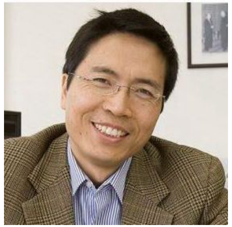
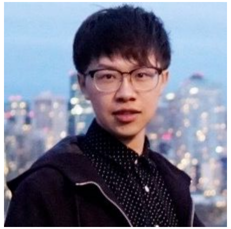
SQA3D: Situated Question Answering in 3D Scenes

Xiaojian Ma *[1,2], Silong Yong *[1,3], Zilong Zheng [1], Qing Li [1], Yitao Liang [1,4],
Song-Chun Zhu [1,2,3,4], Siyuan Huang [1]

[1] BIGAI [2] UCLA [3] Tsinghua University [4] Peking University * Equal contribution

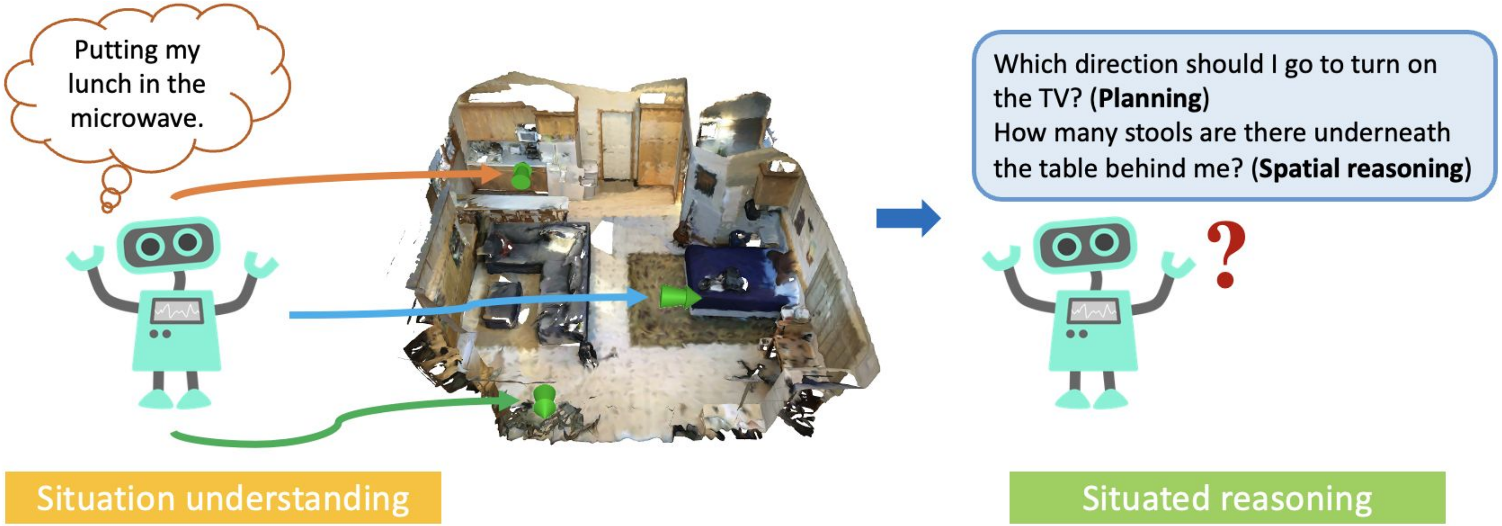
ICLR 2023 arxiv.org/abs/2210.07474





Motivation

We study the problem of **embodied scene understanding** to bridge the gap between *embodied AI* and 3D *scene understanding*: an agent need to understand its surroundings (situations) from a *dynamic & egocentric* view, then accomplish reasoning & planning tasks *accordingly* (situated reasoning).

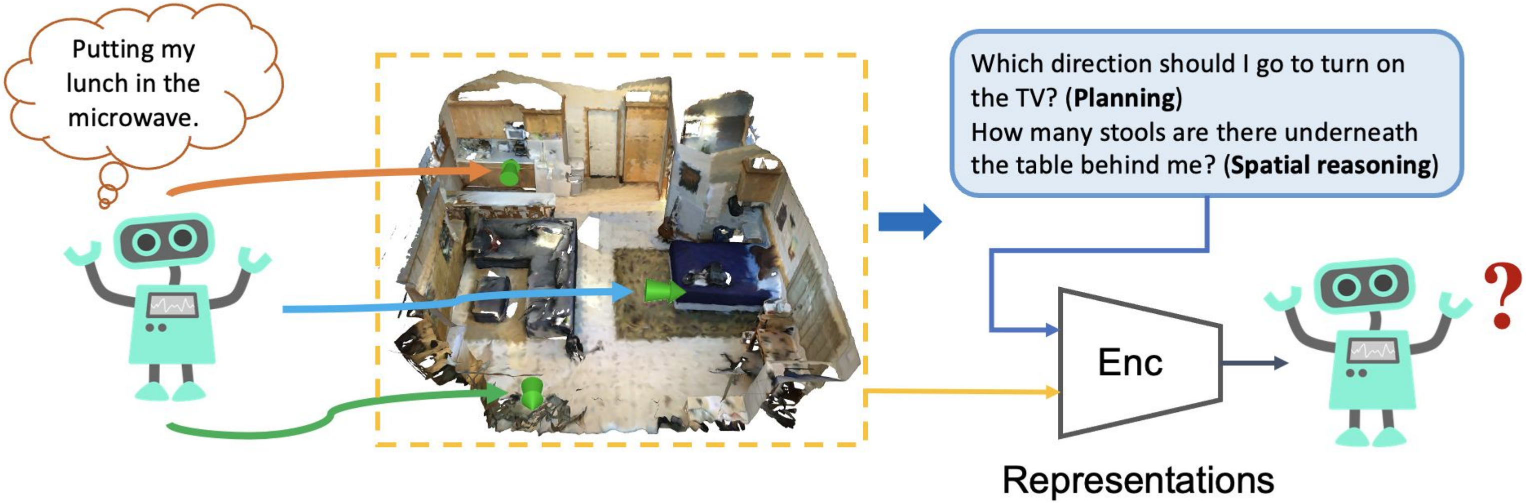


Putting my lunch in the microwave.

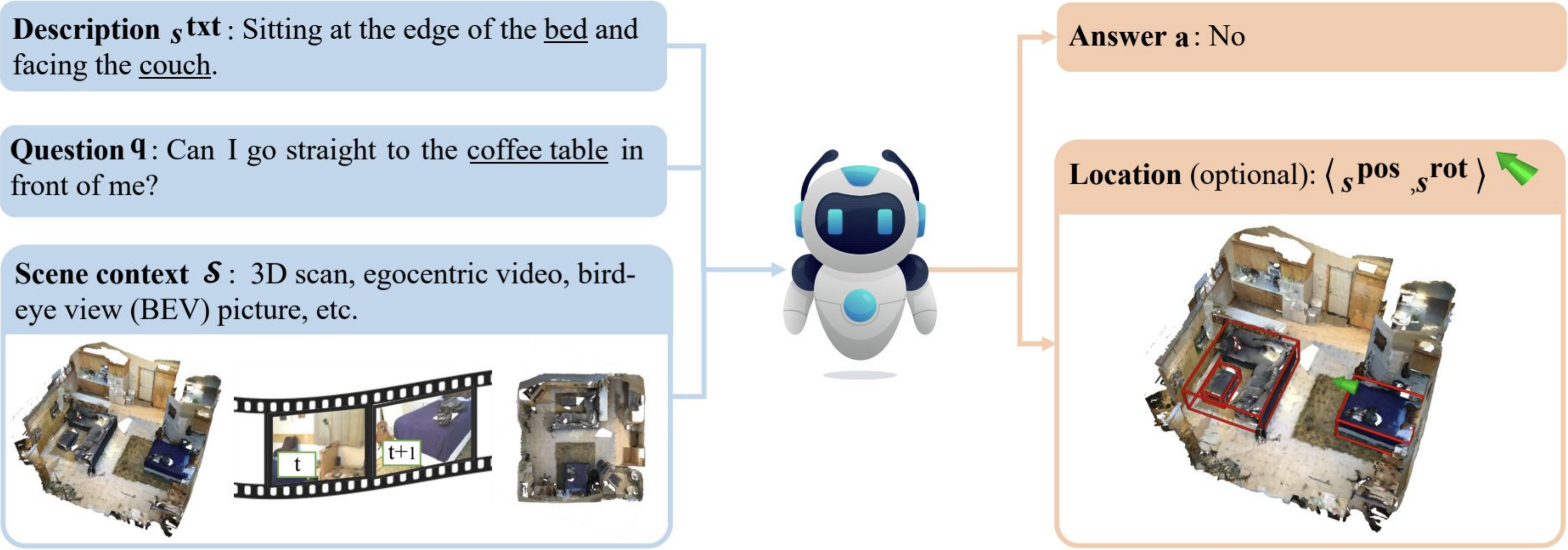
Which direction should I go to turn on the TV? (**Planning**)
How many stools are there underneath the table behind me? (**Spatial reasoning**)

Motivation

We believe, truly **generalist representations** should support such challenging **situation understanding** and **situated reasoning** in **embodied**, **3D** scenes.


Putting my lunch in the microwave.
Which direction should I go to turn on the TV? (**Planning**)
How many stools are there underneath the table behind me? (**Spatial reasoning**)
Enc
?
Representations

What is SQA3D?



Description s^{txt}: Sitting at the edge of the bed and facing the couch.

Question q: Can I go straight to the coffee table in front of me?

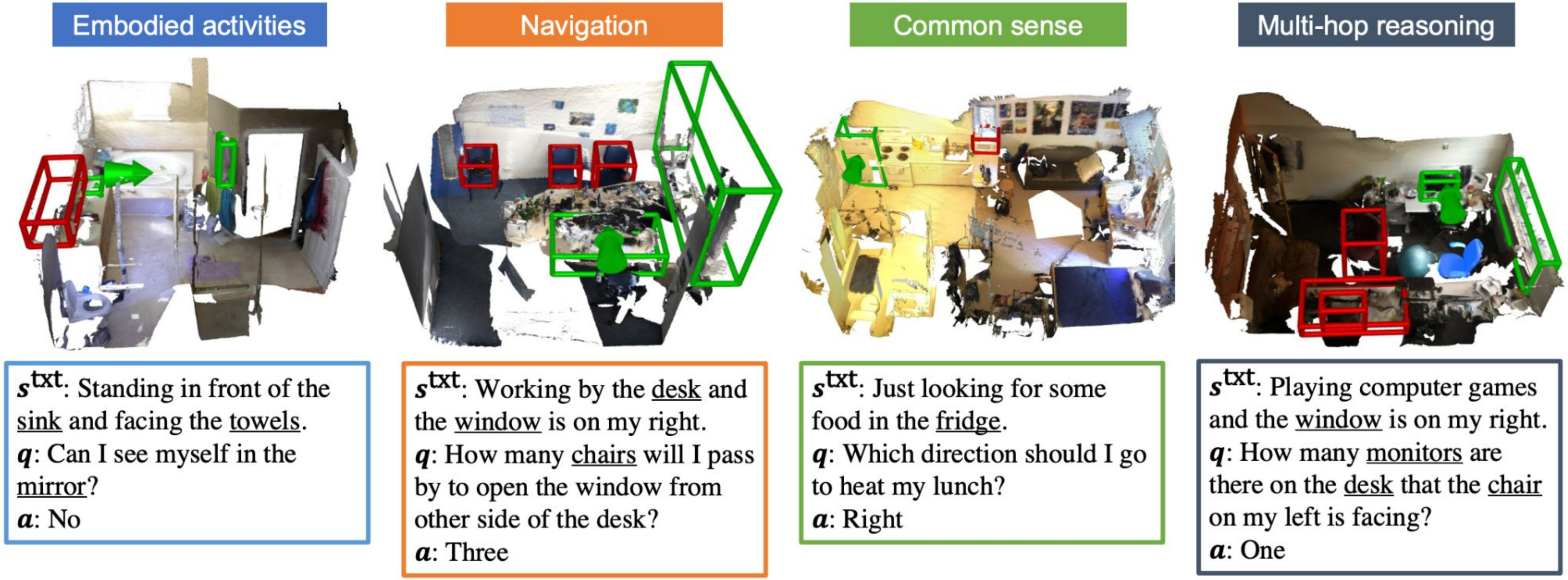
Scene context S: 3D scan, egocentric video, bird-eye view (BEV) picture, etc.

Answer a: No

Location (optional): $\langle s^{\text{pos}}, s^{\text{rot}} \rangle$

Given a **scene context** (3D scan, egocentric video, BEV pictures…), the agent needs to understand its situation from a **description**, then answer a **question**.

Examples from SQA3D



Embodied activities

s^{txt}: Standing in front of the sink and facing the towels.
q: Can I see myself in the mirror?
a: No

Navigation

s^{txt}: Working by the desk and the window is on my right.
q: How many chairs will I pass by to open the window from other side of the desk?
a: Three

Common sense

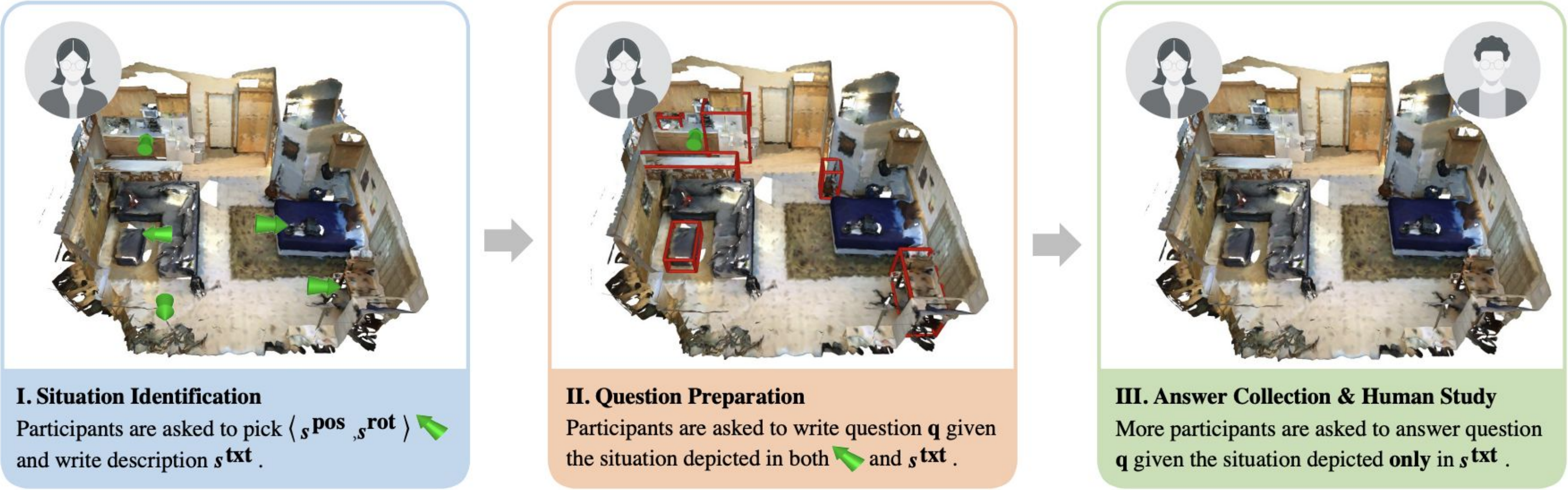
s^{txt}: Just looking for some food in the fridge.
q: Which direction should I go to heat my lunch?
a: Right

Multi-hop reasoning

s^{txt}: Playing computer games and the window is on my right.
q: How many monitors are there on the desk that the chair on my left is facing?
a: One

The green boxes indicate relevant objects in situation description while red boxes are for the questions. The virtual avatar ➡ marks the actual location of the agent.

Building SQA3D



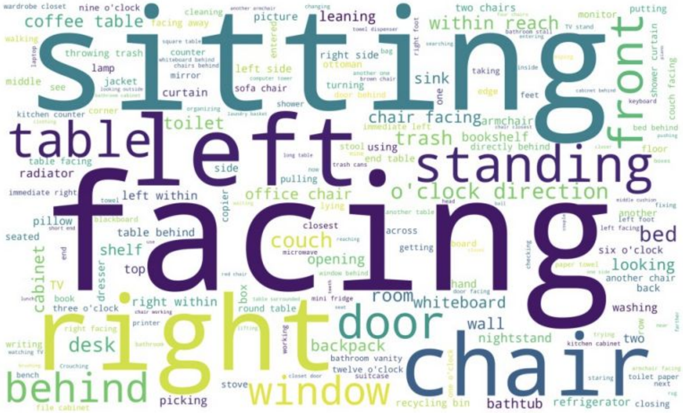
I. Situation Identification
Participants are asked to pick $\langle s^{\textbf{pos}}, s^{\textbf{rot}} \rangle$ ↖ and write description $s^{\textbf{txt}}$.

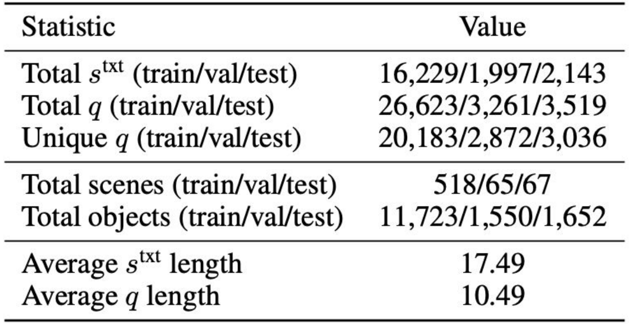
II. Question Preparation
Participants are asked to write question **q** given the situation depicted in both ↖ and $s^{\textbf{txt}}$.

III. Answer Collection & Human Study
More participants are asked to answer question **q** given the situation depicted **only** in $s^{\textbf{txt}}$.

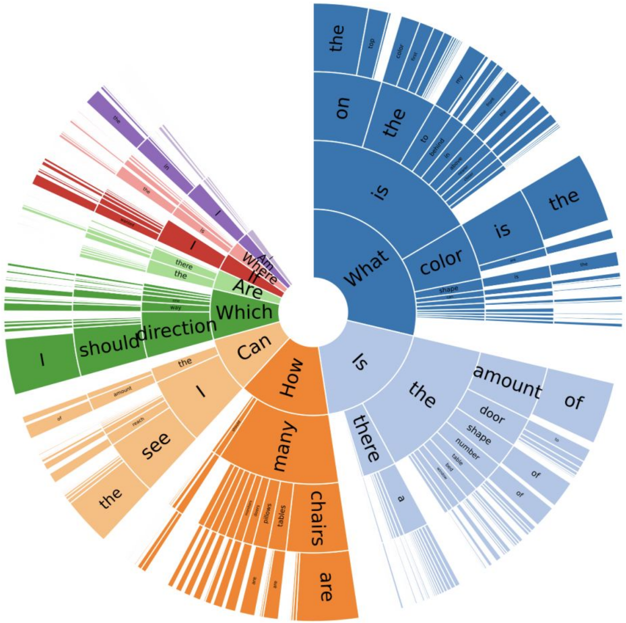
We recruit our workforces from Amazon Mechanical Turk (AMT). A multi-staged collection strategy is adopted to ensure manageable workload and higher data quality.

Dataset statistics



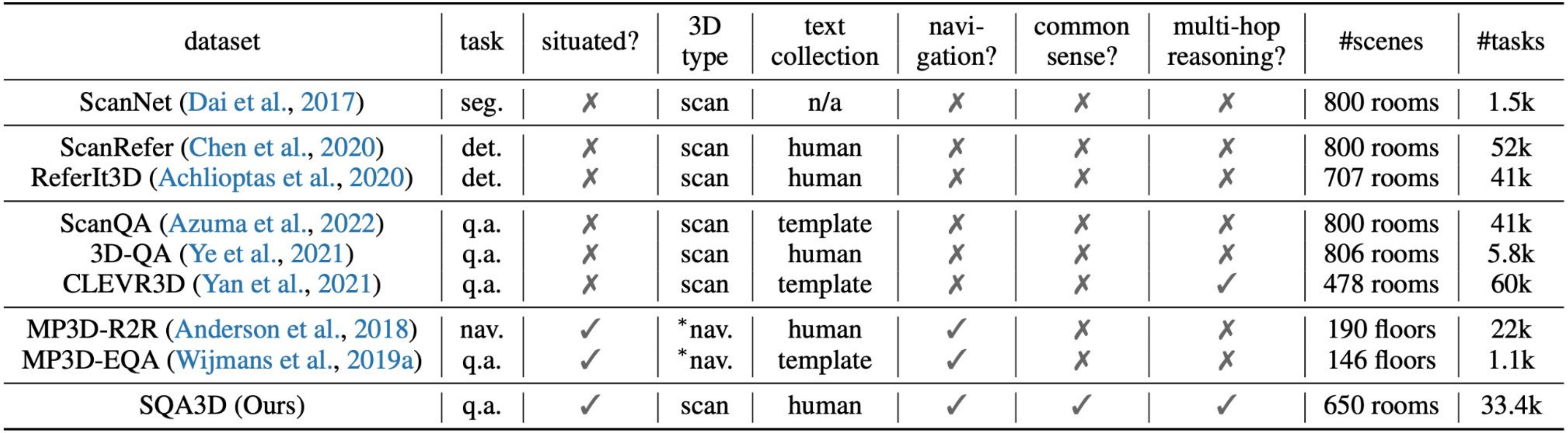


Statistic	Value
Total s^{txt} (train/val/test)	16,229/1,997/2,143
Total q (train/val/test)	26,623/3,261/3,519
Unique q (train/val/test)	20,183/2,872/3,036
Total scenes (train/val/test)	518/65/67
Total objects (train/val/test)	11,723/1,550/1,652
Average s^{txt} length	17.49
Average q length	10.49



Compared to counterparts with template-based text, SQA3D offers more **diverse** questions thanks to our AMT workforces.

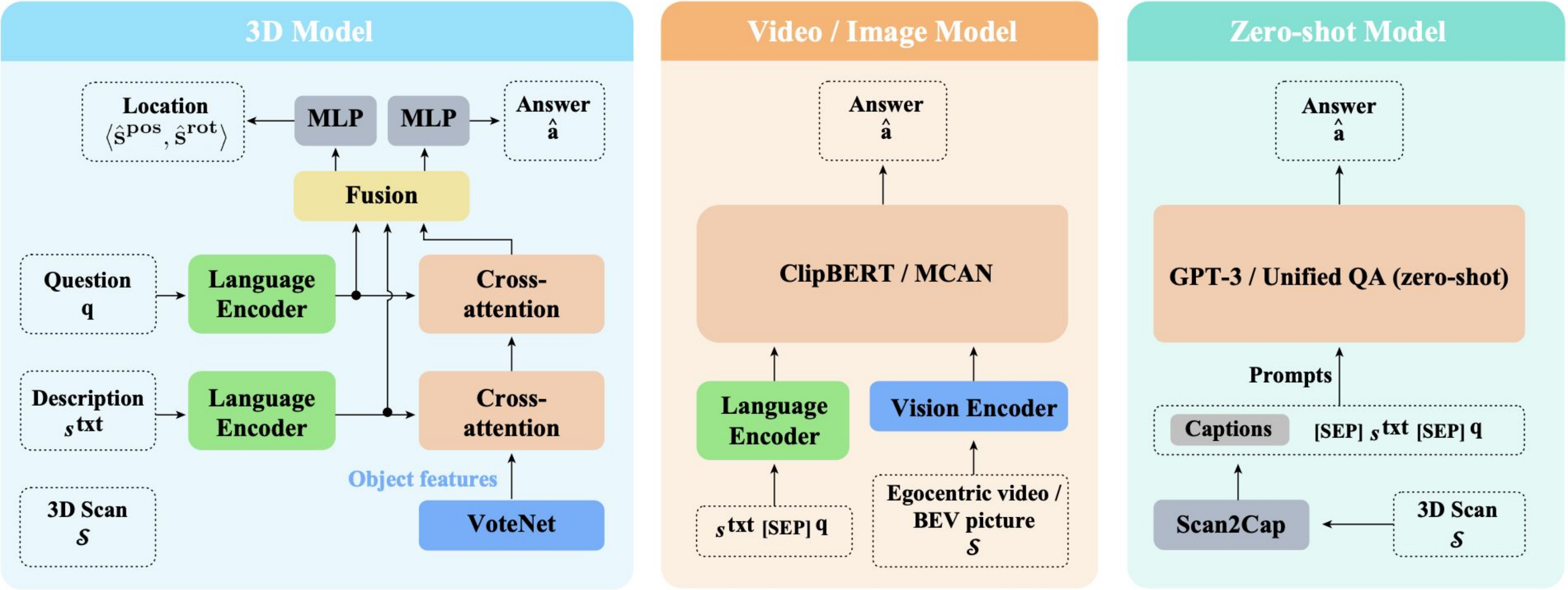
Comparison to related benchmarks



dataset	task	situated?	3D type	text collection	navigation?	common sense?	multi-hop reasoning?	#scenes	#tasks
ScanNet (Dai et al., 2017)	seg.	✗	scan	n/a	✗	✗	✗	800 rooms	1.5k
ScanRefer (Chen et al., 2020)	det.	✗	scan	human	✗	✗	✗	800 rooms	52k
ReferIt3D (Achlioptas et al., 2020)	det.	✗	scan	human	✗	✗	✗	707 rooms	41k
ScanQA (Azuma et al., 2022)	q.a.	✗	scan	template	✗	✗	✗	800 rooms	41k
3D-QA (Ye et al., 2021)	q.a.	✗	scan	human	✗	✗	✗	806 rooms	5.8k
CLEVR3D (Yan et al., 2021)	q.a.	✗	scan	template	✗	✗	✓	478 rooms	60k
MP3D-R2R (Anderson et al., 2018)	nav.	✓	*nav.	human	✓	✗	✗	190 floors	22k
MP3D-EQA (Wijmans et al., 2019a)	q.a.	✓	*nav.	template	✓	✗	✗	146 floors	1.1k
SQA3D (Ours)	q.a.	✓	scan	human	✓	✓	✓	650 rooms	33.4k

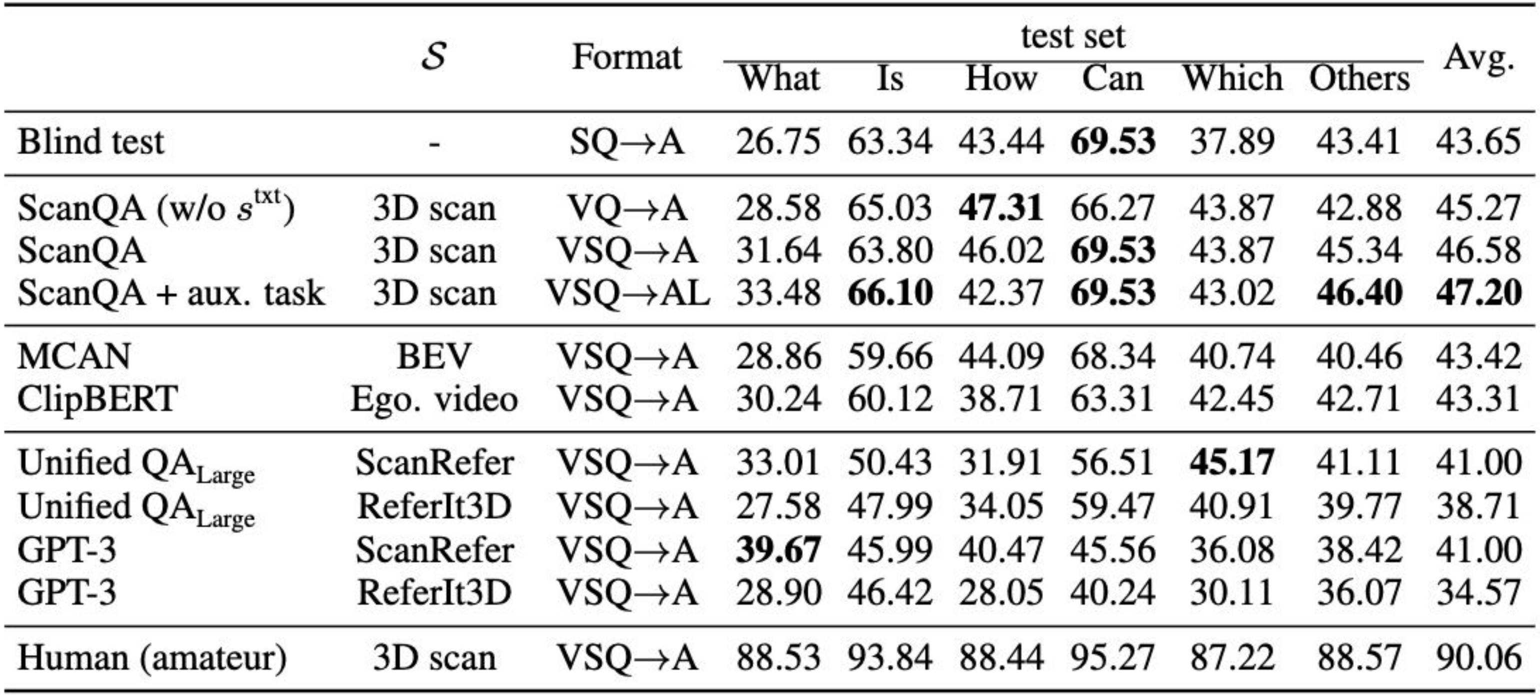
To the best of our knowledge, SQA3D is the **largest** dataset combines the best of both worlds: **situated reasoning**, **human-written text**, and **diverse & challenging problems**.

Models for SQA3D?



Canonical question answering models for 3D scan, video and image input are evaluated. We further explore **zero-shot large models** (GPT-3, Unified QA) by converting the 3D scene into *captions*.

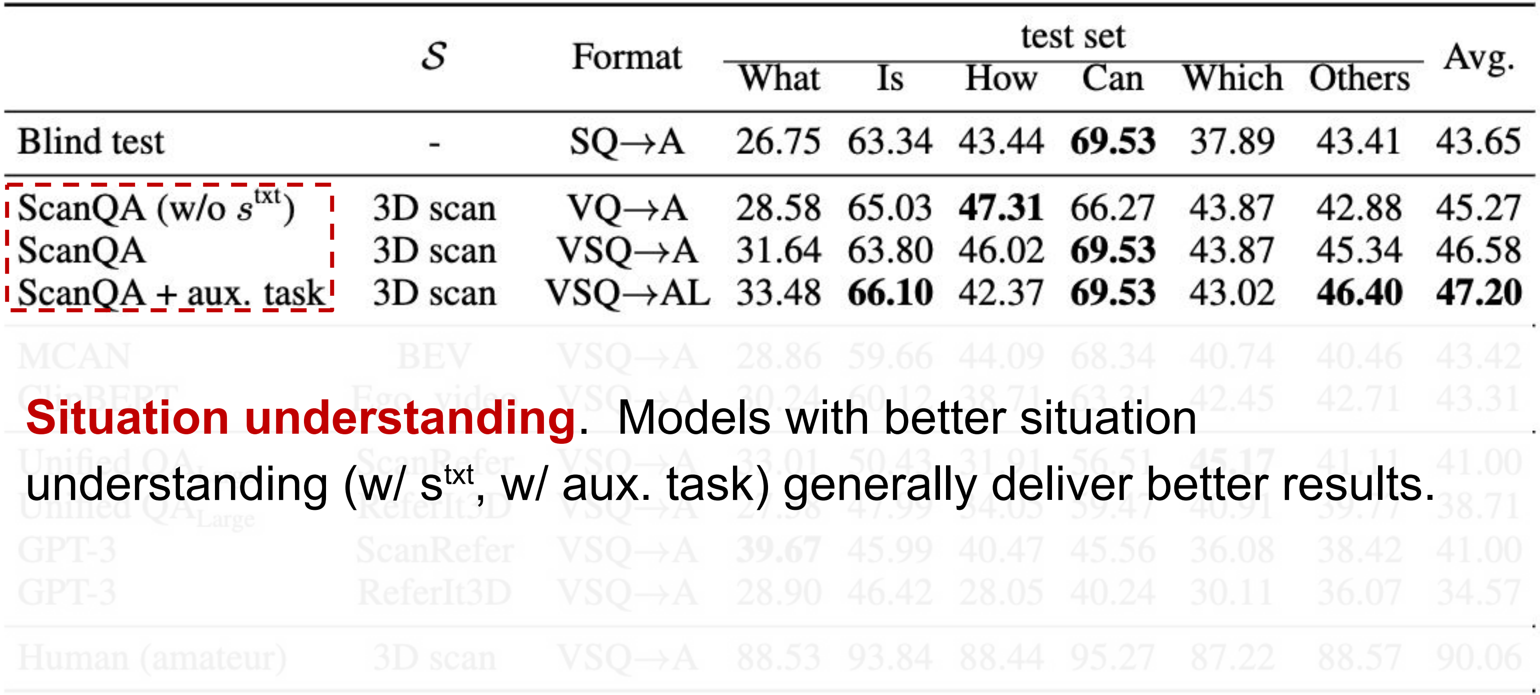
Benchmarking: quantitative results



	$\mathcal{S}$	Format	test set						Avg.
			What	Is	How	Can	Which	Others	
Blind test	-	SQ→A	26.75	63.34	43.44	**69.53**	37.89	43.41	43.65
ScanQA (w/o s^{txt})	3D scan	VQ→A	28.58	65.03	**47.31**	66.27	43.87	42.88	45.27
ScanQA	3D scan	VSQ→A	31.64	63.80	46.02	**69.53**	43.87	45.34	46.58
ScanQA + aux. task	3D scan	VSQ→AL	33.48	**66.10**	42.37	**69.53**	43.02	**46.40**	**47.20**
MCAN	BEV	VSQ→A	28.86	59.66	44.09	68.34	40.74	40.46	43.42
ClipBERT	Ego. video	VSQ→A	30.24	60.12	38.71	63.31	42.45	42.71	43.31
Unified QA$_{\text{Large}}$	ScanRefer	VSQ→A	33.01	50.43	31.91	56.51	**45.17**	41.11	41.00
Unified QA$_{\text{Large}}$	ReferIt3D	VSQ→A	27.58	47.99	34.05	59.47	40.91	39.77	38.71
GPT-3	ScanRefer	VSQ→A	**39.67**	45.99	40.47	45.56	36.08	38.42	41.00
GPT-3	ReferIt3D	VSQ→A	28.90	46.42	28.05	40.24	30.11	36.07	34.57
Human (amateur)	3D scan	VSQ→A	88.53	93.84	88.44	95.27	87.22	88.57	90.06

***aux. task**: we introduce an additional location prediction task to encourage better situation understanding.

Benchmarking: quantitative results

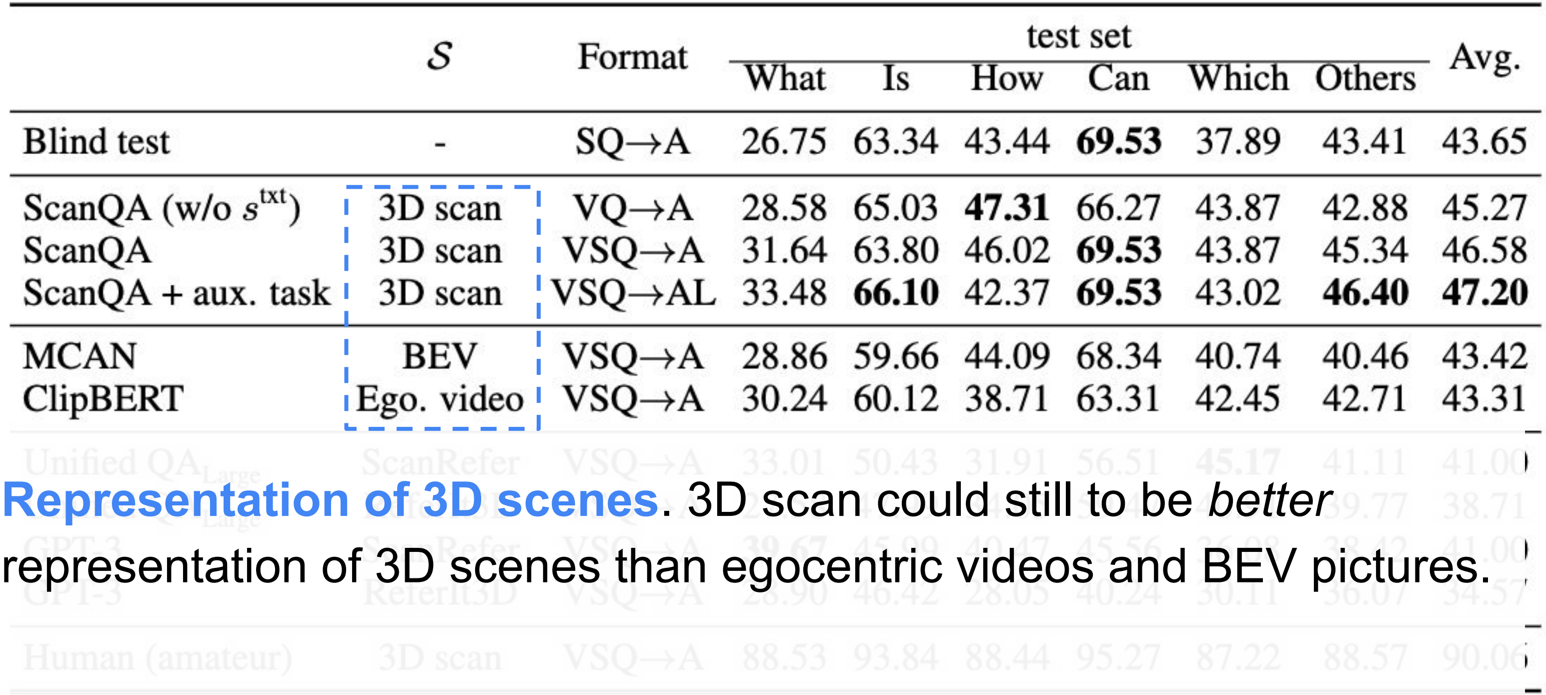


	$\mathcal{S}$	Format	test set						Avg.
			What	Is	How	Can	Which	Others	
Blind test	-	SQ→A	26.75	63.34	43.44	**69.53**	37.89	43.41	43.65
ScanQA (w/o s^{txt})	3D scan	VQ→A	28.58	65.03	**47.31**	66.27	43.87	42.88	45.27
ScanQA	3D scan	VSQ→A	31.64	63.80	46.02	**69.53**	43.87	45.34	46.58
ScanQA + aux. task	3D scan	VSQ→AL	33.48	**66.10**	42.37	**69.53**	43.02	**46.40**	**47.20**
MCAN	BEV	VSQ→A	28.86	59.66	44.09	68.34	40.74	40.46	43.42
ClipBERT	Ego. video	VSQ→A	30.24	60.12	38.71	63.31	42.45	42.71	43.31
Unified QA$_{\text{Large}}$	ScanRefer	VSQ→A	33.01	50.43	31.91	56.51	**45.17**	41.11	41.00
Unified QA$_{\text{Large}}$	ReferIt3D	VSQ→A	27.58	47.99	34.05	59.47	40.91	39.77	38.71
GPT-3	ScanRefer	VSQ→A	**39.67**	45.99	40.47	45.56	36.08	38.42	41.00
GPT-3	ReferIt3D	VSQ→A	28.90	46.42	28.05	40.24	30.11	36.07	34.57
Human (amateur)	3D scan	VSQ→A	88.53	93.84	88.44	95.27	87.22	88.57	90.06

Situation understanding. Models with better situation understanding (w/ s^{txt}, w/ aux. task) generally deliver better results.

***aux. task**: we introduce an additional location prediction task to encourage better situation understanding.

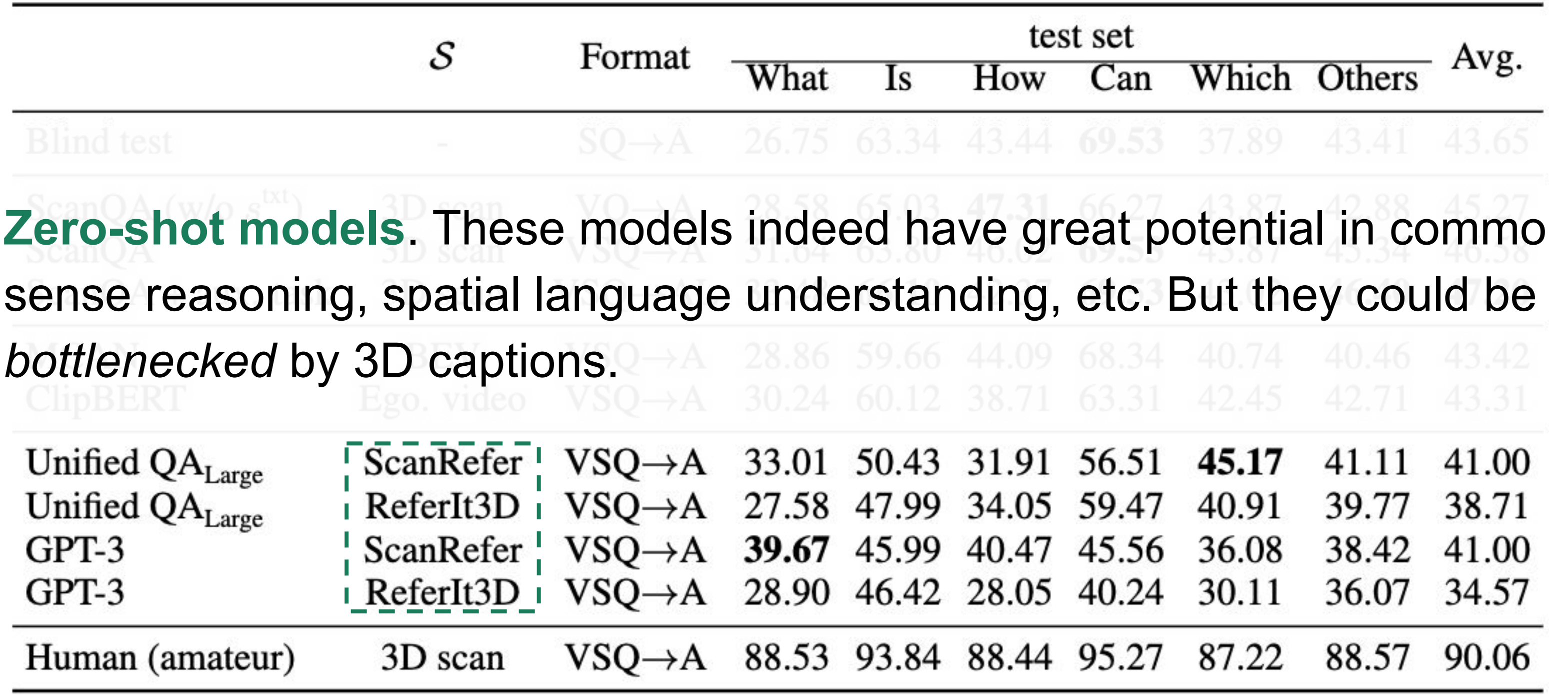
Benchmarking: quantitative results



	$\mathcal{S}$	Format	test set						Avg.
			What	Is	How	Can	Which	Others	
Blind test	-	SQ→A	26.75	63.34	43.44	**69.53**	37.89	43.41	43.65
ScanQA (w/o s^{txt})	3D scan	VQ→A	28.58	65.03	**47.31**	66.27	43.87	42.88	45.27
ScanQA	3D scan	VSQ→A	31.64	63.80	46.02	**69.53**	43.87	45.34	46.58
ScanQA + aux. task	3D scan	VSQ→AL	33.48	**66.10**	42.37	**69.53**	43.02	**46.40**	**47.20**
MCAN	BEV	VSQ→A	28.86	59.66	44.09	68.34	40.74	40.46	43.42
ClipBERT	Ego. video	VSQ→A	30.24	60.12	38.71	63.31	42.45	42.71	43.31
Unified QA$_{\text{Large}}$	ScanRefer	VSQ→A	33.01	50.43	31.91	56.51	**45.17**	41.11	41.00
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	39.77	38.71
GPT-3	ScanRefer	VSQ→A	**39.67**	45.99	40.47	45.56	36.08	38.42	41.00
GPT-3	ReferIt3D	VSQ→A	26.90	46.42	28.05	40.24	30.11	36.07	34.57
Human (amateur)	3D scan	VSQ→A	88.53	93.84	88.44	95.27	87.22	88.57	90.06

Representation of 3D scenes. 3D scan could still to be *better* representation of 3D scenes than egocentric videos and BEV pictures.

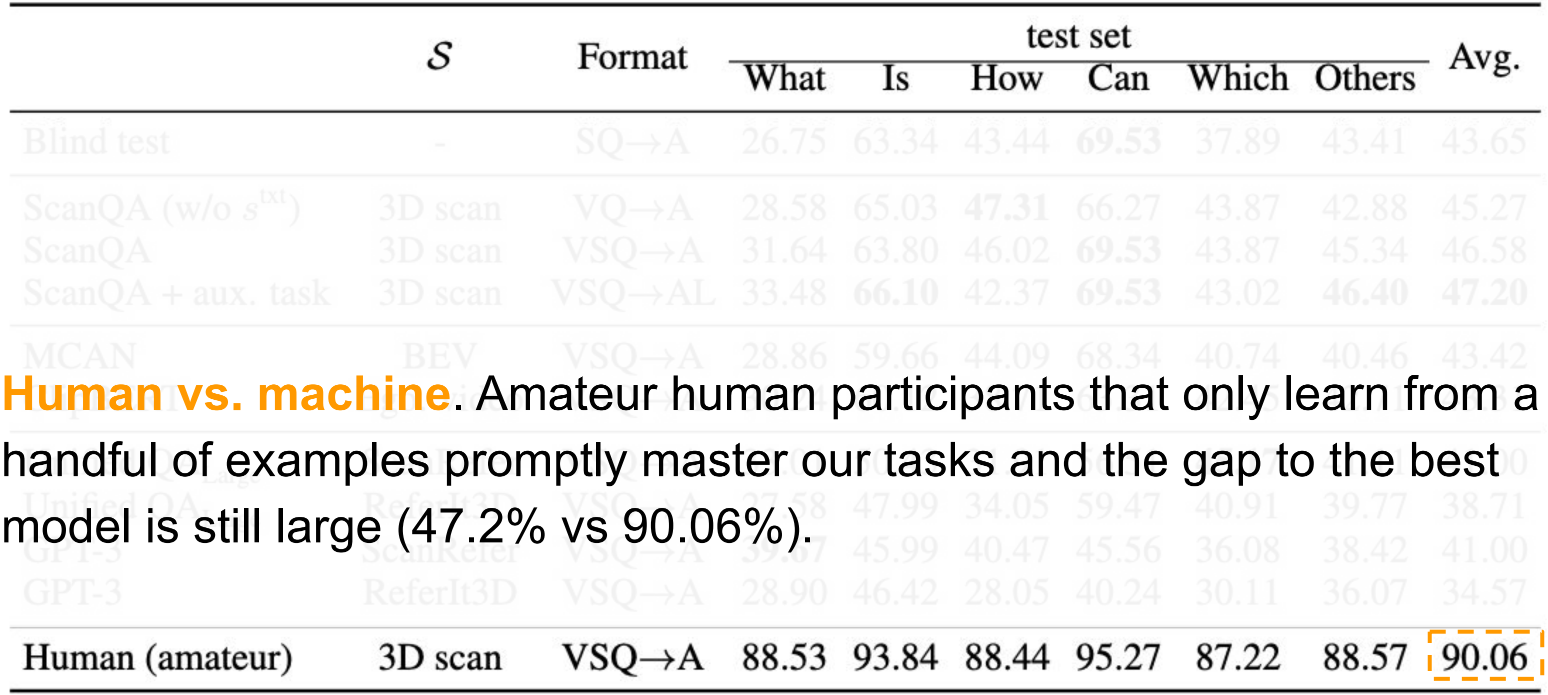
Benchmarking: quantitative results



	$\mathcal{S}$	Format	test set						Avg.
			What	Is	How	Can	Which	Others	
Blind test	-	SQ→A	26.75	63.34	43.44	**69.53**	37.89	43.41	43.65
ScanQA (w/o s^{txt})	3D scan	VQ→A	28.58	65.03	**47.31**	66.27	43.87	42.88	45.27
ScanQA	3D scan	VSQ→A	31.64	63.80	46.02	**69.53**	43.87	45.34	46.58
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	BEV	VSQ→A	28.86	59.66	44.09	68.34	40.74	40.46	43.42
ClipBERT	Ego. video	VSQ→A	30.24	60.12	38.71	63.31	42.45	42.71	43.31
Unified QA_{Large}	ScanRefer	VSQ→A	33.01	50.43	31.91	56.51	**45.17**	41.11	41.00
Unified QA_{Large}	ReferIt3D	VSQ→A	27.58	47.99	34.05	59.47	40.91	39.77	38.71
GPT-3	ScanRefer	VSQ→A	**39.67**	45.99	40.47	45.56	36.08	38.42	41.00
GPT-3	ReferIt3D	VSQ→A	28.90	46.42	28.05	40.24	30.11	36.07	34.57
Human (amateur)	3D scan	VSQ→A	88.53	93.84	88.44	95.27	87.22	88.57	90.06

Zero-shot models. These models indeed have great potential in common sense reasoning, spatial language understanding, etc. But they could be *bottlenecked* by 3D captions.

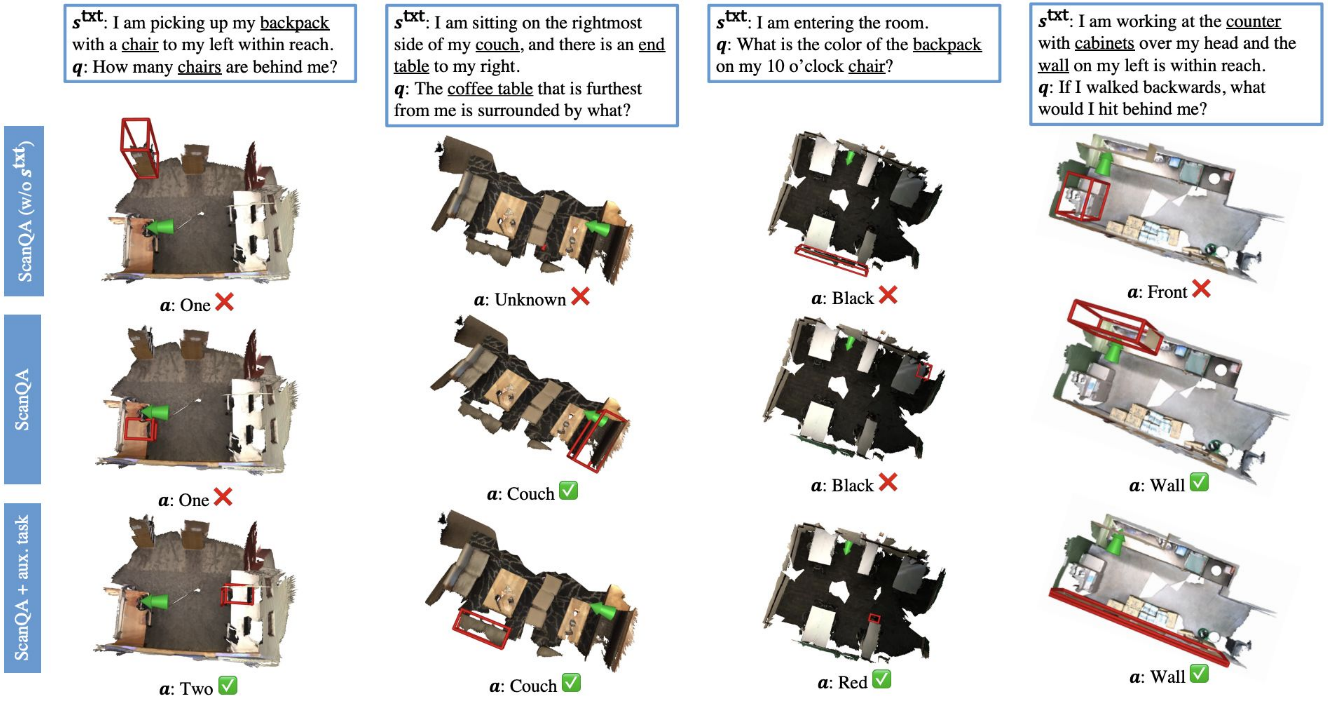
Benchmarking: quantitative results



	$\mathcal{S}$	Format	test set						Avg.
			What	Is	How	Can	Which	Others	
Blind test	-	SQ→A	26.75	63.34	43.44	**69.53**	37.89	43.41	43.65
ScanQA (w/o s^{txt})	3D scan	VQ→A	28.58	65.03	**47.31**	66.27	43.87	42.88	45.27
ScanQA	3D scan	VSQ→A	31.64	63.80	46.02	**69.53**	43.87	45.34	46.58
ScanQA + aux. task	3D scan	VSQ→AL	33.48	**66.10**	42.37	**69.53**	43.02	**46.40**	**47.20**
MCAN	BEV	VSQ→A	28.86	59.66	44.09	68.34	40.74	40.46	43.42
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Unified QA	ReferIt3D	VSQ→A	27.58	47.99	34.05	59.47	40.91	39.77	38.71
GPT-3	ScanRefer	VSQ→A	39.67	45.99	40.47	45.56	36.08	38.42	41.00
GPT-3	ReferIt3D	VSQ→A	28.90	46.42	28.05	40.24	30.11	36.07	34.57
Human (amateur)	3D scan	VSQ→A	88.53	93.84	88.44	95.27	87.22	88.57	90.06

Human vs. machine. Amateur human participants that only learn from a handful of examples promptly master our tasks and the gap to the best model is still large (47.2% vs 90.06%).

Benchmarking: qualitative results & failure modes



s^{txt}: I am picking up my backpack with a chair to my left within reach.
$\boldsymbol{q}$: How many chairs are behind me?

s^{txt}: I am sitting on the rightmost side of my couch, and there is an end table to my right.
$\boldsymbol{q}$: The coffee table that is furthest from me is surrounded by what?

s^{txt}: I am entering the room.
$\boldsymbol{q}$: What is the color of the backpack on my 10 o'clock chair?

s^{txt}: I am working at the counter with cabinets over my head and the wall on my left is within reach.
$\boldsymbol{q}$: If I walked backwards, what would I hit behind me?

ScanQA (w/o s^{txt})
ScanQA
ScanQA + aux. task
$\boldsymbol{a}$: One ✗
$\boldsymbol{a}$: Unknown ✗
$\boldsymbol{a}$: Black ✗
$\boldsymbol{a}$: Front ✗
$\boldsymbol{a}$: One ✗
$\boldsymbol{a}$: Couch ✅
$\boldsymbol{a}$: Black ✗
$\boldsymbol{a}$: Wall ✅
$\boldsymbol{a}$: Two ✅
$\boldsymbol{a}$: Couch ✅
$\boldsymbol{a}$: Red ✅
$\boldsymbol{a}$: Wall ✅

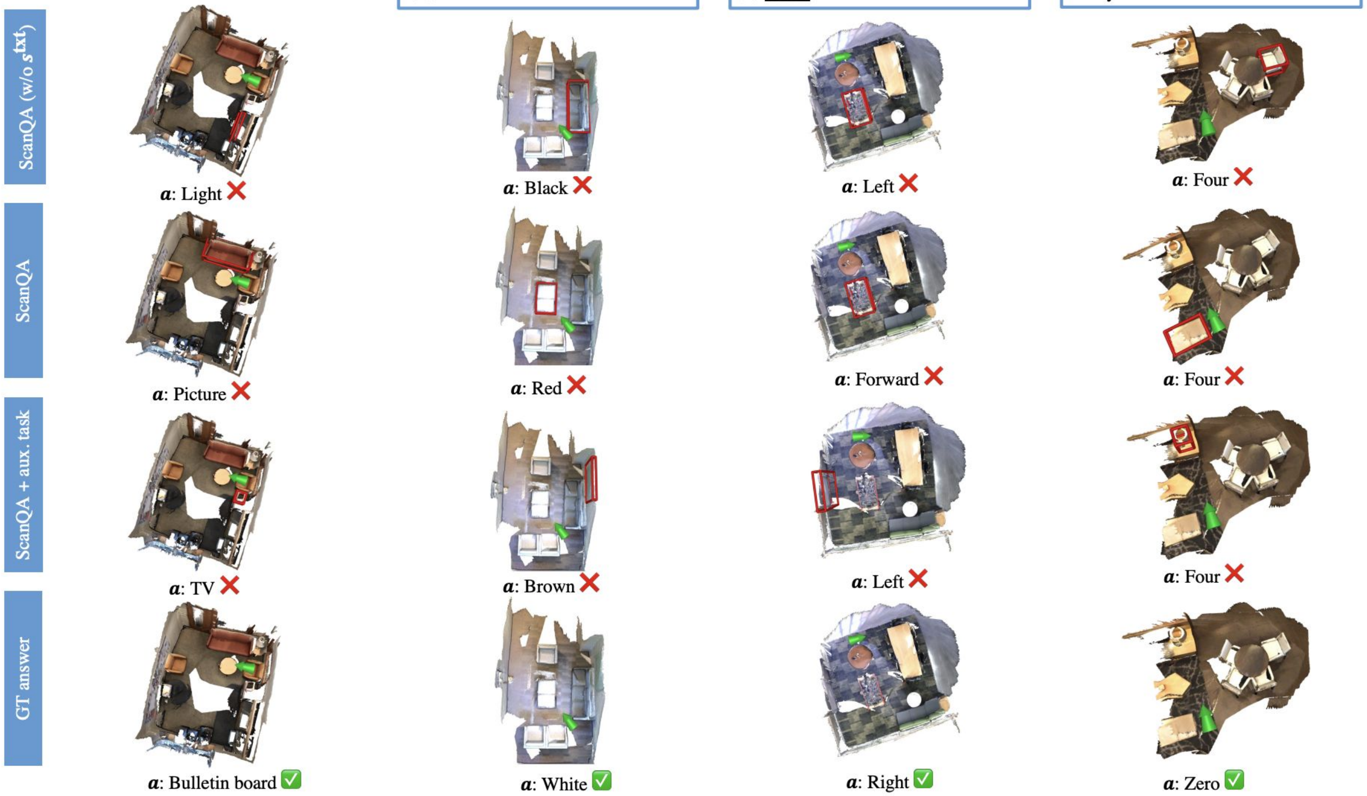
Most-attended bbox is highlighted in red. Our best model (ScanQA + aux. task) are more likely to attend to the **relevant** objects and provide the **correct** answer.

s^{txt}: I am sitting on the armchair in front of the window.
q: What is above the armchair that is far away in front of me?

s^{txt}: I am facing an ottoman with a couch to my right within reach and an armchair to my left.
q: What color is the armchair to my left?

s^{txt}: I am facing the table and there is a coffee table and a foosball table to my left.
q: Which way should I go to sit on the couch?

s^{txt}: I am facing an end table and there is a couch on my left within reach.
q: How many chairs does the table on my left have?



When the model **fails** to attend to the relevant objects, there is a good chance it will also provide the **wrong** answer.

Takeaway

We present SQA3D, a new benchmark for **embodied scene understanding**, aiming at bridging the gap between 3D scene understanding and embodied AI.

SQA3D is the **largest** dataset combines the best of both worlds: situated reasoning, human-written text, and diverse & challenging problems.

State-of-the-art multi-modal QA models and zero-shot large models struggle on SQA3D and the gap to amateur human participants is also considerable.

Code & benchmark will be online soon!

SQA3D: Situated Question Answering in 3D Scenes