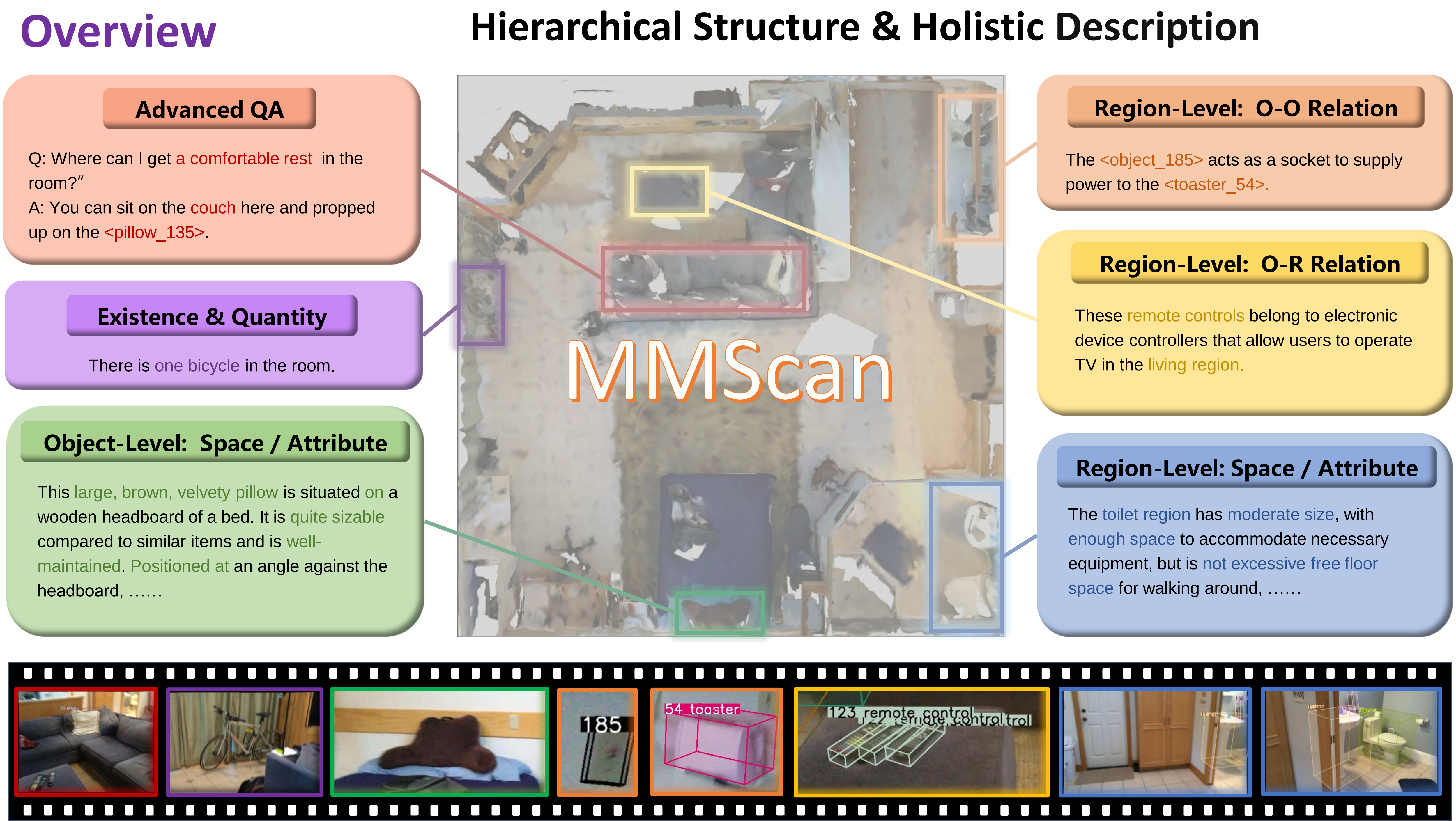
MMScan: A Multi-Modal 3D Scene Dataset with Hierarchical Grounded Language Annotations

Ruiyuan Lyu*, Jingli Lin*, Tai Wang*, Shuai Yang*, Xiaohan Mao, Yilun Chen,

Runsen Xu, Haifeng Huang, Chenming Zhu, Dahua Lin, Jiangmiao Pang†


Overview
Hierarchical Structure & Holistic Description
Advanced QA
Q: Where can I get a comfortable rest in the room?"
A: You can sit on the couch here and propped up on the <pillow_135>.
Existence & Quantity
There is one bicycle in the room.
Object-Level: Space / Attribute
This large, brown, velvety pillow is situated on a wooden headboard of a bed. It is quite sizable compared to similar items and is well-maintained. Positioned at an angle against the headboard, ……
MMScan
Region-Level: O-O Relation
The <object_185> acts as a socket to supply power to the <toaster_54>.
Region-Level: O-R Relation
These remote controls belong to electronic device controllers that allow users to operate TV in the living region.
Region-Level: Space / Attribute
The toilet region has moderate size, with enough space to accommodate necessary equipment, but is not excessive free floor space for walking around, ……
185
54 toaster
123 remote control

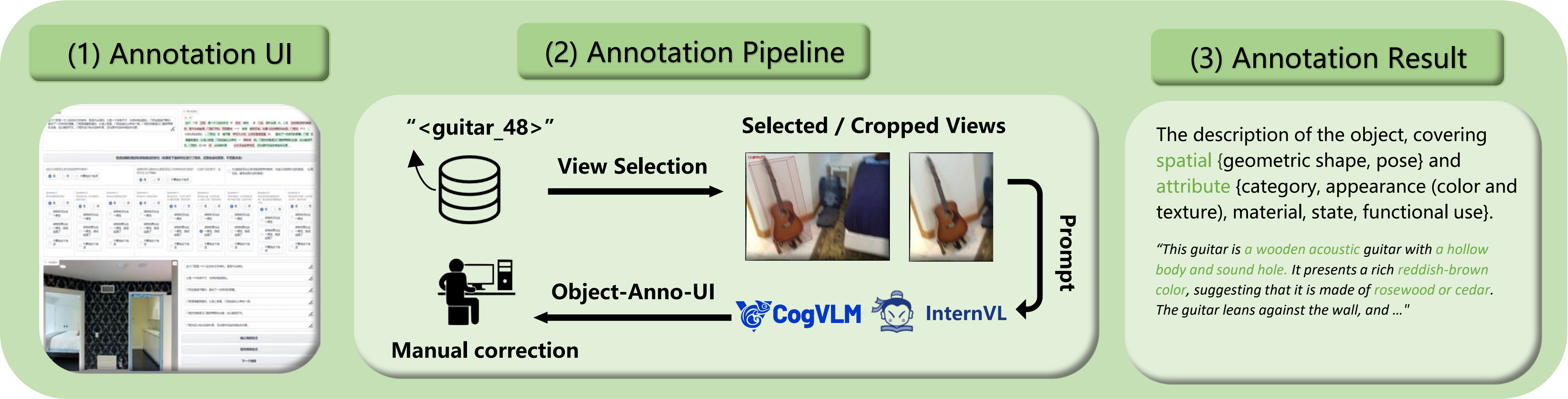
Overview

Largest Ever Multi-modal 3D Scene Dataset

Dataset	#Scans	#Language	#Tokens	Correspondence	Focus	Annotation
ScanRefer [12]	0.7k	11k	1.18M	Sent.-Obj.	Natural	Human
Nr3D [6]	0.7k	42k	0.62M	Sent.-Obj.	Natural	Human
Sr3D [6]	0.7k	115k	1.48M	Sent.-Obj.	OO-Space	Template
ScanQA [7]	0.8k	41k	-	Sent.-Obj.	QA	AutoGen+Human
SQA3D [35]	0.7k	53.8k	-	Sent.-Obj.	Situated QA	Human
ScanScribe [57]	1.2k	278K	18.49M	Sent.-Obj.	Description	GPT
Multi3DRef [52]	0.7k	62K	1.2M	Sent.-Multi-Obj.	Multi-Obj.	GPT+Human
EmbodiedScan [47]	5.2k	970k	-	Sent.-Obj.	OO-Space	Template
RIORefer [36]	1.4k	63.6k	0.94M	Sent.-Obj.	Natural	Human
ARKitSceneRefer [32]	1.6k	15.6k	0.22M	Sent.-Obj.	Natural	Human
MMScan (Ours)	5.2k	6.9M	114M	Pharse-Obj./Reg.	Holistic	GPT+Temp.+Human

Dataset: Meta-annotation

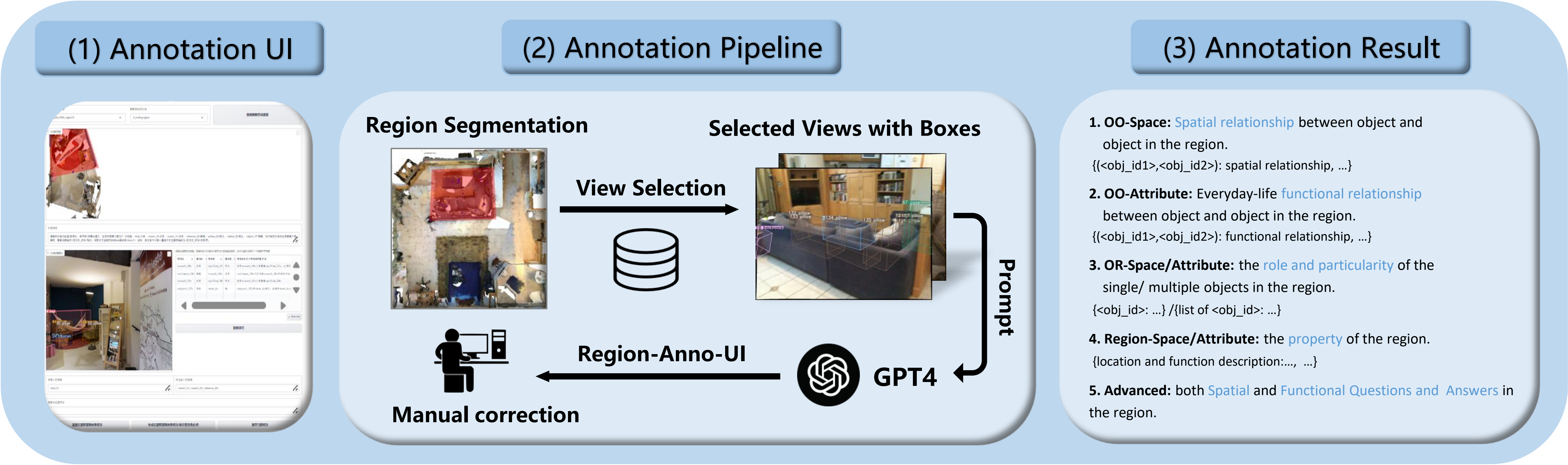
Top-down Logic, Human-in-the-loop Pipeline



The annotation result includes both spatial (geometric shape, pose) and attribute (category, appearance, material, state, etc.) descriptions for the object.

Dataset: Meta-annotation

Top-down Logic, Human-in-the-loop Pipeline



The annotation result includes regions' inherent properties, object-object/region relationships and advanced QA.

Dataset: Post-processing

Object / Region- level, Single / Inter- target, Space / Attribute

QA (12.79%): Answering attribute relation of two objects; Answering one object in attribute relation to the other.

VG (14.43%): Grounding one object in attribute relation to the anchor.

OO-Attribute

QA (23.02%): Answering space relation of two objects; Answering one object in space relation to the other.

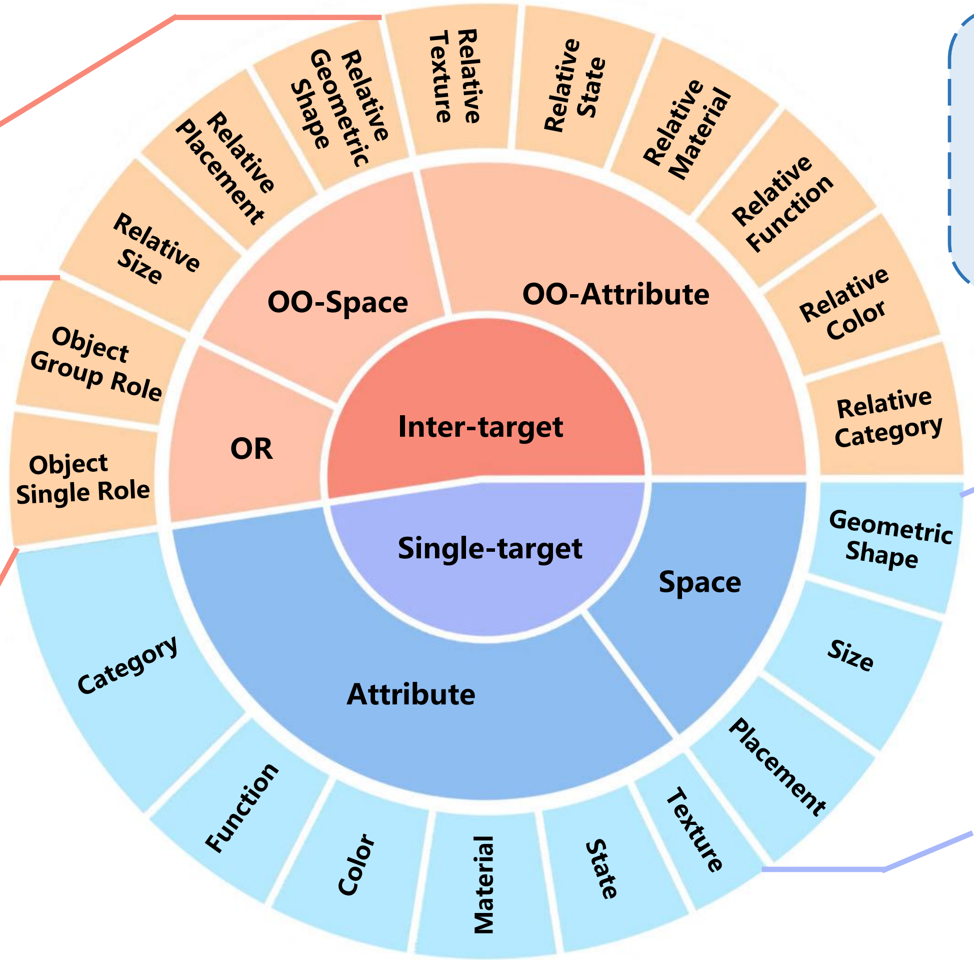
VG (39.30%): Grounding one object in space relation to the anchor.

OO-Space

QA (5.68%): Answering (single/group) role of (single/multiple) objects.

VG (5.18%): Grounding (single/multiple) objects based on (single/group) role.

OR-Space/Attribute



QA (8.43%): Answering existence /quantity of objects with specific spatial/attribute characteristics.

Existence & Quantity (Special Case)

QA (17.13%): Answering the spatial features of the entities.

VG (8.93%): Grounding single /multiple entities based on the spatial features.

Space (O/R)

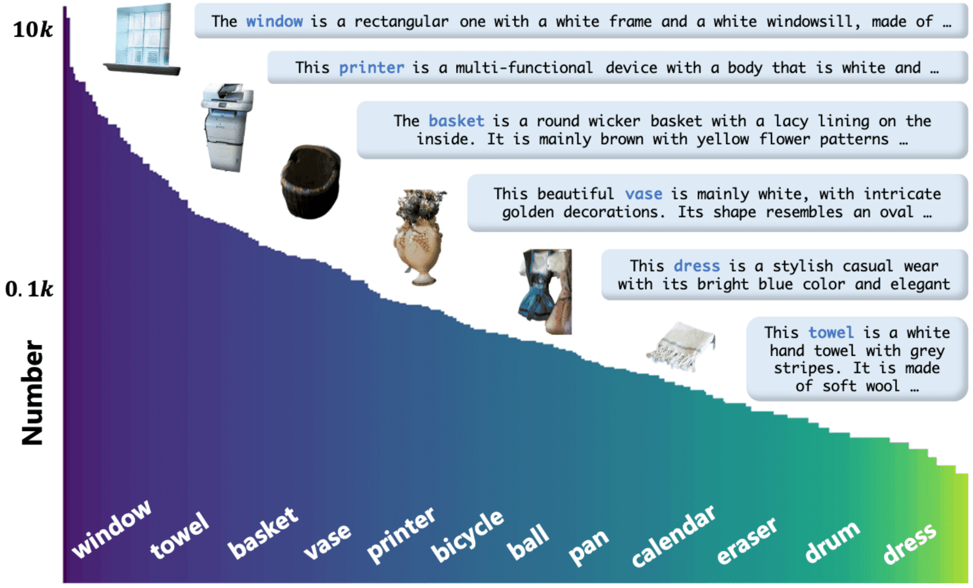
QA (30.78%): Answering the attributes of the entities.

VG (32.16%): Grounding single /multiple entities based on the attributes.

Attribute (O/R)

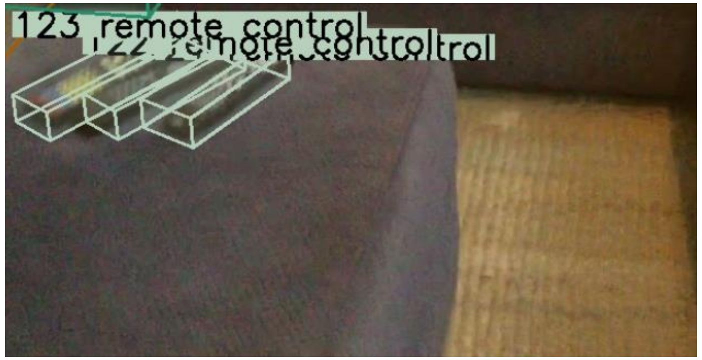
Dataset: Statistic

6.9M Language Annotations & 114M Tokens



Benchmark

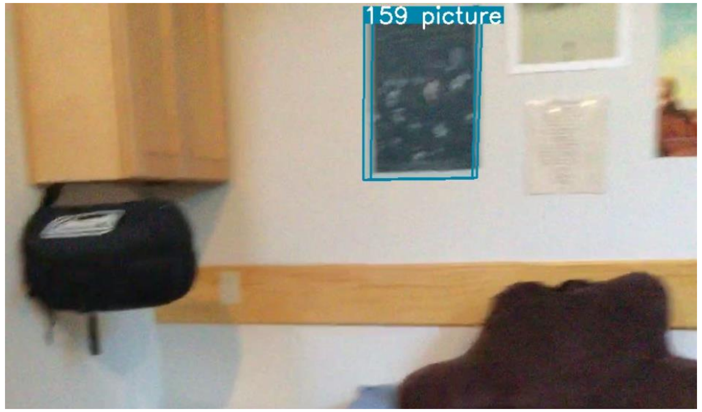
Complex semantic information / 9-DoF box / Uncertain number of GTs



Grounding all electronic products in the room

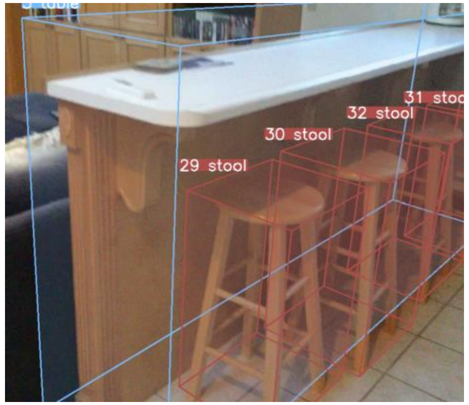
Are these two objects different in material?

What can I see when I enter a room and look up?



Grounding the dark-color picture in the bedroom

Grounding items used to ensure safety when going up the stairs



How many stools are there below the table?

Benchmark

Table 1: 3D visual grounding benchmark on MMScan.

Methods	Overall				Single-target		Inter-target		
	AP_{25}	AR_{25}	AP_{50}	AR_{50}	ST-attr	ST-space	OO-attr	OO-space	OR
ScanRefer [8]	3.83	42.40	1.37	20.96	1.44	2.84	5.22	4.32	1.12
BUTD-DETR [23]	2.29	65.61	0.84	33.11	4.79	2.04	1.49	1.75	11.87
ViL3DRef [11]	5.17	**72.50**	2.07	**51.61**	6.29	4.20	7.89	5.29	6.81
EmbodiedScan [39]	**10.49**	47.21	**2.94**	21.76	7.44	7.53	13.65	11.19	7.74

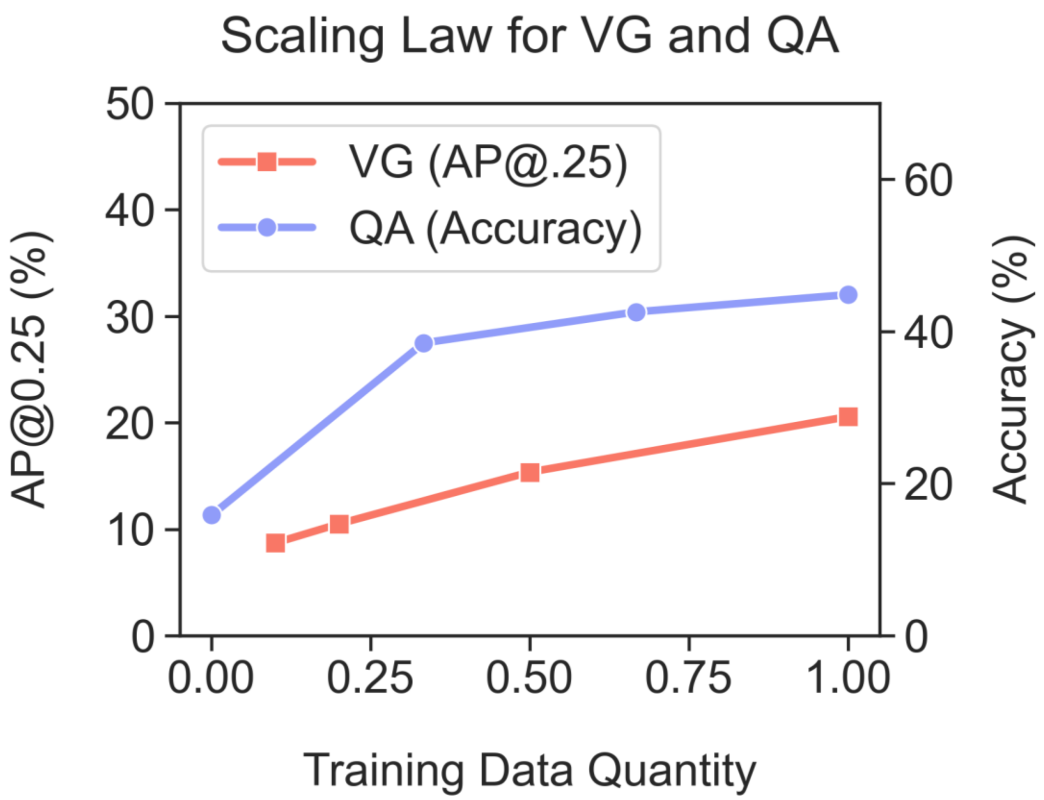
Table 2: 3D question-answering benchmark on MMScan.

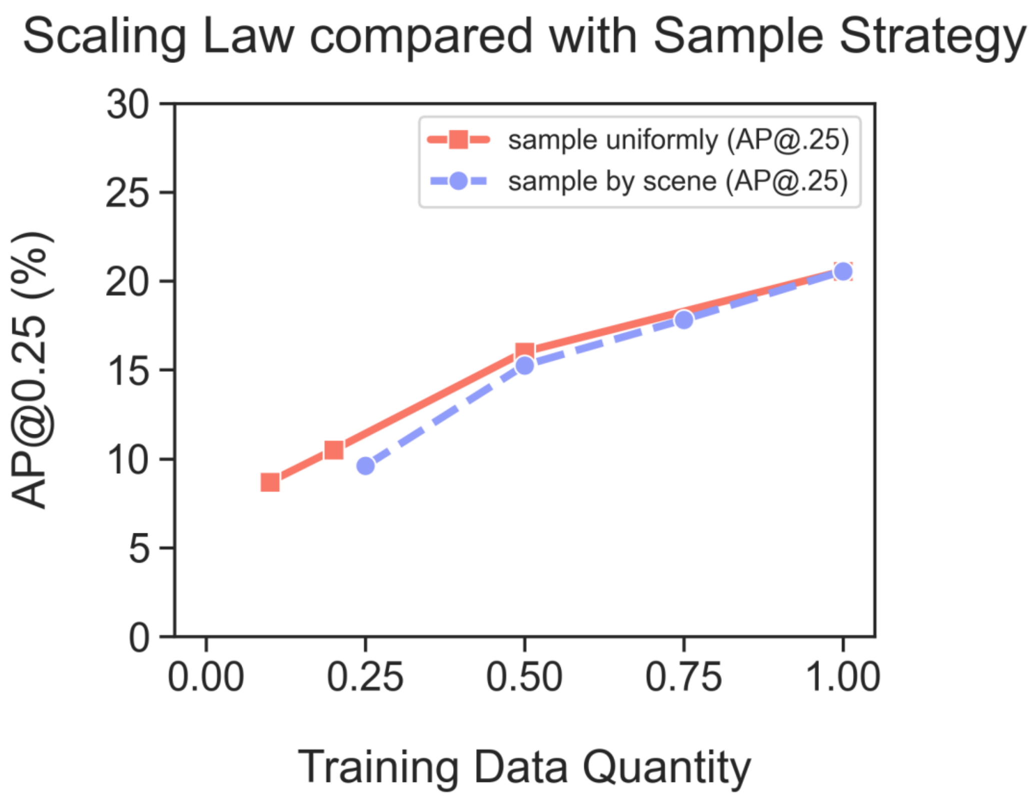
Methods	Setting	Overall	Single-target		Inter-target			Advanced	Data-driven Metrics		Traditional Metrics				
			ST-attr	ST-space	OO-attr	OO-space	OR		SimCSE	S.-BERT	B-1.	B-4.	R.-L	MET.	EM@1
3D-LLM [20]	Zero-Shot	28.6	37.8	18.8	13.7	26.3	15.4	20.8	40.4	40.3	13.4	1.5	17.3	6.0	6.2 (19.6)
Chat3D-v2 [21]	Zero-Shot	27.9	38.1	18.3	9.3	22.4	13.5	25.4	45.4	46.3	18.0	3.0	22.9	7.5	10.2 (19.6)
LL3DA [10]	Zero-Shot	15.8	15.5	14.7	14.2	25.2	4.3	6.4	40.7	43.6	5.4	2.1	16.4	4.4	8.3 (19.4)
LEO [22]	Zero-Shot	22.2	28.9	17.6	18.1	20.4	15.0	16.3	40.4	41.0	11.0	0.7	17.1	4.9	9.6 (18.7)
LL3DA [10]	Fine-tuning	38.5	40.4	46.2	14.7	47.1	26.4	7.1	65.3	67.0	26.4	8.5	44.3	14.7	30.2 (37.6)
LEO [22]	Fine-tuning	**47.8**	55.5	49.5	36.1	45.6	32.1	38.4	71.2	72.2	32.0	12.5	52.1	17.7	36.6 (44.5)

We can observe that the new benchmark is challenging due to more complex language understanding.

Our data also shows great importance in tuning current 3D-LLMs to have more satisfactory performance.

Experiment: Scaling Law





On both two benchmarks, models' performance improve while the training data quantity increase.
Scene diversity matters and scaling up scene diversity is more effective.

Experiment: Training Stronger Models

3D Question Answering Baseline on Traditional Benchmarks

Methods	ScanQA (val)				SQA3D (test)
	B-4.	R.-L.	MET.	EM@1	EM@1
baseline	10.5	39.2	15.1	23.1 (39.0)	51.6 (54.1)
+ meta-ann. captions	10.7	41.2	14.2	23.3 (39.3)	52.7 (54.8)
+ scene captions	12.3	46.4	18.1	24.3 (46.6)	53.2 (55.4)
+ all captions	12.7	48.1	19.8	24.7 (48.9)	54.1 (56.8)

3D Visual Grounding Baseline on Traditional Benchmarks

Methods	HF	Overall	
		AP_{25}	AP_{50}
baseline	-	37.27	17.78
pre-training	✗	42.18	21.84
co-training	✗	42.96	22.77
pre-training	✓	42.49	22.17
co-training	✓	**44.44**	**23.69**

Both VG/QA baselines have a great improvement after training with MMScan dataset.

Experiment: In-the-Wild Test

Test our models trained with MMScan.

Where is the plant in the room?

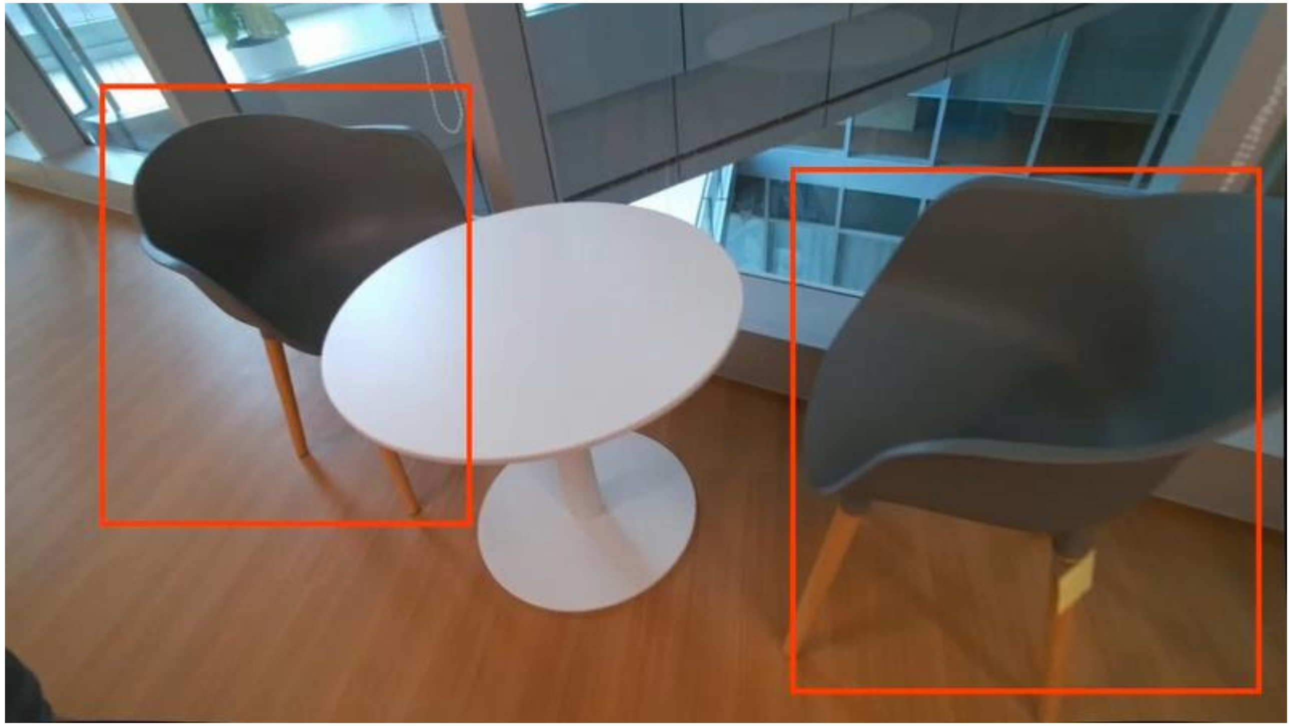
On top of the end table

How many white round tables are there?

Two

Experiment: In-the-Wild Test

Test our models trained with MMScan.



Can you describe the chair close to the white round table?

The chairs have grey seats and backrests, which appear to be made of a molded plastic or composite material with a matte finish. The backrest is also contoured to provide support.

What color is the chair next to the white round table?

Gray

Experiment: In-the-Wild Test

Test our models trained with MMScan.

What is on the blue cabinet?

Microwave

Where can I heat the food?

Microwave

Can you describe the object on the white wooden table?

The object on the table is a plotted plant with variegated leaves, likely a Sansevieria, commonly known as a snake plant

Experiment: In-the-Wild Test

Test our models trained with MMScan.

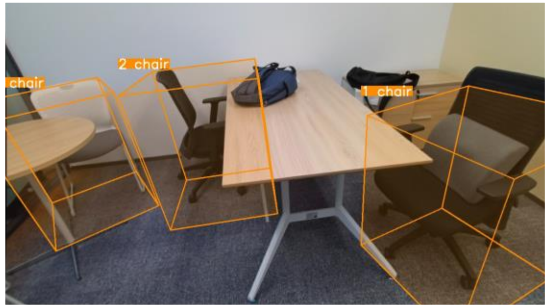
I'm facing the front of blue cabinet, which direction should I go if I want to sit on the sofa?

Left

Experiment: In-the-Wild Test

Test our models trained with MMScan.

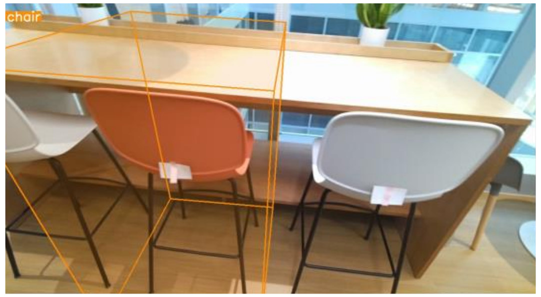
The X is an ergonomically designed furniture that provides a comfortable sitting experience. Please find the X.

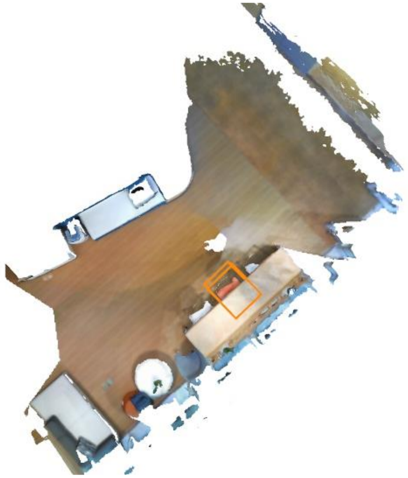


Find all the items that can entertain us in the room.

I am thirsty. Find all the items that can help me with it.

Find the orange chair in the room.





Thank you!