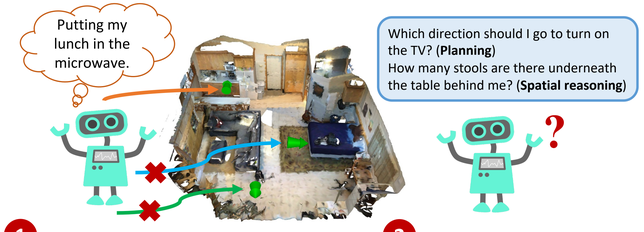
SQA3D: Situated Question Answering in 3D Scenes

Xiaojian Ma*, Silong Yong*, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, Siyuan Huang

*equal contribution

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

I Embodied AI + Scene Understanding = Embodied Scene Understanding

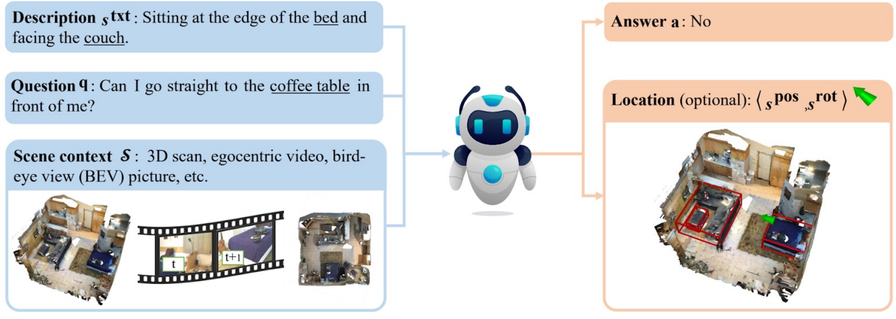


We propose to bridge **embodied AI** and **3D scene understanding** with a new quest: **embodied scene understanding**, which includes two tasks:

1. Situated understanding: agents understand the surroundings (situations) from a *dynamic*, *ego-centric* view.

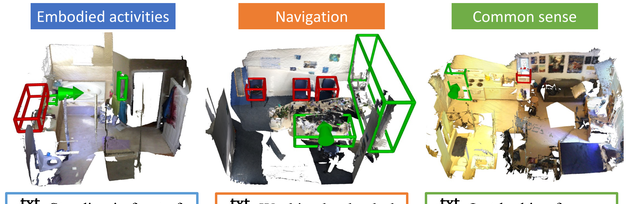
2. Situated reasoning: agents accomplish *reasoning & planning* according to current situation.

II What is SQA3D?



- Given a **3D scene context** $\mathcal{S}$, **situation description** s^{txt}, the agents needs to first infer the corresponding position in the 3D scene (*situation understanding*), then answer a question q (*situated reasoning*).
- The **3D scene context** $\mathcal{S}$ can be 3D scans, egocentric videos, or bird-eye view (BEV) pictures.
- The position is represented as: **position** s^{pos} (xyz) and **orientation** s^{rot} (quaternion). Predicting them is *optional*.

III SQA3D examples



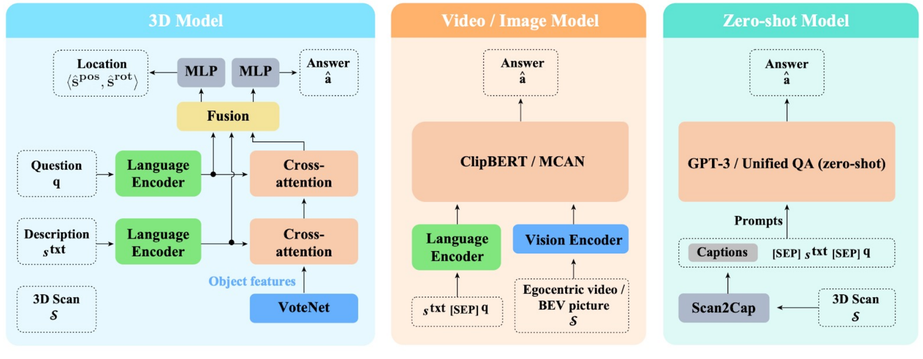
s^{txt}: Standing in front of the sink and facing the towels.
q: Can I see myself in the mirror?
a: No

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

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

*The categories here do not mean to be exhaustive.

IV Possible models for SQA3D?



- **3D** multimodal models: cross-attention transformer, 3D scans are processed by *VoteNet*.
- **Video/image** multimodal models: State-of-the-art video transformer (*ClipBERT*) and image transformer (*MCAN*).
- **Zero-shot** models: Large language models (LLMs), 3D scene is converted into text using scene captioning.

V Benchmarking results

	$\mathcal{S}$	Format	test set What	Is	How	Can	Which	Others	Avg.
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_{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

What can we learn from the results on SQA3D so far?

- **Situation understanding**. Models with better situation understanding (w/ s^{txt}, w/ aux. task) generally deliver better results (*aux. task means predicting the position from s^{txt}).
- **Representation of 3D scenes**. 3D scan could still to be *better* representation of 3D scenes than egocentric videos and BEV pictures.
- **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.
- **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%).

Code

Project page

Paper