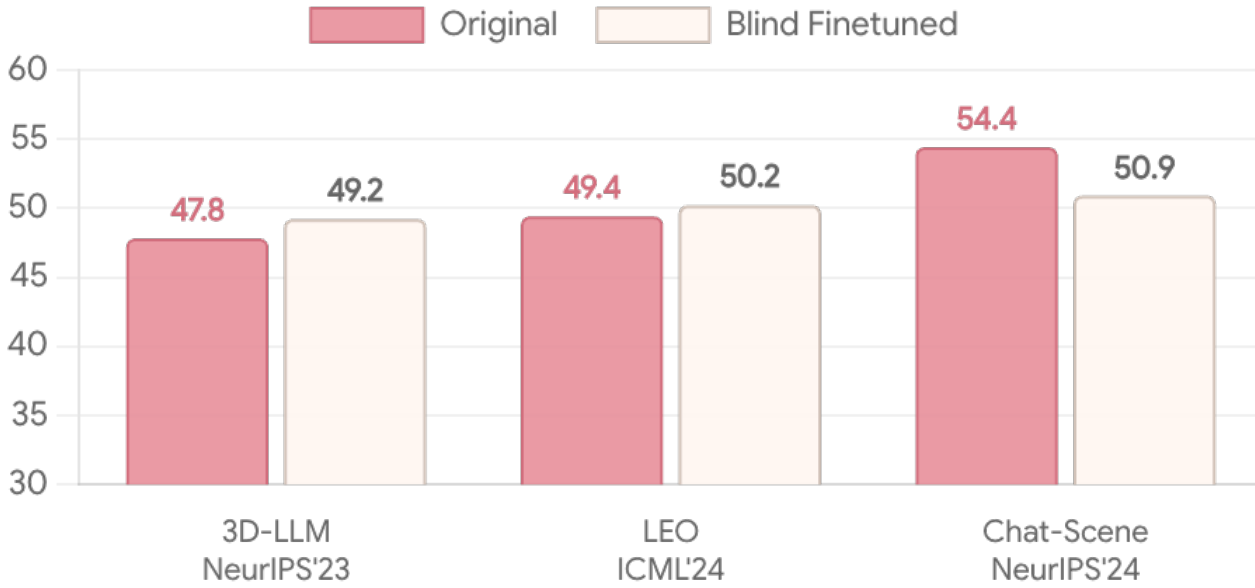
S: I am standing by the ottoman on my right facing a couple of toolboxes.

S: I am facing the armchair with a backpack and the couch behind me.

Q: What instrument in front of you is ebony and ivory?

Q: What is the first black object you ' d run into behind you?

Questions from SQA3D [Ma et al. ICLR 2023]



Models can do well without any visual input!

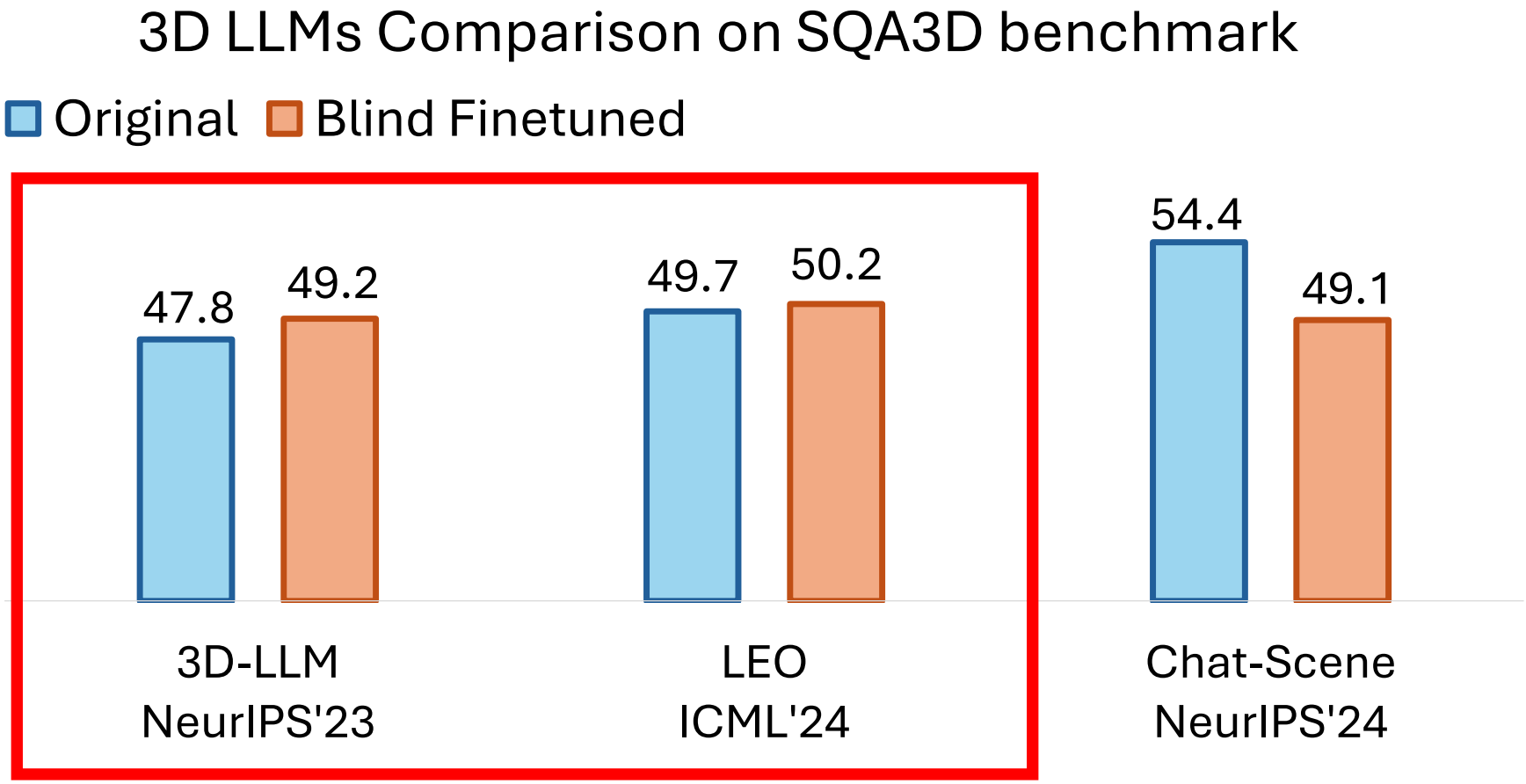
Improved 3D aware dataset with rotational viewpoint augmentation

Do 3D Large Language Models Really Understand 3D Spatial Relationships?

Workshop on 3D-LLM/VLA [Ma et al. ICLR 2026] https://real-3dqa.github.io/

Prior benchmarks are limited

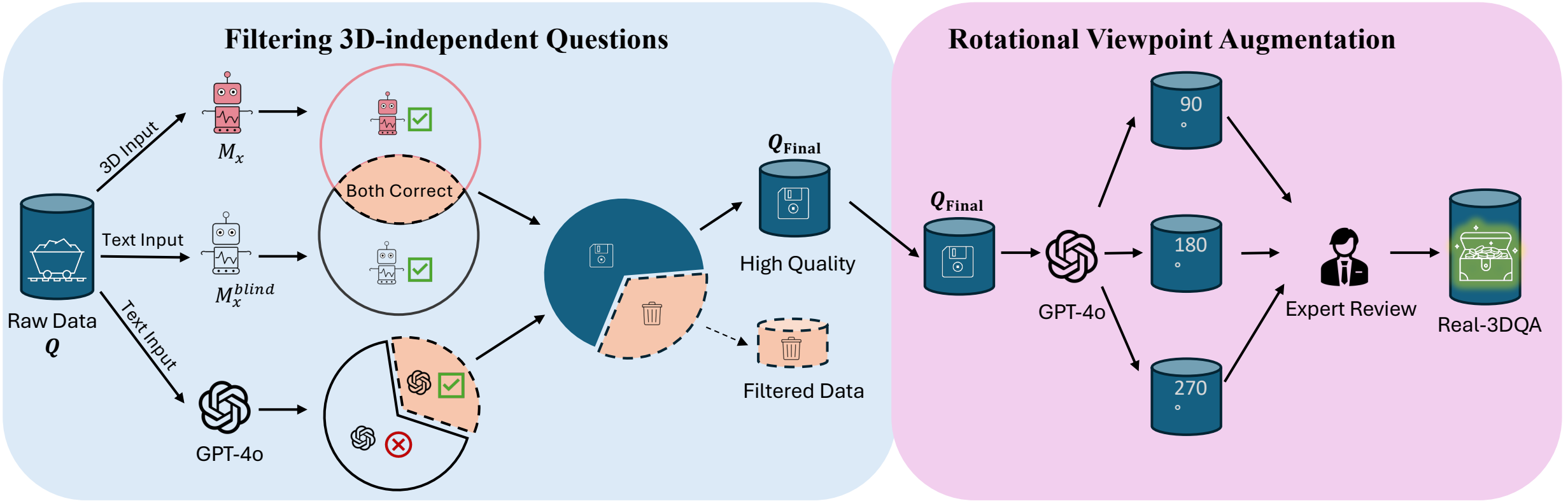
Blind-finetuned models can perform very well



Do 3D Large Language Models Really Understand 3D Spatial Relationships?

[Ma et al. ICLR 2026] https://real-3dqa.github.io/

Real-3DQA: new benchmark

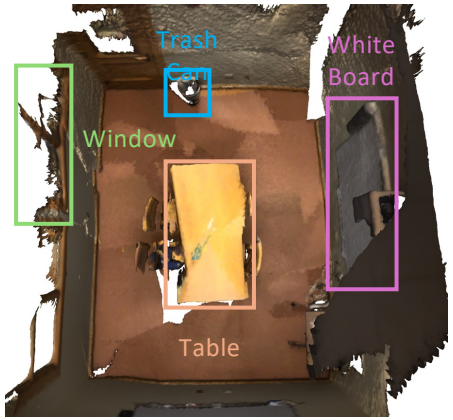


Filter out 3D independent questions

Viewpoint augmentation

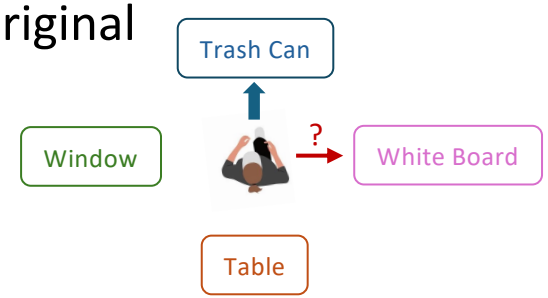
Example of rotation augmentation

Augmented SQA



Sample Room Layout

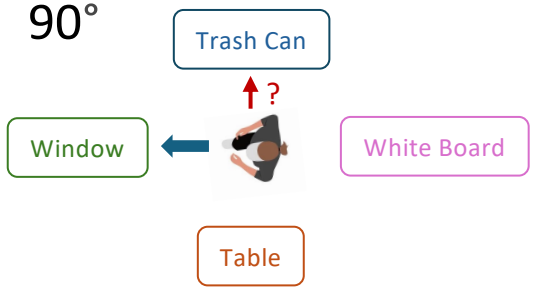
Original



Situation: I am facing a trash can, with a table behind me.

Question: What is on my right?
Answer: White board.

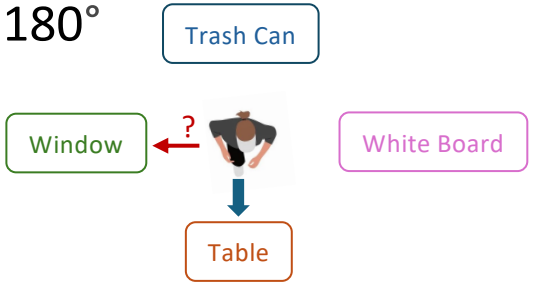
90°



Situation: I am opening the window, with a table on my left.

Question: What is on my right?
Answer: Trash can.

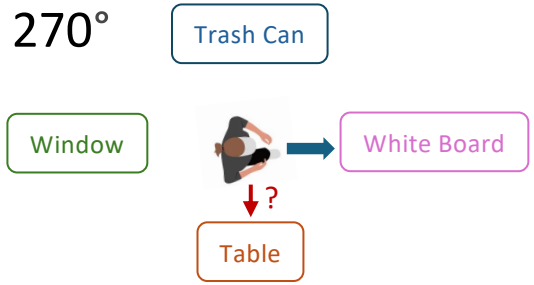
180°



Situation: I am facing a table, with a white board on my left.

Question: What is on my right?
Answer: Window.

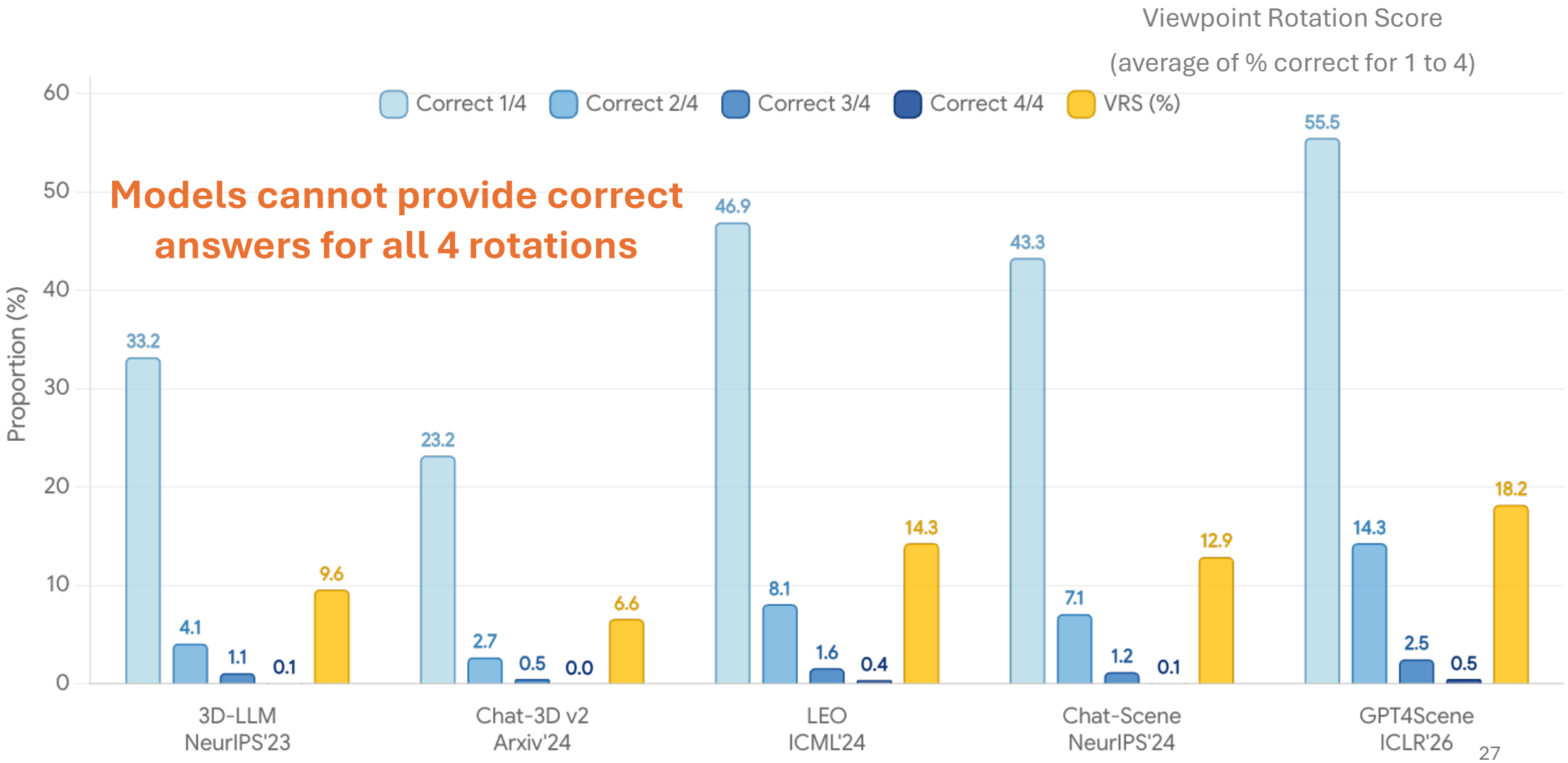
270°



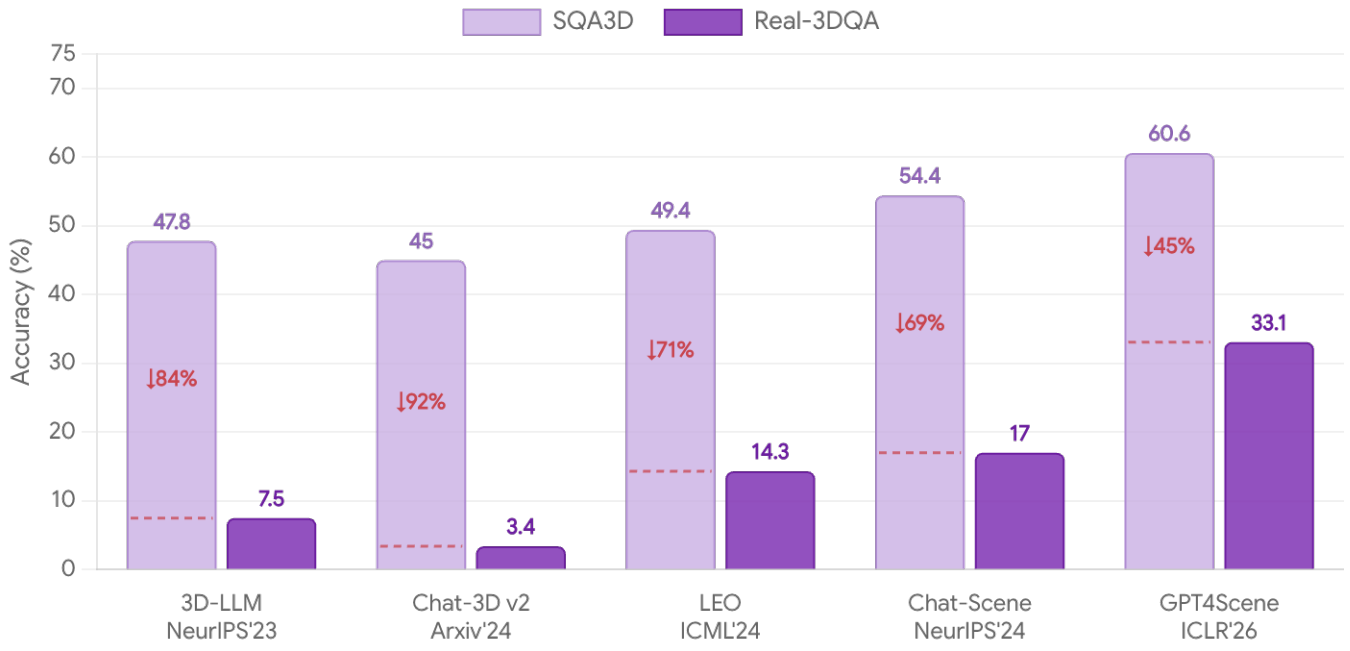
Situation: I am trying to clean the white board, with a trash can on my left.

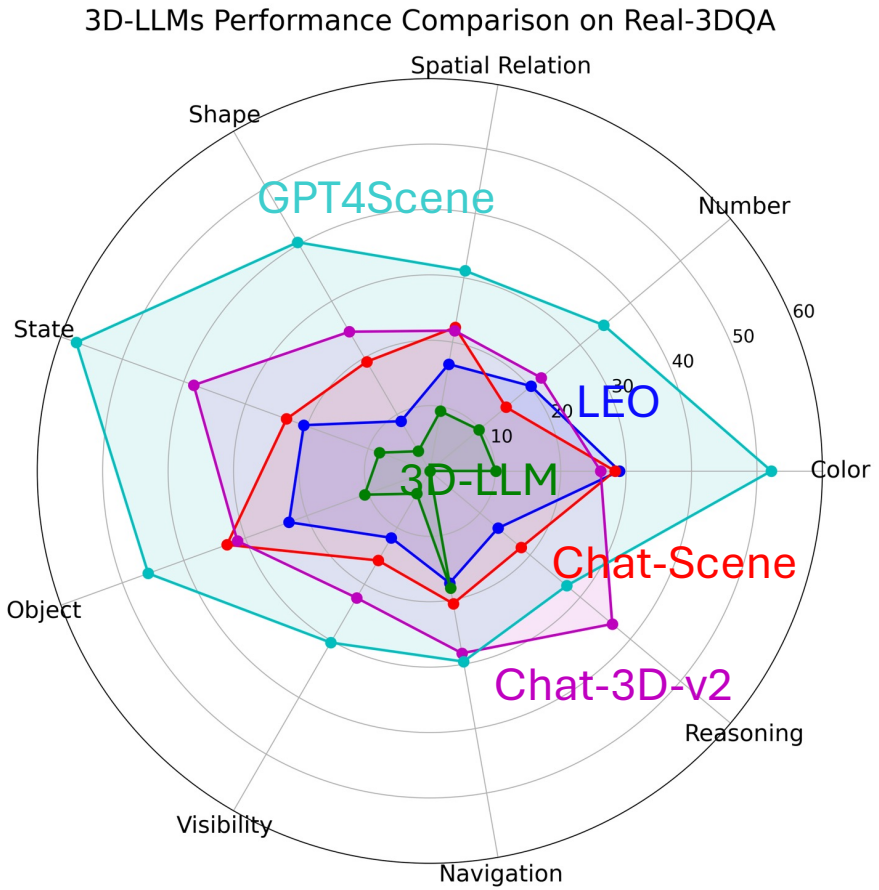
Question: What is on my right?
Answer: Table.

How well does current 3D LLMs do?



How well does current 3D LLMs do?



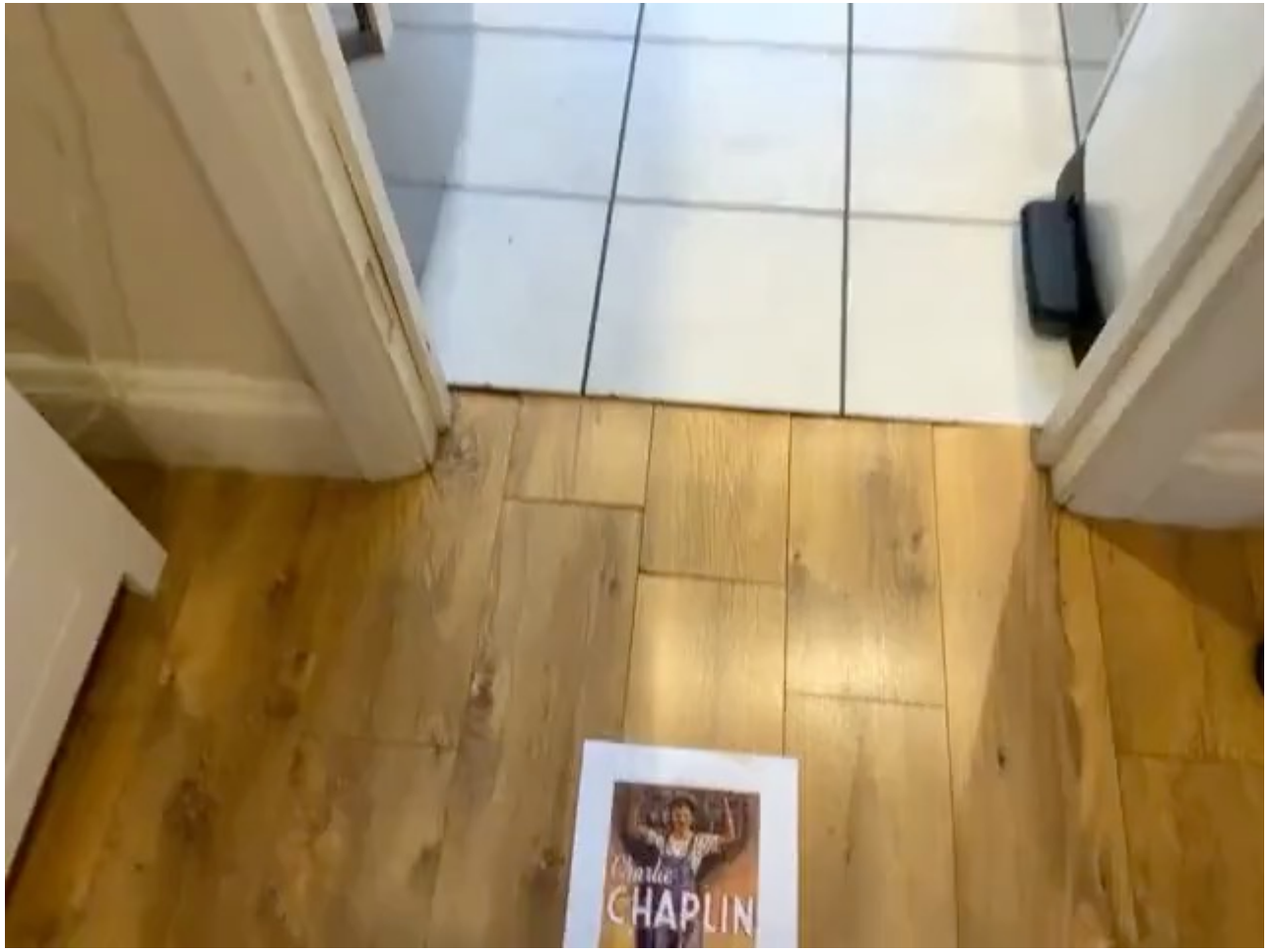


Real-3DQA is challenging!

Large drop in performance across all methods

What about spatial understanding by VLMs?

- Can we answer questions about a 3D space by just watching a video?
- Is it necessary to reconstruct to 3D?



Thinking in Space: How Multimodal Large Language Models See, Remember, and Recall Spaces [Yang et al., 2025]

What about spatial understanding by VLMs?

3D Visual Spatial Intelligence Benchmark: VSI-Bench

Object Count

How many chairs are there in this room?

Answer: 4

Relative Distance

Measuring from the closest point of each object, which of these objects (refrigerator, sofa, ceiling light, cutting board) is the closest to the printer?

A. refrigerator B. sofa C. ceiling light ***D. cutting board***

Appearance Order

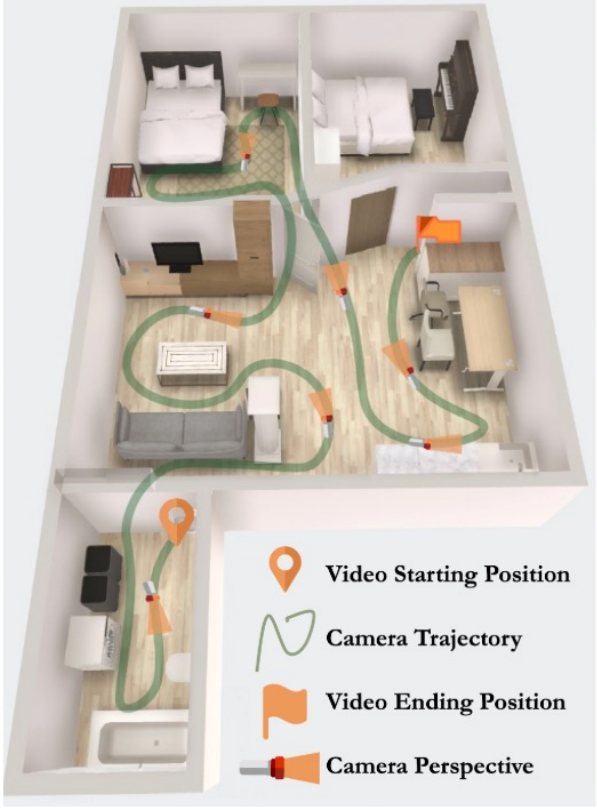
What will be the first-time appearance order of the following categories in the video: basket, printer, refrigerator, kettle?

A. kettle, basket, printer, refrigerator
B. refrigerator, printer, basket, kettle
C. basket, printer, refrigerator, kettle
D. basket, refrigerator, kettle, printer

Relative Direction

If I am standing by the refrigerator and facing the sofa, is the kettle to my left, right, or back?

A. left B. right ***C. back***



Object Size

What is the length of the longest dimension (length, width, or height) of the refrigerator in centimeters?

Answer: 119

Absolute Distance

Measuring from the closest point of each object, what is the distance between the bed and the sofa in meters?

Answer: 3.2

Room Size

What is the size of this room (in square meters)? If multiple rooms are shown, estimate the size of the combined space.

Answer: 57.6

Route Plan

You are a robot beginning at the toilet and facing the washer. Navigate to the pan. Fill in this route: 1. Go forward until the washing machine 2. [?] 3. Go forward until the sofa 4. [?] 5. Go forward until the pan.

A. Turn Left, Turn Left B. Turn Left, Turn Right
C. Turn Back, Turn Right ***D. Turn Right, Turn Right***

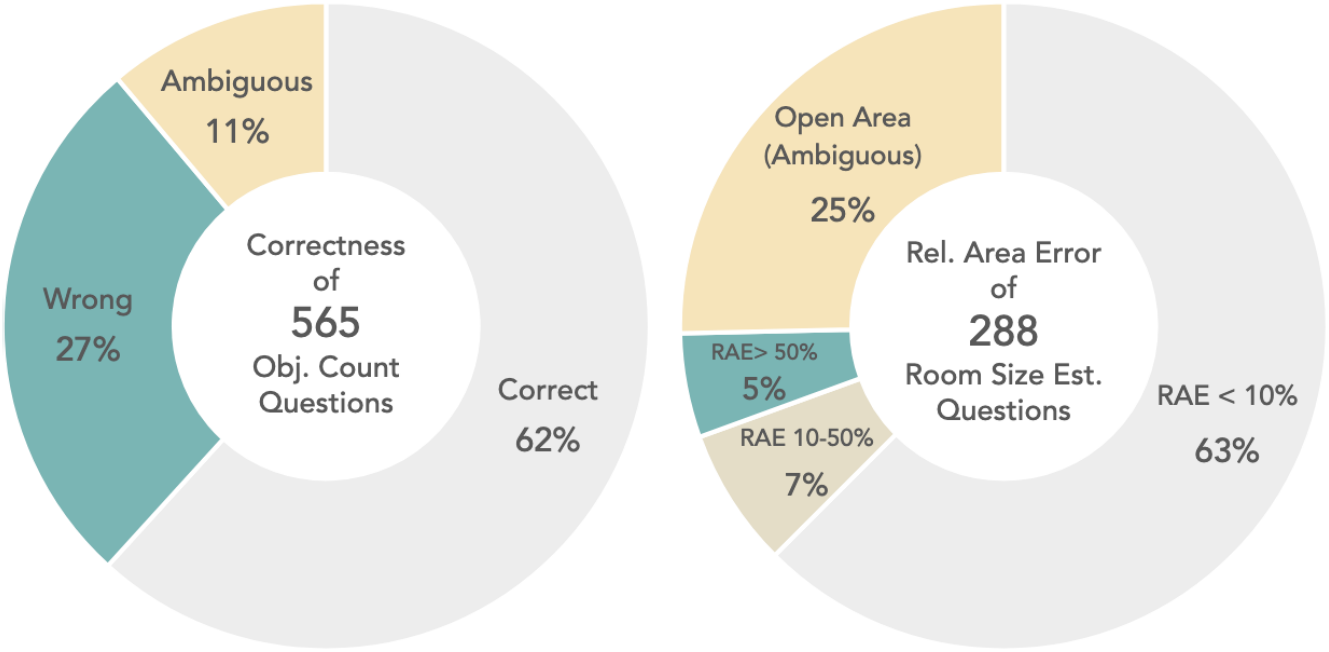
Thinking in Space: How Multimodal Large Language Models See, Remember, and Recall Spaces [Yang et al., 2025]

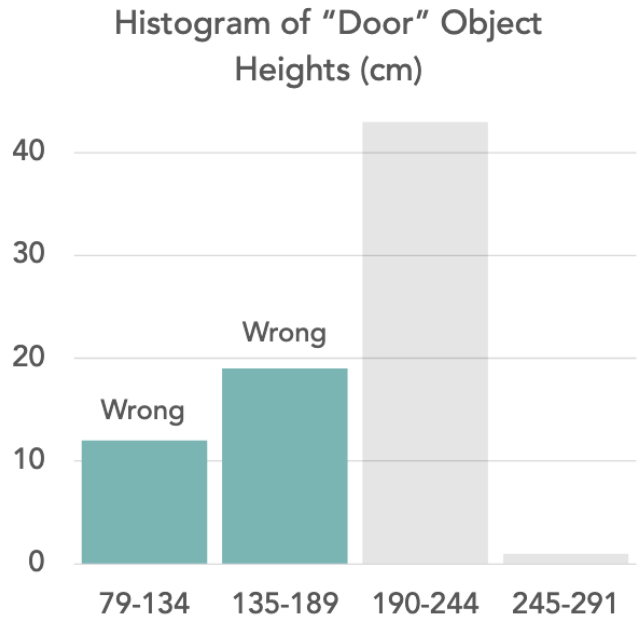
Pitfalls in Prior 3D Visual Spatial Intelligence Evaluations

https://3dlg-hcvc.github.io/revsi/

Yiming Zhang

Inaccurate Ground-Truth: 30-40% of answers in VSI-Bench is wrong or ambiguous!





ReVSI: Rebuilding Visual Spatial Intelligence Evaluation for Accurate Assessment of VLM 3D Reasoning [Zhang et al. ICML 2026]

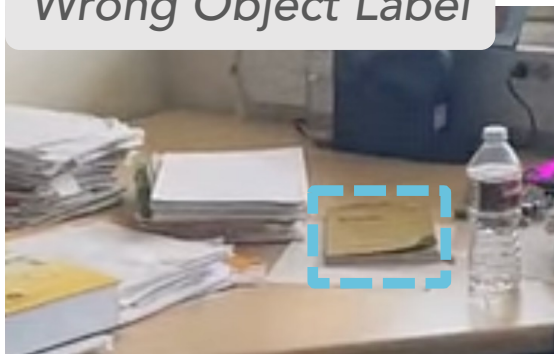
Pitfalls in Prior 3D Visual Spatial Intelligence Evaluations

Mismatch in what answers should be based on **video** and **incomplete, noisy 3D data**

- Template-based QA generation pipeline (missing human verification)
- GT is calculated using 3D annotations (mismatch video content)

Inaccurate object count and sizes due to missing geometry

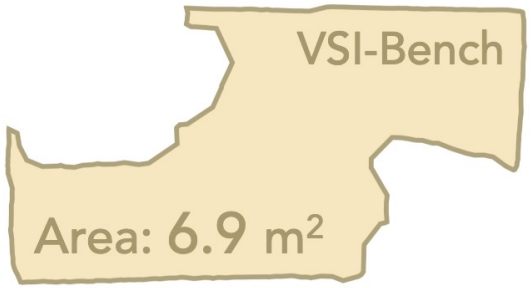
Wrong Object Label



Cup → Notebook

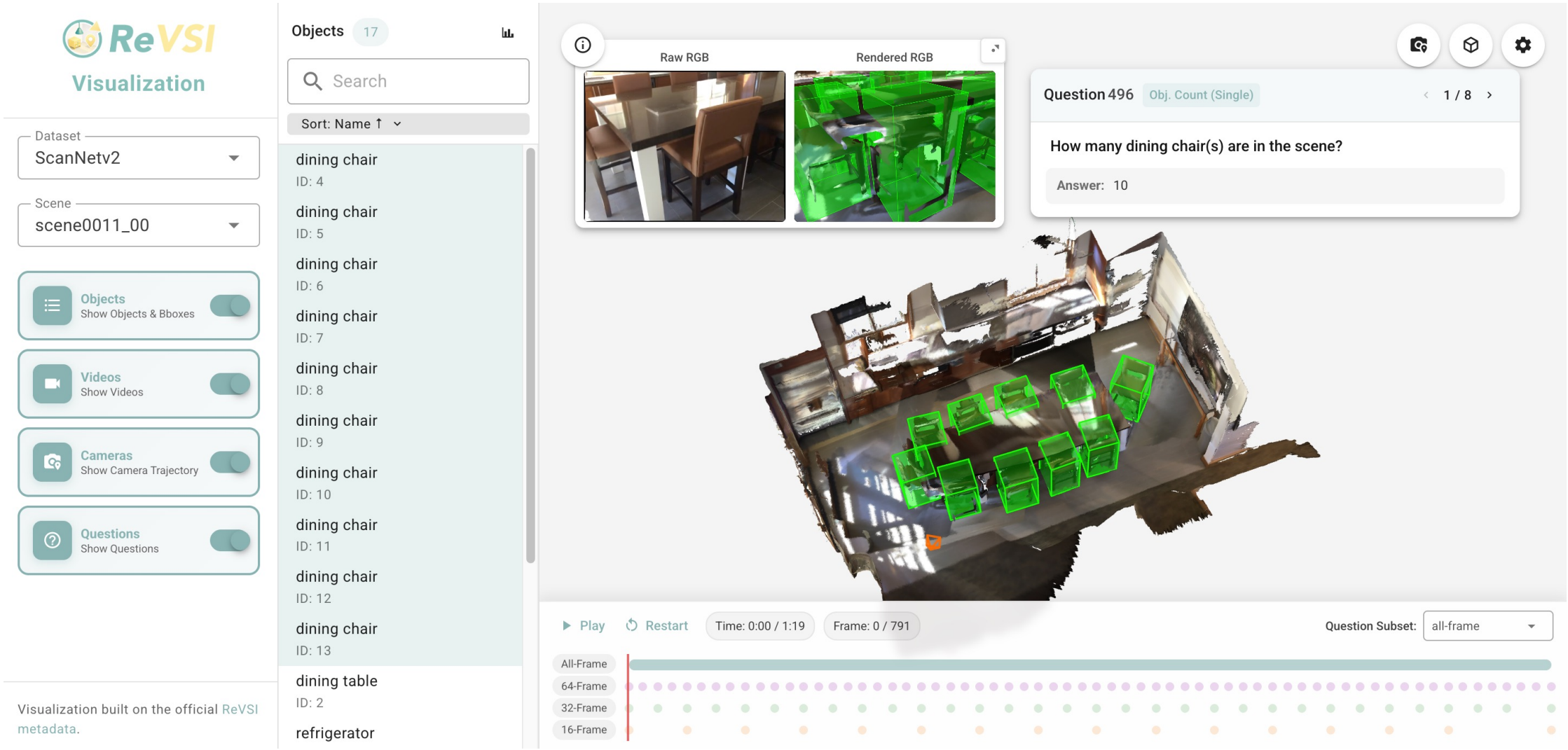
Keyboard → Cables

Inaccurate room size estimate



ReVSI: we re-construct and re-annotate the dataset

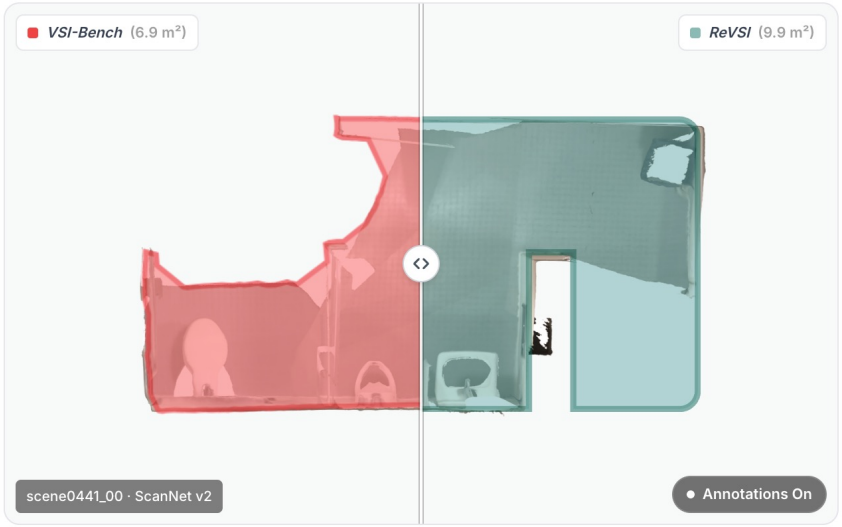
Build 3D/video interface to inspect and re-annotate the dataset

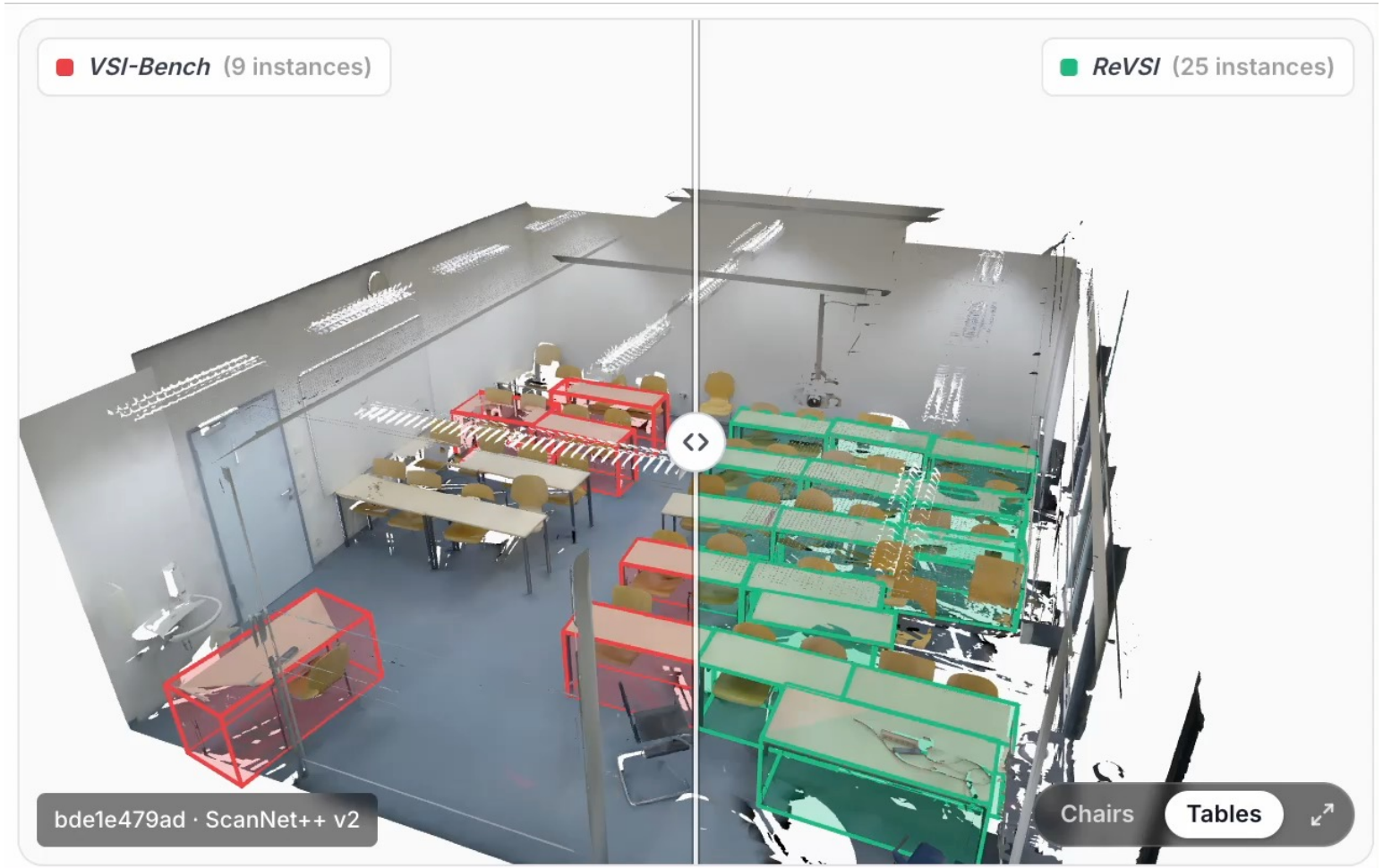


ReVSI: we re-construct and re-annotate the dataset

Re-annotate and inspect with 3D/video interface

1. **Correct object annotation:** 3D object bounding boxes and object labels
2. **Room boundary annotation**





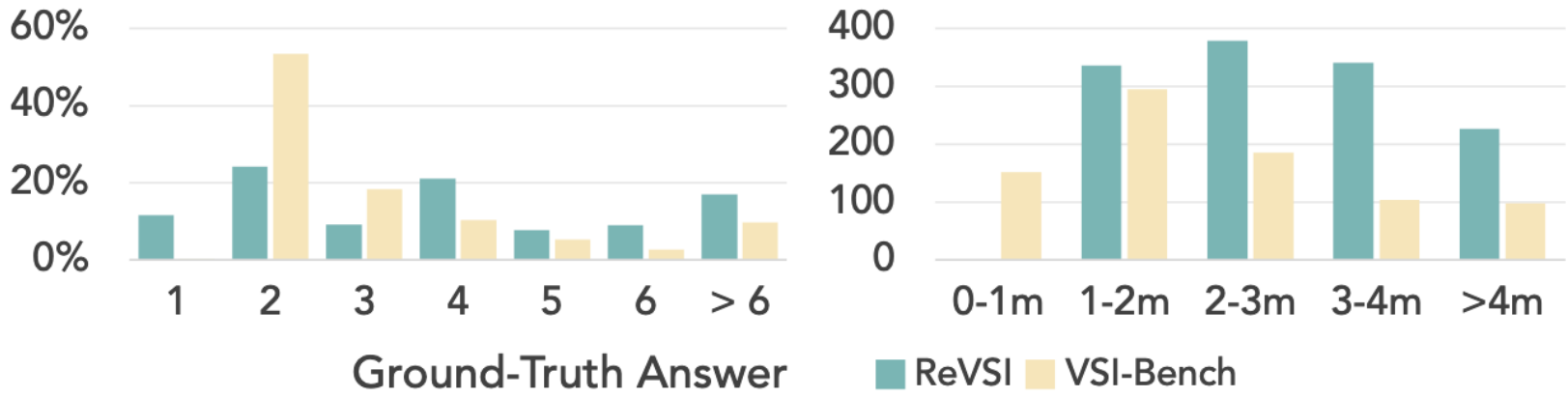
ReVSI: we re-construct and re-annotate the dataset

Eliminate mismatch in frame-based evaluation and answer based on full 3D scene

How Many Pillows Are In This Room? *VSI-Bench GT: 6*

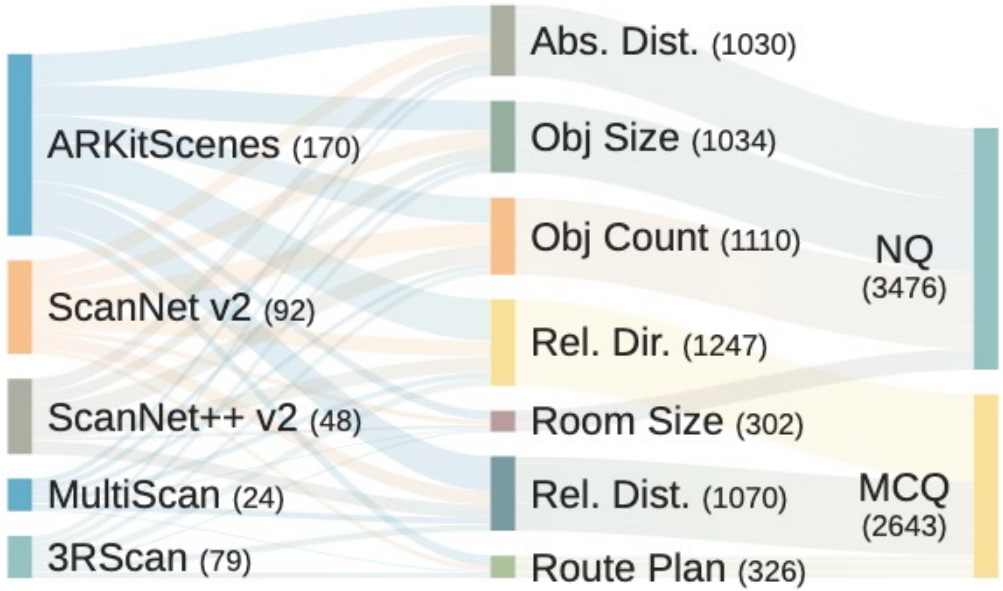
Actual GT: ALL-FRAMES (>1000): 6 | 64-FRAMES: 4 | 32-FRAMES: 3 | 16-FRAMES: 1

Avoid strong biases in answer distribution



ReVSI: we re-construct and re-annotate the dataset

	Scenes	Objects	Obj. Labels	Open-Vocab
VSI-Bench	288	3185	65	
ReVSI	**413**	**5436**	**466**	✓



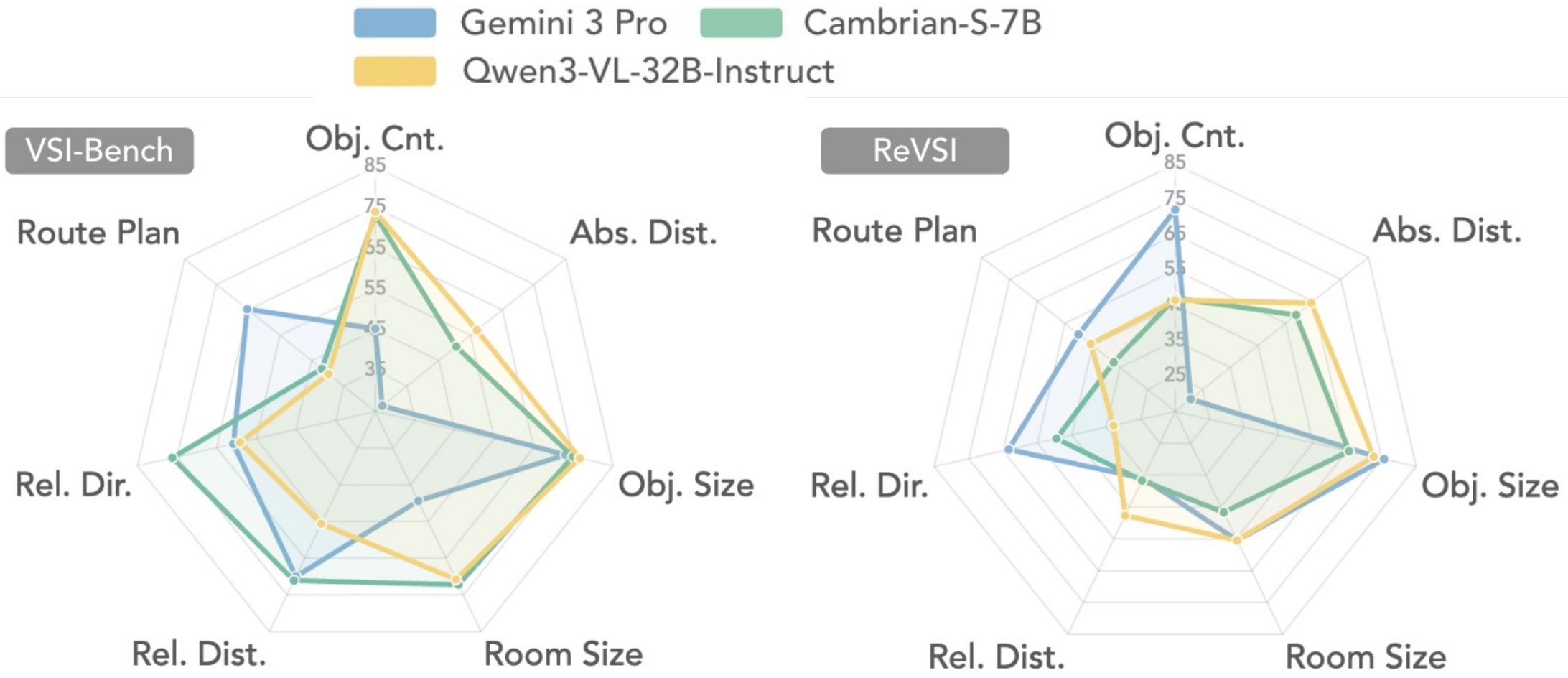
Re-annotate with 3D/video interface

1. **Correct object annotation:** 3D object bounding boxes
2. **Frame-adaptive QA pairs base on object visibility**
3. **Balanced Data Distributions:**
4. **Fully manually verified**

Increase diversity:

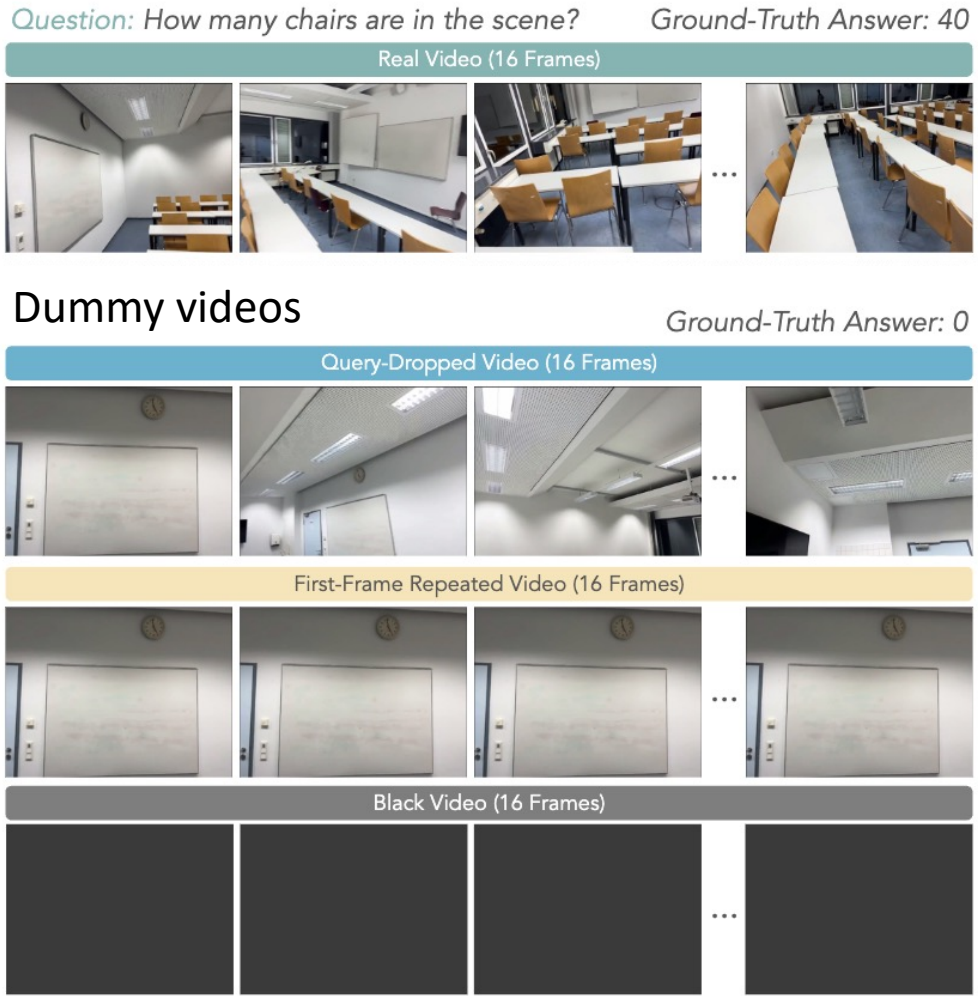
1. **Open Vocabulary:** More human-annotated object labels.
2. **More Scenes:** Include 5 indoor scene datasets.
3. **More Question Variants:** More challenging questions.

ReVSI: Re-evaluate and re-visit prior findings



Performance of proprietary VLMs are under-estimated by prior benchmark!

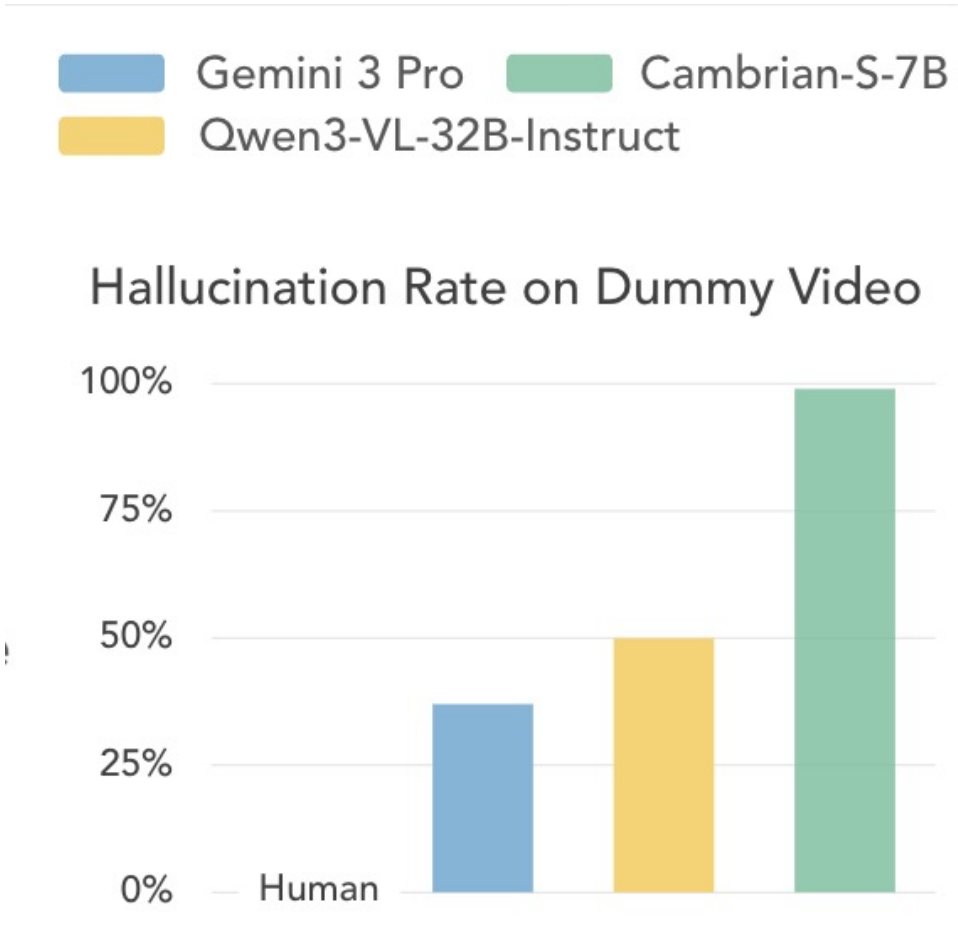
ReVSI: Re-evaluate and re-visit prior findings



Target object not present

First frame repeated

Black frames



Open-source fine-tuned models are more prone to hallucinations.

Summary

- Current VLMs and 3D-LLM still lack 3D spatial understanding
  - Hallucinations
  - Cannot reason about rotation, spatial relations, comparatives, negation…
  - Cannot handle linguistic diversity
- Fine-tuning lead to over-reliance on memorized priors
- Need continual development and refinement of datasets and benchmarks
  - Does it need vision? Compare to blind agents!
  - Design probing questions and scenarios

Some problems my group work on

- Language understanding in 3D scenes
- 3D scene and shape generation
- Modeling interactive (e.g. articulated) objects

Toward large-scale interactive scene generation
Workshop on Urban Scene Modeling
Mile High 3B - Afternoon (4pm)
https://usm3d.github.io/

JRM: Joint Reconstruction Model for Multiple Objects without Alignment

[Wu et al. CVPR 2026]

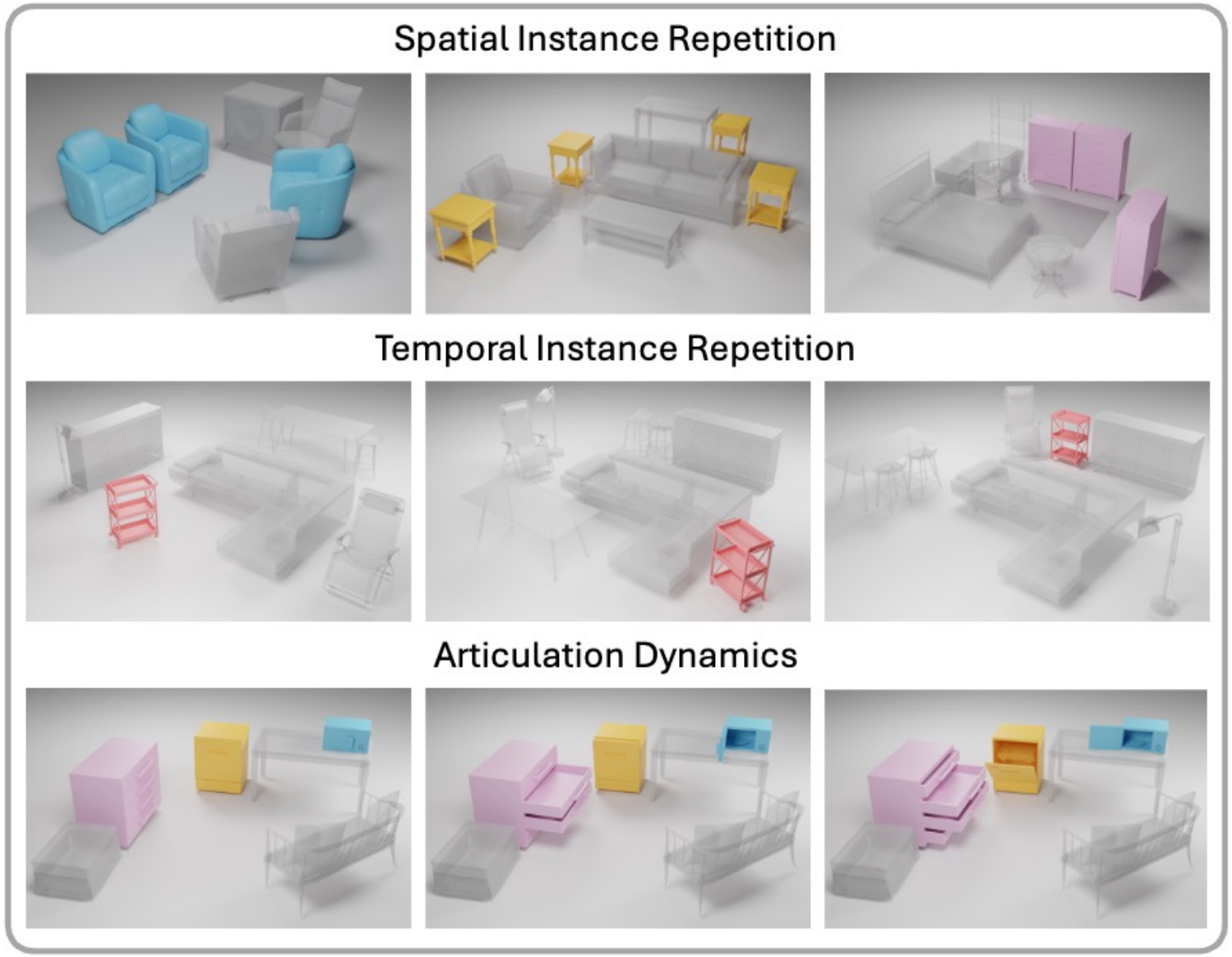
Qirui Wu

Meta

Similar objects / geometry re-observed across space and time

Leverage information across observations for improved reconstruction

Poster Session 1 – Fri am
#30



Static-to-Articulated: Artiverse [Iliash et al. CVPR 2026]

Dataset of 5K articulated objects with 25K articulated parts over 88 categories

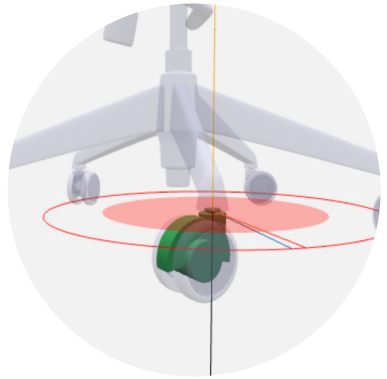
Object Attributes

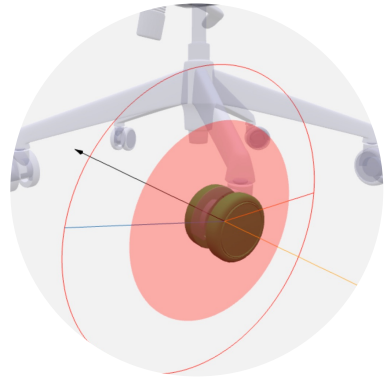
Category: *swivel_chair*
Weight: *15.2 kg*
Physical dimension: *105 x 58 x 58 cm*

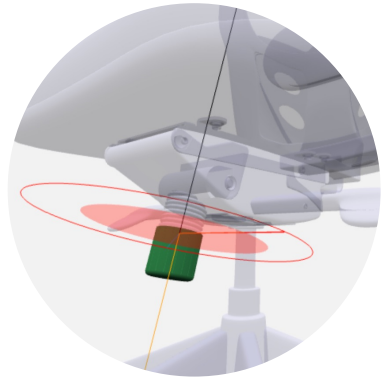
Part Attributes

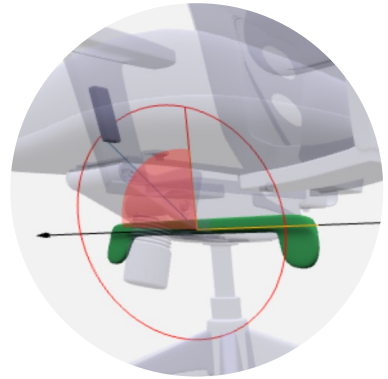
Part label: *seat*
Material: *foam*
Density: *0.045 g/cm³*
Weight: *0.81 kg*

Part label: *caster_connector*
Material: *plastic*
Density: *1.4 g/cm³*
Weight: *0.022 kg*









Physical properties

Textured object

Functional part segmentation

Articulated parts and motion joints

Poster Session 2 – Fri pm
#164

Denys Iliash, Jiayi Liu, Egor Forkin, Qirui Wu

3DV 2027

Thessaloniki, Greece
April 6-9, 2027
International Conference on 3D Vision

General Chairs

Angela Yao
National University of Singapore

Dimitris Tzionas
University of Amsterdam;
Aristotle University of Thessaloniki

Martin R. Oswald
University of Amsterdam

Program Chairs

Angel X. Chang
Simon Fraser University

Iro Armeni
Stanford University

Or Litany
Technion; Nvidia

Torsten Sattler
Czech Technical University

Important Dates

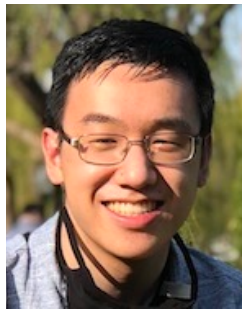
Paper Submission	Aug 28, 2026
Suppl Material	Sep 02, 2026
Preliminary Notification	Oct 27, 2026
Final Notification	Dec 02, 2026
Main Conference	Apr 6-9, 2027

For more information:
https://3dvconf.github.io

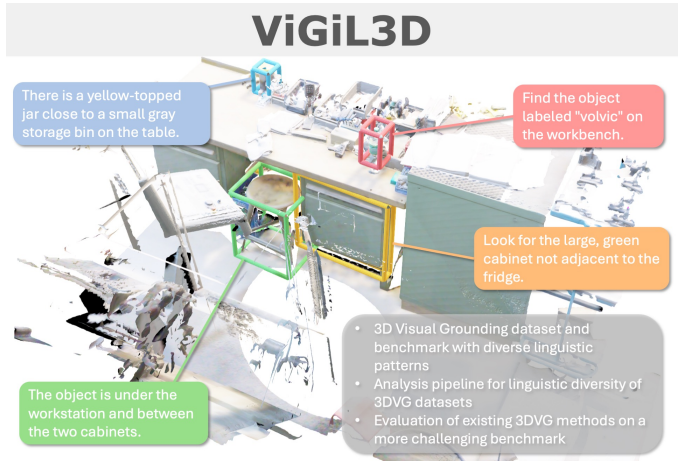
Thank you! Questions?

3DLG | 3D Language & Generation

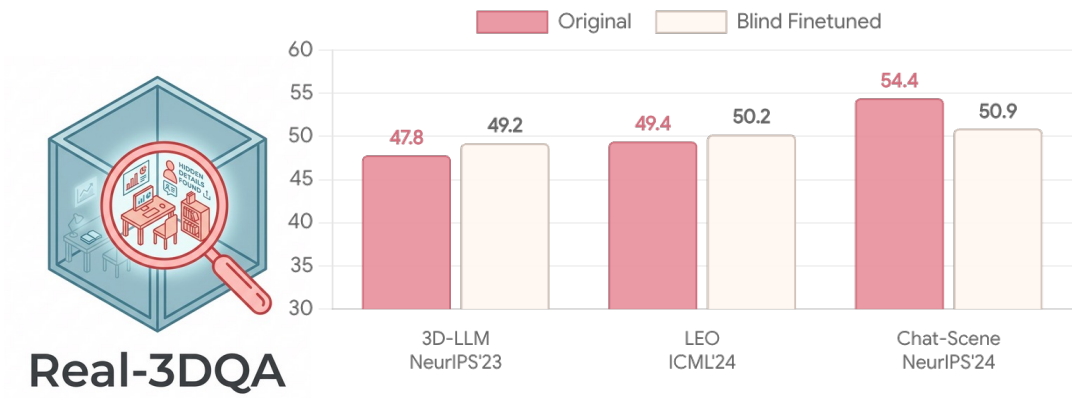
https://3dlg-hcvc.github.io/



Austin Wang



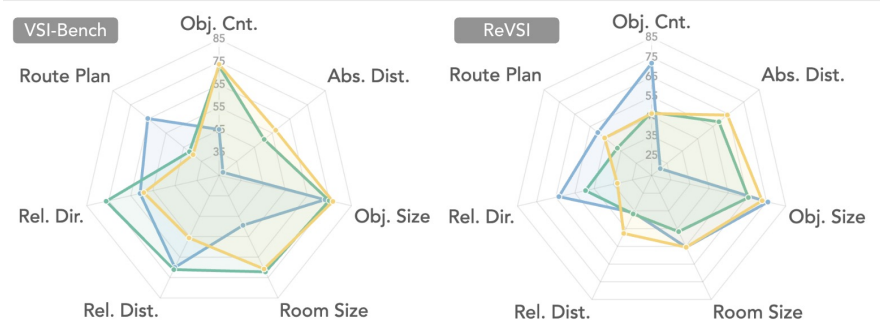
https://3dlg-hcvc.github.io/vigil3d/



https://real-3dqa.github.io/

Yiming Zhang

ReVSI



https://3dlg-hcvc.github.io/revsi/

Workshop on 3D-LLM/VLA

June 3rd pm - Mile High 1CD

UNIVERSITY OF OXFORD

Xianzheng Ma