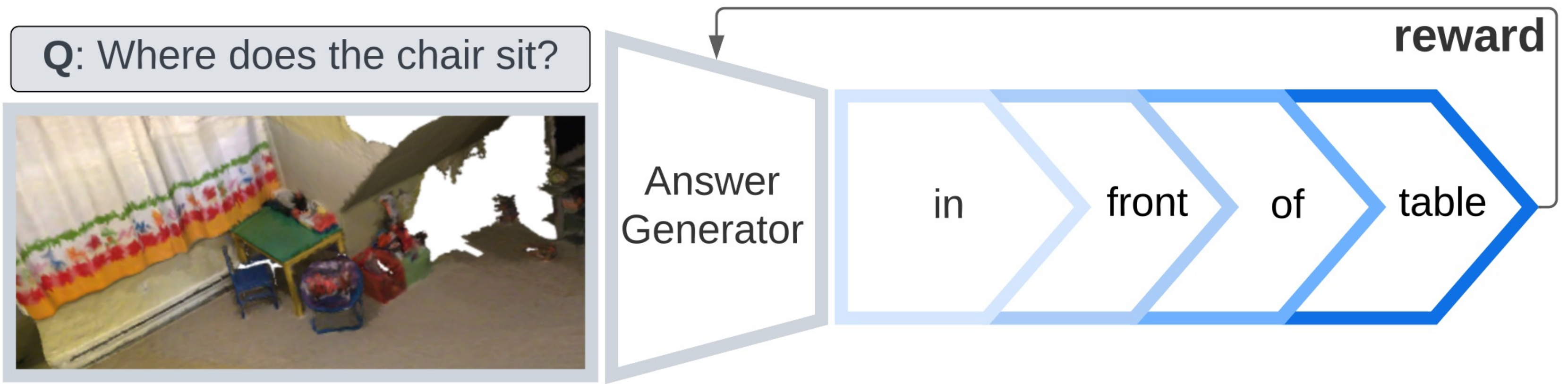
Generating Context-Aware Natural Answers for Questions in 3D Scenes

Mohammed Munzer Dwedari Zhenyu Chen Matthias Niessner
Technical University of Munich

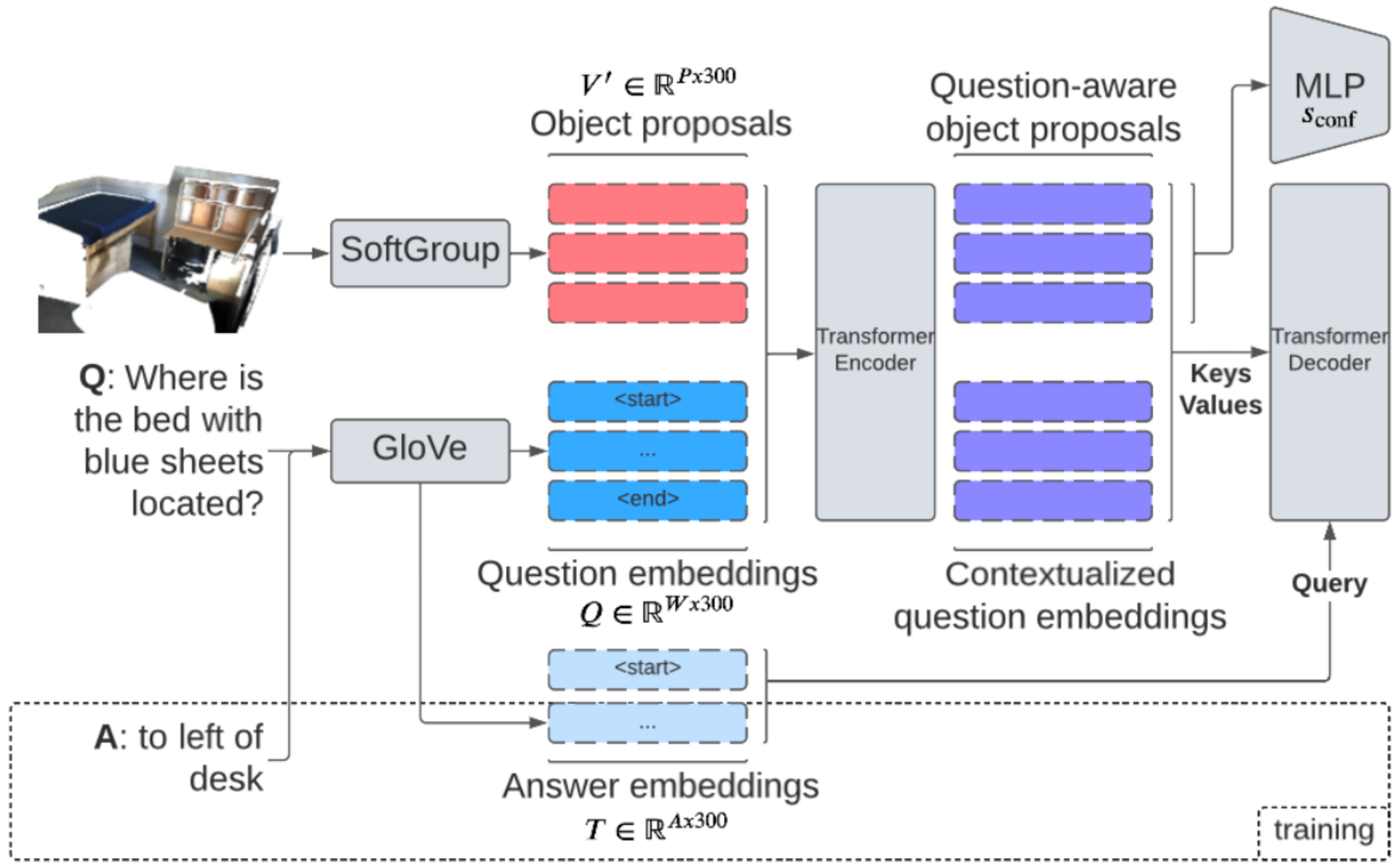
BMVC 2023

Introduction



In this work, we tackle the task of question answering in 3D indoor environments. Previous methods are restricted to a pre-defined answer space. We propose **Gen3DQA**, an end-to-end transformer-based architecture to generate, rather than predict, natural answers for questions in 3D scenes. Our method directly optimizes the global semantics of the generated sentences via the language rewards.

Method



After encoding the input scene and question into object proposals and questions embeddings, they are combined into one sequence and fed into a transformer encoder. The contextualized sequence is then fed into a transformer decoder to generate the answer.

Check out our code and pretrained models!

github.com/MunzerDw/Gen3DQA

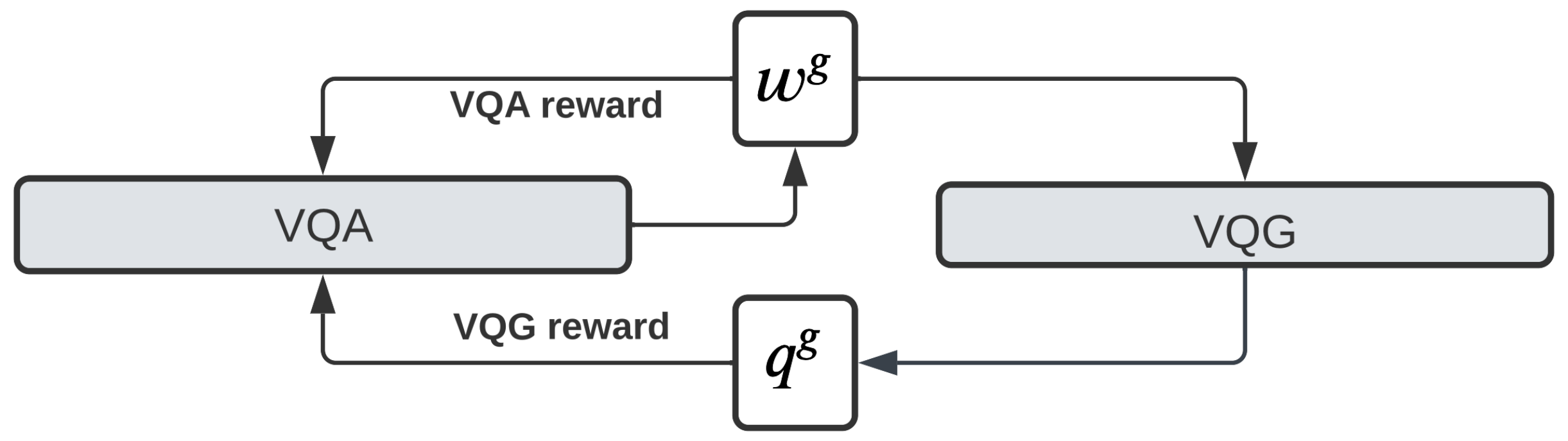
Training

1) Object localization cross entropy loss + word-level cross entropy loss

2) Object localization cross entropy loss + reinforcement learning:

$$L_{\text{cider}}(\theta) = -\mathbb{E}_{w^g \sim p_\theta}[r(w^g)]$$

$$\nabla_\theta L_{\text{cider}}(\theta) = -((r^g_{\text{VQA}} - r^b_{\text{VQA}}) + (r^g_{\text{VQG}} - r^b_{\text{VQG}}))\nabla_\theta log p_\theta(w^g)$$



Model	BLEU-1	BLEU-4	ROUGE	METEOR	CIDEr
Valid					
Gen3DQA (w/o VQG reward)	39.12	**13.2**	35.48	14.89	71.39
Gen3DQA (w/ VQG reward)	**39.53**	12.7	**35.97**	**15.11**	**71.97**
Test w/ object IDs					
Gen3DQA (w/o VQG reward)	38.89	**12.67**	35.35	14.82	71.09
Gen3DQA (w/ VQG reward)	**39.30**	12.24	**35.78**	**14.99**	**72.22**
Test w/o object IDs					
Gen3DQA (w/o VQG reward)	37.61	**12.00**	32.57	14.09	65.58
Gen3DQA (w/ VQG reward)	**38.07**	11.61	**33.03**	**14.28**	**66.57**

Performance of our method with and without VQG reward.

Quantitative Results

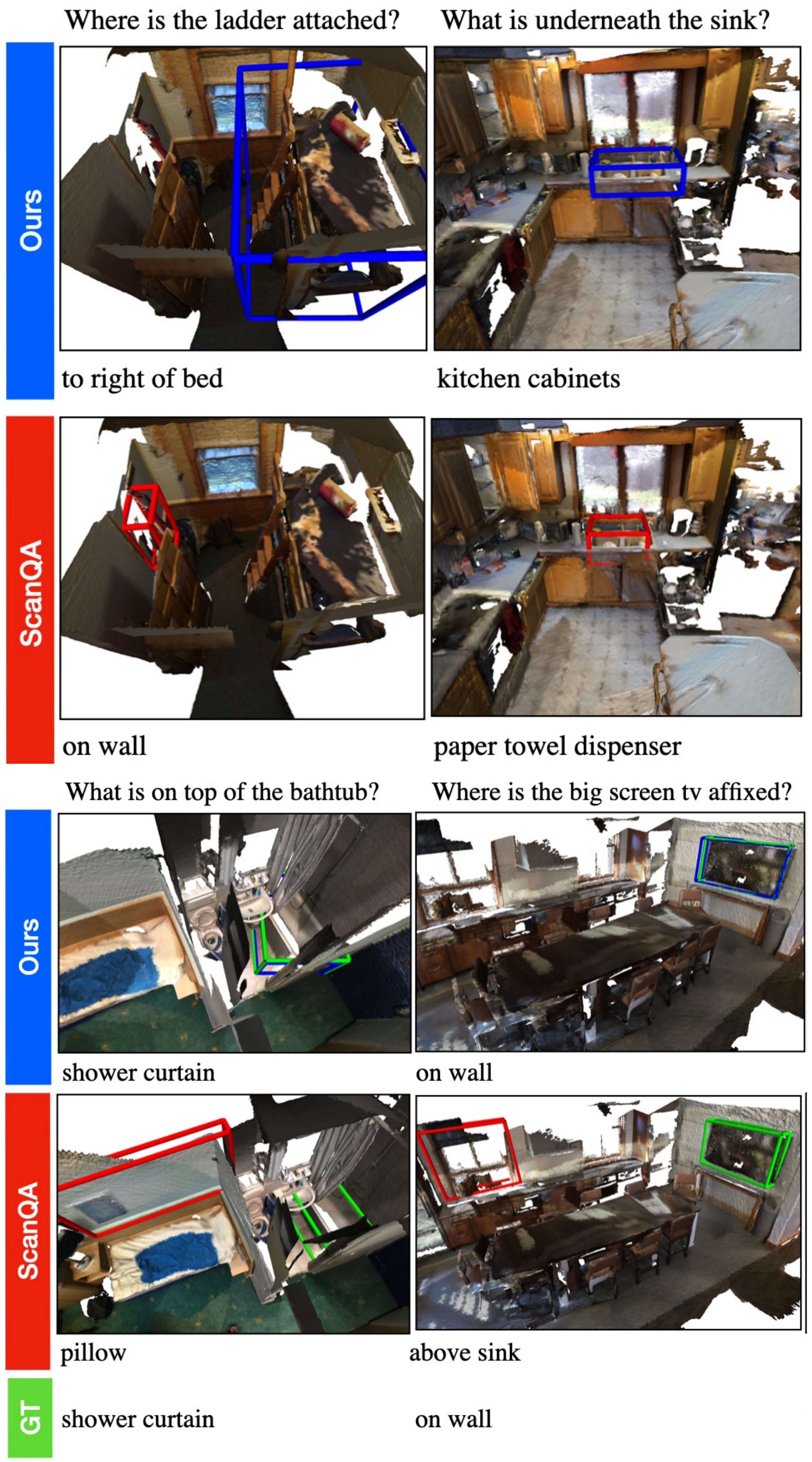
Model	BLEU-1	BLEU-4	ROUGE	METEOR	CIDEr
Test w/ object IDs					
ScanQA [4]	31.56	12.04	34.34	13.55	67.29
CLIP-guided [32]	32.72	**14.64**	35.15	13.94	69.53
Gen3DQA (XE loss)	35.24	10.79	33.50	13.61	64.83
Gen3DQA	**39.30**	12.24	**35.78**	**14.99**	**72.22**
Test w/o object IDs					
ScanQA [4]	30.68	10.75	31.09	12.59	60.24
CLIP-guided [32]	32.70	**11.73**	32.41	13.28	62.83
Gen3DQA (XE loss)	35.08	10.62	30.99	12.87	60.05
Gen3DQA	**38.07**	11.61	**33.03**	**14.28**	**66.57**

ScanQA benchmark scores of previous methods and ours. Our method performs better, especially on the more challenging CIDEr score.

	ScanQA [4]	CLIP-guided [32]	Gen3DQA
Acc@0.5	15.42	21.22	**23.79**

Object localization accuracy Acc@0.5 of our method and previous ones.

Qualitative Results


Where is the ladder attached?
What is underneath the sink?
Ours
to right of bed
kitchen cabinets
ScanQA
on wall
paper towel dispenser
What is on top of the bathtub?
Where is the big screen tv affixed?
Ours
shower curtain
on wall
ScanQA
pillow
above sink
GT
shower curtain
on wall