---

title: "Bug Journal 2025-08-25" date: 2025-08-25T22:37:11-04:00 keywords:

- Bug Journal summary: SOTA VLA draft: false

总结

目前在 Benchmark 1: SQA3D上 GPT4Scene: Understand 3D Scenes from Videos with Vision-Language Models 效果最好。达到了 Exact Maching 62.4% 的准确率。 在

Benchmark 2: ScanQA上 Bridging the Gap between 2D and 3D Visual Question Answering: A Fusion Approach for 3D VQA (AAAI 2024) 效果最好。达到了 Exact Maching 31.29% 的准确率。

动机

想看看VLA / 空间推理VLM 发展得怎么样，SOTA是什么

调研方式

对于这个Task,我打算从CVPR 2025入手，看看最新的VLA都是如何实现的，又是如何比较的

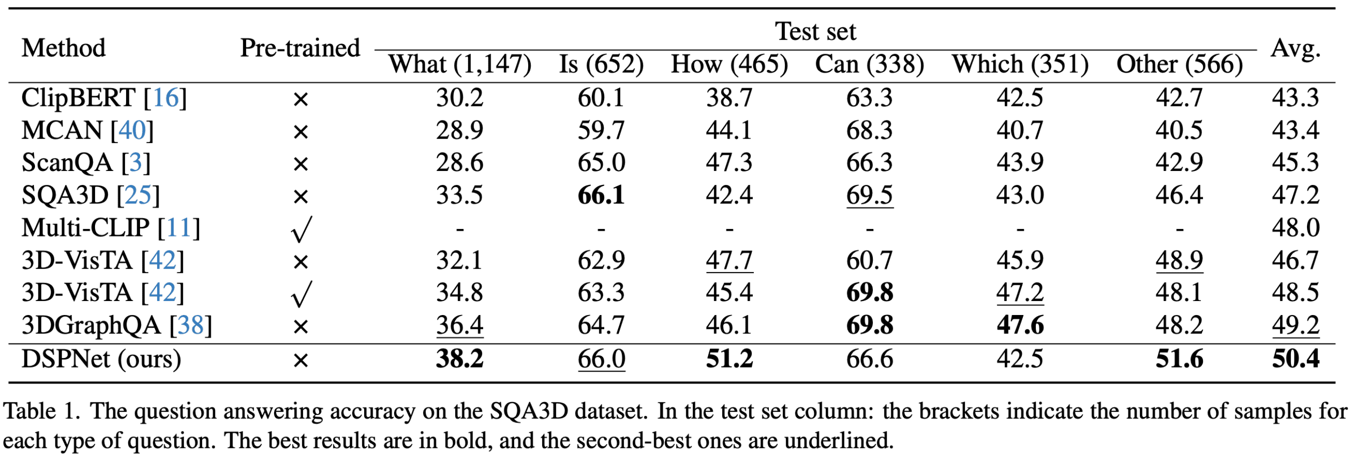
Plan 1: 首先，我会先寻找一下CVPR 2025中和VLA有关的任务，并且看看他们的表现

Paper 1:

DSPNet: Dual-vision Scene Perception for Robust 3D Question Answering 链接 From 中大 HCP

结果：

结果一:

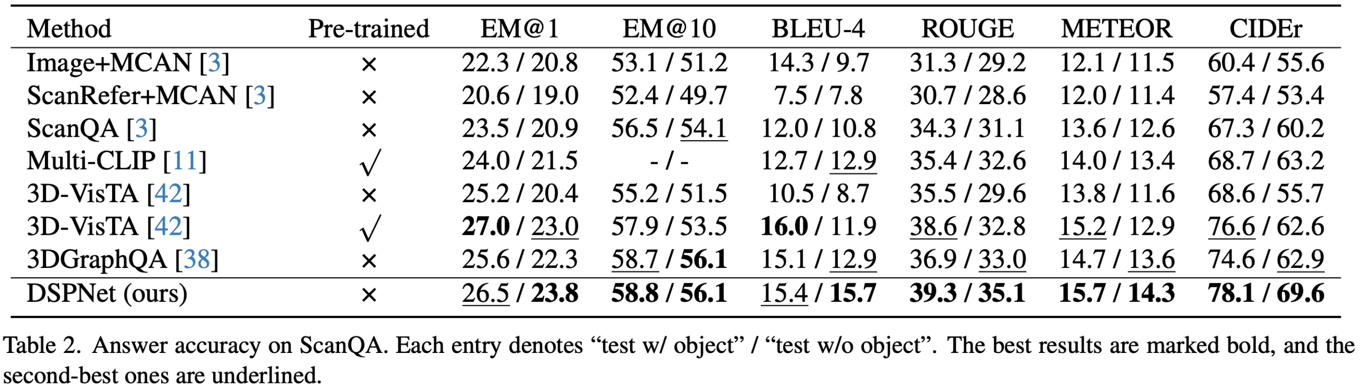


Method	Pre-trained	Test set						Avg.
		What (1,147)	Is (652)	How (465)	Can (338)	Which (351)	Other (566)	
ClipBERT [16]	×	30.2	60.1	38.7	63.3	42.5	42.7	43.3
MCAN [40]	×	28.9	59.7	44.1	68.3	40.7	40.5	43.4
ScanQA [3]	×	28.6	65.0	47.3	66.3	43.9	42.9	45.3
SQA3D [25]	×	33.5	**66.1**	42.4	69.5	43.0	46.4	47.2
Multi-CLIP [11]	√	-	-	-	-	-	-	48.0
3D-VisTA [42]	×	32.1	62.9	47.7	60.7	45.9	48.9	46.7
3D-VisTA [42]	√	34.8	63.3	45.4	**69.8**	47.2	48.1	48.5
3DGraphQA [38]	×	36.4	64.7	46.1	**69.8**	**47.6**	48.2	49.2
DSPNet (ours)	×	**38.2**	66.0	**51.2**	66.6	42.5	**51.6**	**50.4**

Table 1. The question answering accuracy on the SQA3D dataset. In the test set column: the brackets indicate the number of samples for each type of question. The best results are in bold, and the second-best ones are underlined.

这张图用的是 这个benchmark (SQA3D)

结果二：



Method	Pre-trained	EM@1	EM@10	BLEU-4	ROUGE	METEOR	CIDEr
Image+MCAN [3]	×	22.3 / 20.8	53.1 / 51.2	14.3 / 9.7	31.3 / 29.2	12.1 / 11.5	60.4 / 55.6
ScanRefer+MCAN [3]	×	20.6 / 19.0	52.4 / 49.7	7.5 / 7.8	30.7 / 28.6	12.0 / 11.4	57.4 / 53.4
ScanQA [3]	×	23.5 / 20.9	56.5 / 54.1	12.0 / 10.8	34.3 / 31.1	13.6 / 12.6	67.3 / 60.2
Multi-CLIP [11]	√	24.0 / 21.5	- / -	12.7 / 12.9	35.4 / 32.6	14.0 / 13.4	68.7 / 63.2
3D-VisTA [42]	×	25.2 / 20.4	55.2 / 51.5	10.5 / 8.7	35.5 / 29.6	13.8 / 11.6	68.6 / 55.7
3D-VisTA [42]	√	**27.0** / 23.0	57.9 / 53.5	**16.0** / 11.9	38.6 / 32.8	15.2 / 12.9	76.6 / 62.6
3DGraphQA [38]	×	25.6 / 22.3	58.7 / **56.1**	15.1 / 12.9	36.9 / 33.0	14.7 / 13.6	74.6 / 62.9
DSPNet (ours)	×	26.5 / **23.8**	**58.8 / 56.1**	15.4 / **15.7**	**39.3 / 35.1**	**15.7 / 14.3**	**78.1 / 69.6**

Table 2. Answer accuracy on ScanQA. Each entry denotes "test w/ object" / "test w/o object". The best results are marked bold, and the second-best ones are underlined.

这张图用的是 这个benchmark (ScanQA)

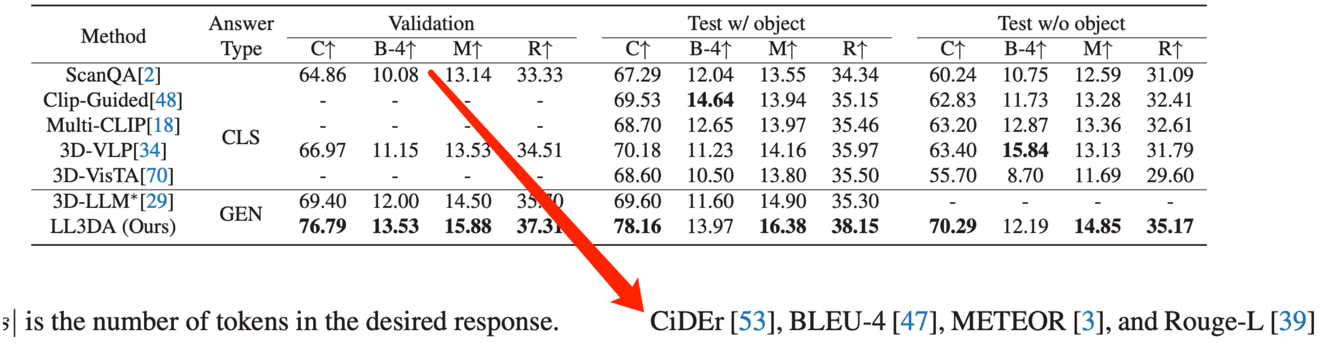
Paper 2:

LL3DA: Visual Interactive Instruction Tuning for Omni-3D Understanding, Reasoning, and Planning (CVPR 2024)

链接

From Fudan, Tencent, and National University of Singapore

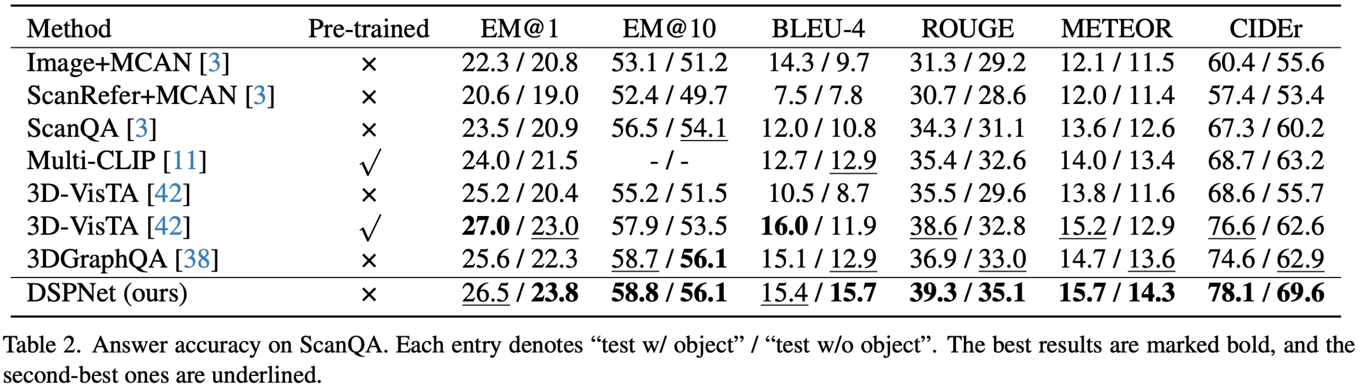
结果



Method	Answer Type	Validation				Test w/ object				Test w/o object			
		C↑	B-4↑	M↑	R↑	C↑	B-4↑	M↑	R↑	C↑	B-4↑	M↑	R↑
ScanQA[2]	CLS	64.86	10.08	13.14	33.33	67.29	12.04	13.55	34.34	60.24	10.75	12.59	31.09
Clip-Guided[48]		-	-	-	-	69.53	**14.64**	13.94	35.15	62.83	11.73	13.28	32.41
Multi-CLIP[18]		-	-	-	-	68.70	12.65	13.97	35.46	63.20	12.87	13.36	32.61
3D-VLP[34]		66.97	11.15	13.53	34.51	70.18	11.23	14.16	35.97	63.40	**15.84**	13.13	31.79
3D-VisTA[70]		-	-	-	-	68.60	10.50	13.80	35.50	55.70	8.70	11.69	29.60
3D-LLM*[29]	GEN	69.40	12.00	14.50	35.70	69.60	11.60	14.90	35.30	-	-	-	-
LL3DA (Ours)		**76.79**	**13.53**	**15.88**	**37.31**	**78.16**	13.97	**16.38**	**38.15**	**70.29**	12.19	**14.85**	**35.17**

s| is the number of tokens in the desired response. CiDEr [53], BLEU-4 [47], METEOR [3], and Rouge-L [39]

这张图用的是 这个benchmark (ScanQA)



Method	Pre-trained	EM@1	EM@10	BLEU-4	ROUGE	METEOR	CIDEr
Image+MCAN [3]	×	22.3 / 20.8	53.1 / 51.2	14.3 / 9.7	31.3 / 29.2	12.1 / 11.5	60.4 / 55.6
ScanRefer+MCAN [3]	×	20.6 / 19.0	52.4 / 49.7	7.5 / 7.8	30.7 / 28.6	12.0 / 11.4	57.4 / 53.4
ScanQA [3]	×	23.5 / 20.9	56.5 / 54.1	12.0 / 10.8	34.3 / 31.1	13.6 / 12.6	67.3 / 60.2
Multi-CLIP [11]	√	24.0 / 21.5	- / -	12.7 / 12.9	35.4 / 32.6	14.0 / 13.4	68.7 / 63.2
3D-VisTA [42]	×	25.2 / 20.4	55.2 / 51.5	10.5 / 8.7	35.5 / 29.6	13.8 / 11.6	68.6 / 55.7
3D-VisTA [42]	√	**27.0** / 23.0	57.9 / 53.5	**16.0** / 11.9	38.6 / 32.8	15.2 / 12.9	76.6 / 62.6
3DGraphQA [38]	×	25.6 / 22.3	58.7 / **56.1**	15.1 / 12.9	36.9 / 33.0	14.7 / 13.6	74.6 / 62.9
DSPNet (ours)	×	26.5 / **23.8**	**58.8 / 56.1**	15.4 / **15.7**	**39.3 / 35.1**	**15.7 / 14.3**	**78.1 / 69.6**

Table 2. Answer accuracy on ScanQA. Each entry denotes "test w/ object" / "test w/o object". The best results are marked bold, and the second-best ones are underlined.

如果横向对比这两个模型，其实有些数据还是这个模型高一些

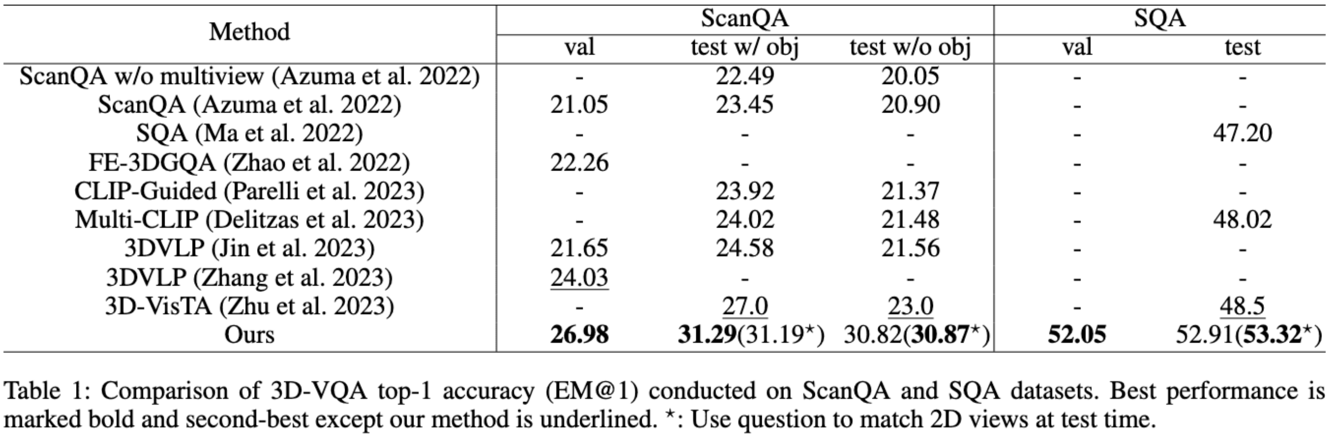
Paper 3:

Bridging the Gap between 2D and 3D Visual Question Answering: A Fusion Approach for 3D VQA (AAAI 2024)

链接

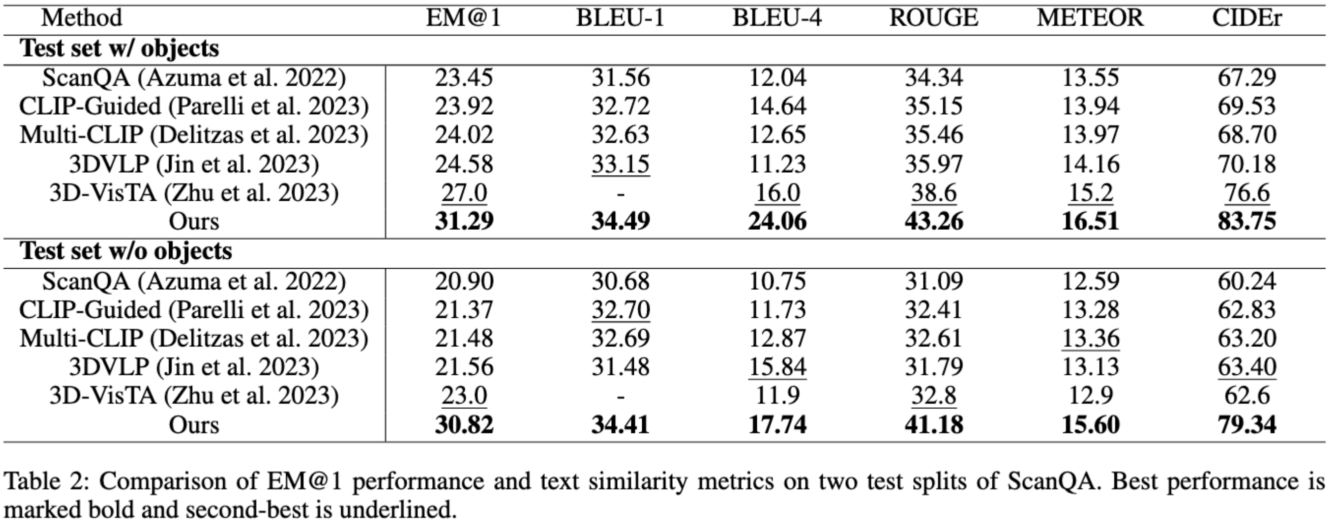
From PKU

结果



Method	ScanQA			SQA	
	val	test w/ obj	test w/o obj	val	test
ScanQA w/o multiview (Azuma et al. 2022)	-	22.49	20.05	-	-
ScanQA (Azuma et al. 2022)	21.05	23.45	20.90	-	-
SQA (Ma et al. 2022)	-	-	-	-	47.20
FE-3DGQA (Zhao et al. 2022)	22.26	-	-	-	-
CLIP-Guided (Parelli et al. 2023)	-	23.92	21.37	-	-
Multi-CLIP (Delitzas et al. 2023)	-	24.02	21.48	-	48.02
3DVLP (Jin et al. 2023)	21.65	24.58	21.56	-	-
3DVLP (Zhang et al. 2023)	24.03	-	-	-	-
3D-VisTA (Zhu et al. 2023)	-	27.0	23.0	-	48.5
Ours	**26.98**	**31.29**(31.19*)	30.82(**30.87***)	**52.05**	52.91(**53.32***)

Table 1: Comparison of 3D-VQA top-1 accuracy (EM@1) conducted on ScanQA and SQA datasets. Best performance is marked bold and second-best except our method is underlined. *: Use question to match 2D views at test time.

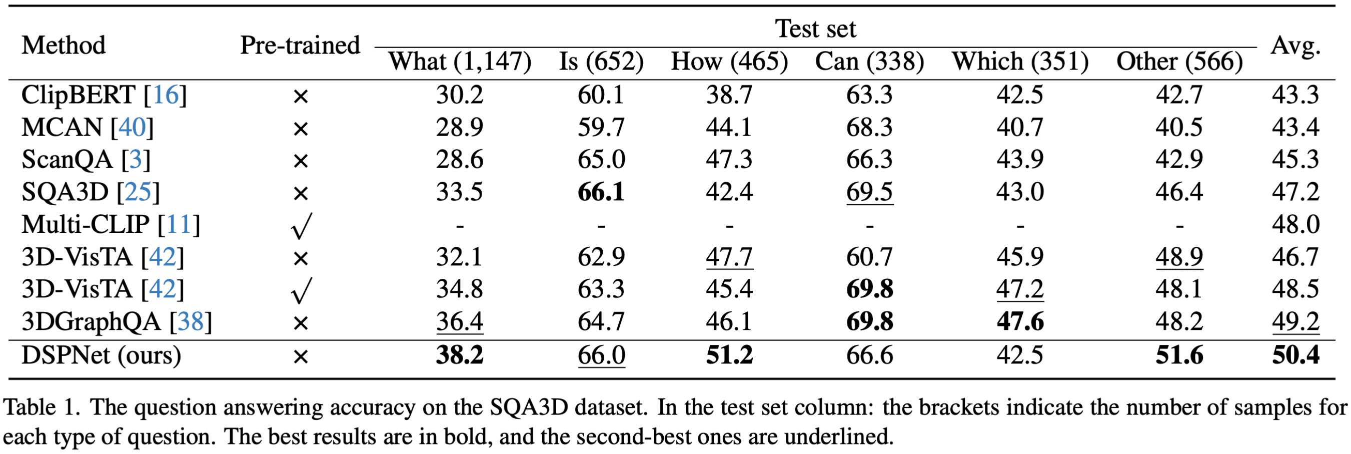


Method	EM@1	BLEU-1	BLEU-4	ROUGE	METEOR	CIDEr
Test set w/ objects						
ScanQA (Azuma et al. 2022)	23.45	31.56	12.04	34.34	13.55	67.29
CLIP-Guided (Parelli et al. 2023)	23.92	32.72	14.64	35.15	13.94	69.53
Multi-CLIP (Delitzas et al. 2023)	24.02	32.63	12.65	35.46	13.97	68.70
3DVLP (Jin et al. 2023)	24.58	33.15	11.23	35.97	14.16	70.18
3D-VisTA (Zhu et al. 2023)	27.0	-	16.0	38.6	15.2	76.6
Ours	**31.29**	**34.49**	**24.06**	**43.26**	**16.51**	**83.75**
Test set w/o objects						
ScanQA (Azuma et al. 2022)	20.90	30.68	10.75	31.09	12.59	60.24
CLIP-Guided (Parelli et al. 2023)	21.37	32.70	11.73	32.41	13.28	62.83
Multi-CLIP (Delitzas et al. 2023)	21.48	32.69	12.87	32.61	13.36	63.20
3DVLP (Jin et al. 2023)	21.56	31.48	15.84	31.79	13.13	63.40
3D-VisTA (Zhu et al. 2023)	23.0	-	11.9	32.8	12.9	62.6
Ours	**30.82**	**34.41**	**17.74**	**41.18**	**15.60**	**79.34**

Table 2: Comparison of EM@1 performance and text similarity metrics on two test splits of ScanQA. Best performance is marked bold and second-best is underlined.

这张图用的是 这个benchmark (ScanQA) 和 这个benchmark (SQA3D)

如果我们比较这一篇和第一篇CVPR2025



Method	Pre-trained	Test set						Avg.
		What (1,147)	Is (652)	How (465)	Can (338)	Which (351)	Other (566)	
ClipBERT [16]	×	30.2	60.1	38.7	63.3	42.5	42.7	43.3
MCAN [40]	×	28.9	59.7	44.1	68.3	40.7	40.5	43.4
ScanQA [3]	×	28.6	65.0	47.3	66.3	43.9	42.9	45.3
SQA3D [25]	×	33.5	**66.1**	42.4	69.5	43.0	46.4	47.2
Multi-CLIP [11]	√	-	-	-	-	-	-	48.0
3D-VisTA [42]	×	32.1	62.9	47.7	60.7	45.9	48.9	46.7
3D-VisTA [42]	√	34.8	63.3	45.4	**69.8**	47.2	48.1	48.5
3DGraphQA [38]	×	36.4	64.7	46.1	**69.8**	**47.6**	48.2	49.2
DSPNet (ours)	×	**38.2**	66.0	**51.2**	66.6	42.5	**51.6**	**50.4**

Table 1. The question answering accuracy on the SQA3D dataset. In the test set column: the brackets indicate the number of samples for each type of question. The best results are in bold, and the second-best ones are underlined.

我们会发现，实际上这篇效果更好

Paper 4:

SceneVerse: Scaling 3D Vision-Language Learning for Grounded Scene Understanding (ECCV 2024)

链接

From BIGAI, Beijing

结果

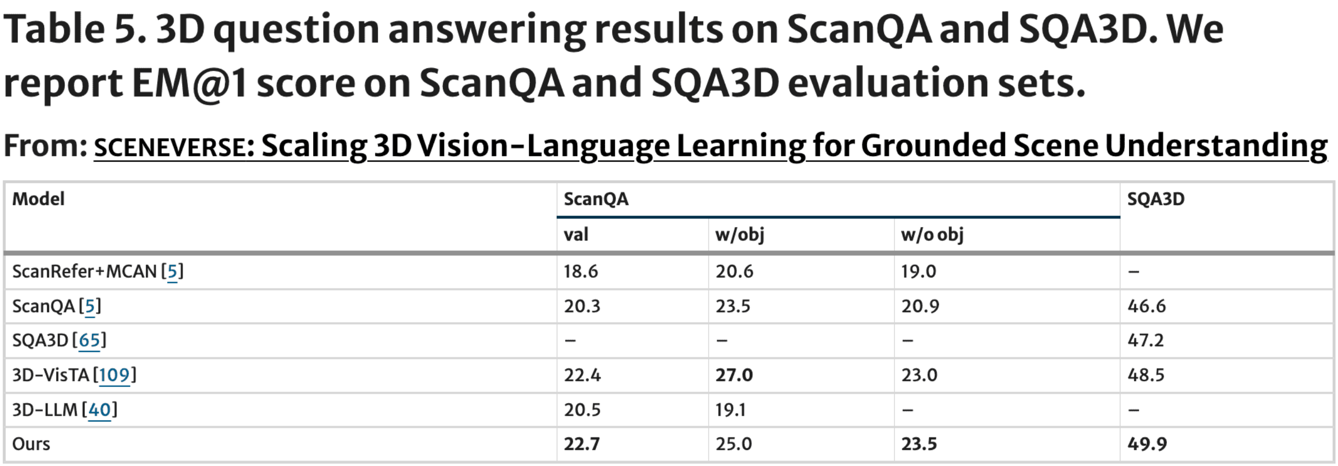


Table 5. 3D question answering results on ScanQA and SQA3D. We report EM@1 score on ScanQA and SQA3D evaluation sets.

From: SCENEVERSE: Scaling 3D Vision-Language Learning for Grounded Scene Understanding

Model	ScanQA			SQA3D
	val	w/obj	w/o obj	
ScanRefer+MCAN [5]	18.6	20.6	19.0	–
ScanQA [5]	20.3	23.5	20.9	46.6
SQA3D [65]	–	–	–	47.2
3D-VisTA [109]	22.4	**27.0**	23.0	48.5
3D-LLM [40]	20.5	19.1	–	–
Ours	**22.7**	25.0	**23.5**	**49.9**

可以看到效果没那么理想

Paper 5:

Scene-LLM: Extending Language Model for 3D Visual Reasoning (WACV 2025) 链接

From Brown University & Meta

结果

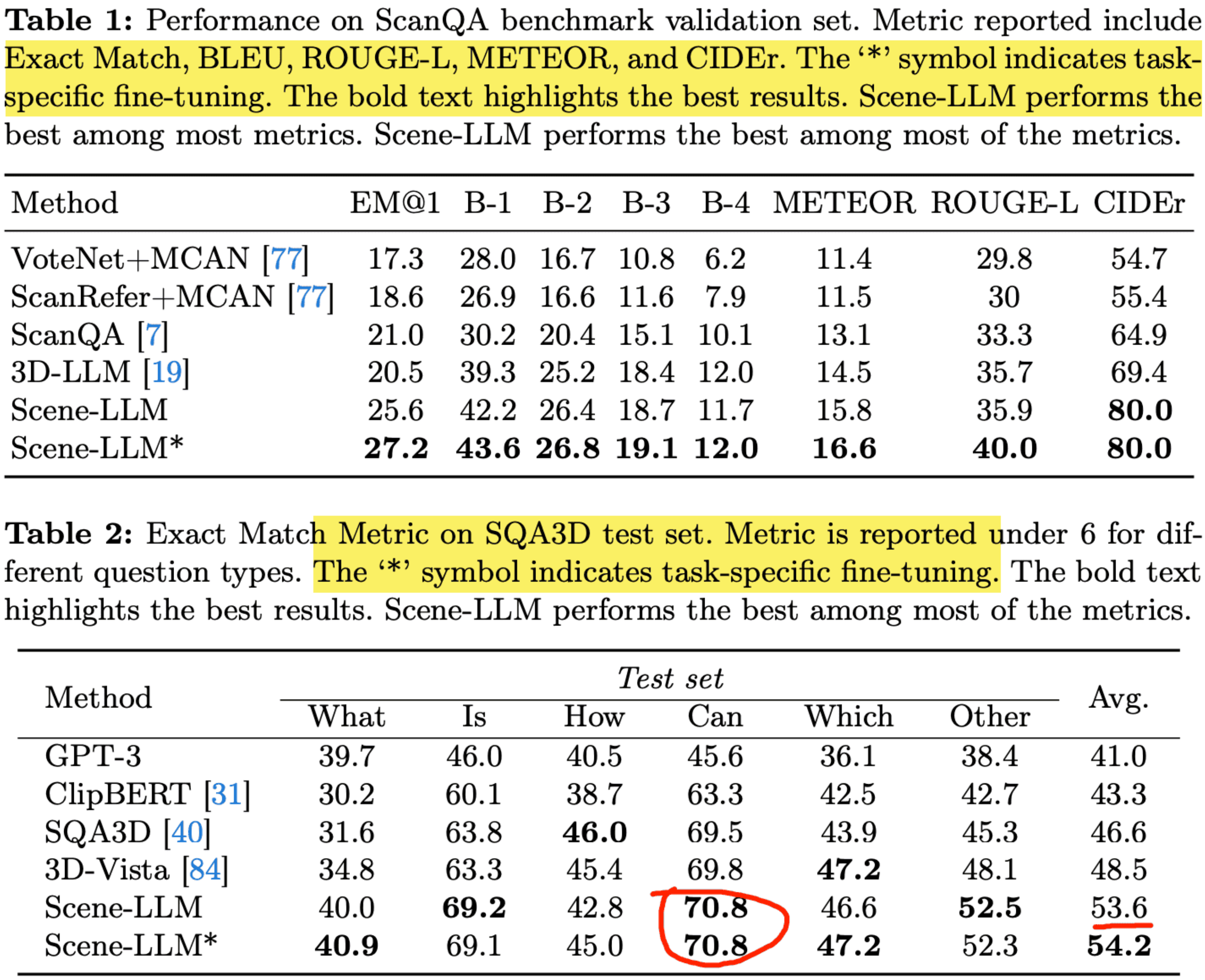


Table 1: Performance on ScanQA benchmark validation set. Metric reported include Exact Match, BLEU, ROUGE-L, METEOR, and CIDEr. The '*' symbol indicates task-specific fine-tuning. The bold text highlights the best results. Scene-LLM performs the best among most metrics. Scene-LLM performs the best among most of the metrics.

Method	EM@1	B-1	B-2	B-3	B-4	METEOR	ROUGE-L	CIDEr
VoteNet+MCAN [77]	17.3	28.0	16.7	10.8	6.2	11.4	29.8	54.7
ScanRefer+MCAN [77]	18.6	26.9	16.6	11.6	7.9	11.5	30	55.4
ScanQA [7]	21.0	30.2	20.4	15.1	10.1	13.1	33.3	64.9
3D-LLM [19]	20.5	39.3	25.2	18.4	12.0	14.5	35.7	69.4
Scene-LLM	25.6	42.2	26.4	18.7	11.7	15.8	35.9	**80.0**
Scene-LLM*	**27.2**	**43.6**	**26.8**	**19.1**	**12.0**	**16.6**	**40.0**	**80.0**

Table 2: Exact Match Metric on SQA3D test set. Metric is reported under 6 for different question types. The '*' symbol indicates task-specific fine-tuning. The bold text highlights the best results. Scene-LLM performs the best among most of the metrics.

Method	*Test set*						Avg.
	What	Is	How	Can	Which	Other	
GPT-3	39.7	46.0	40.5	45.6	36.1	38.4	41.0
ClipBERT [31]	30.2	60.1	38.7	63.3	42.5	42.7	43.3
SQA3D [40]	31.6	63.8	**46.0**	69.5	43.9	45.3	46.6
3D-Vista [84]	34.8	63.3	45.4	69.8	**47.2**	48.1	48.5
Scene-LLM	40.0	**69.2**	42.8	**70.8**	46.6	**52.5**	53.6
Scene-LLM*	**40.9**	69.1	45.0	**70.8**	**47.2**	52.3	**54.2**

可以看到这个模型在SQA3D的表现比其他模型更好，达到了54.2%

Paper 6:

Unifying 3D Vision-Language Understanding via Promptable Queries (ECCV 2024)

From BIGAI, Beijing

链接

结果

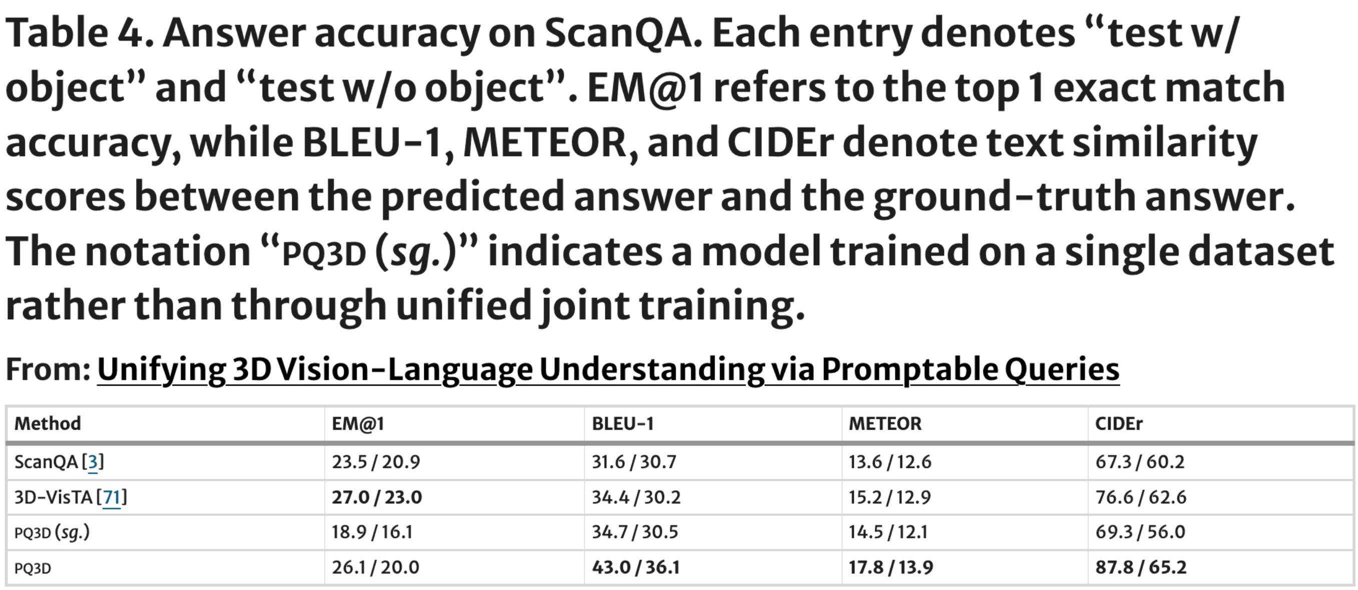


Table 4. Answer accuracy on ScanQA. Each entry denotes "test w/ object" and "test w/o object". EM@1 refers to the top 1 exact match accuracy, while BLEU-1, METEOR, and CIDEr denote text similarity scores between the predicted answer and the ground-truth answer. The notation "PQ3D (*sg.*)" indicates a model trained on a single dataset rather than through unified joint training.

From: Unifying 3D Vision-Language Understanding via Promptable Queries

Method	EM@1	BLEU-1	METEOR	CIDEr
ScanQA [3]	23.5 / 20.9	31.6 / 30.7	13.6 / 12.6	67.3 / 60.2
3D-VisTA [71]	**27.0 / 23.0**	34.4 / 30.2	15.2 / 12.9	76.6 / 62.6
PQ3D (*sg.*)	18.9 / 16.1	34.7 / 30.5	14.5 / 12.1	69.3 / 56.0
PQ3D	26.1 / 20.0	**43.0 / 36.1**	**17.8 / 13.9**	**87.8 / 65.2**

把这个模型放在这里的原因是：这个模型的 MENTOR 和 CIDEr matrix 的表现都比之前的模型好

注：这个模型和 Paper 4 是同一个组做的

Paper 7:

Chat-Scene: Bridging 3D Scene and Large Language Models with Object Identifiers (NIPS 2024)

From 浙大, 上海AI lab, and 字节

链接

结果

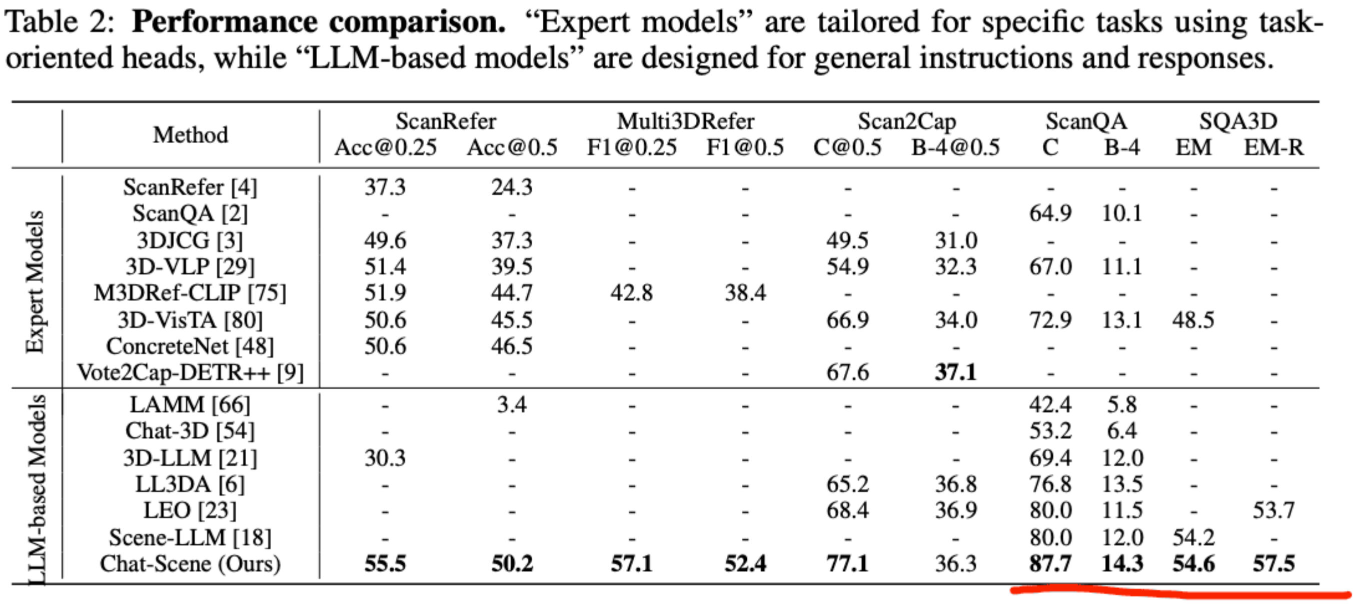


Table 2: **Performance comparison.** "Expert models" are tailored for specific tasks using task-oriented heads, while "LLM-based models" are designed for general instructions and responses.

	Method	ScanRefer		Multi3DRefer		Scan2Cap		ScanQA		SQA3D	
		Acc@0.25	Acc@0.5	F1@0.25	F1@0.5	C@0.5	B-4@0.5	C	B-4	EM	EM-R
Expert Models	ScanRefer [4]	37.3	24.3	-	-	-	-	-	-	-	-
	ScanQA [2]	-	-	-	-	-	-	64.9	10.1	-	-
	3DJCG [3]	49.6	37.3	-	-	49.5	31.0	-	-	-	-
	3D-VLP [29]	51.4	39.5	-	-	54.9	32.3	67.0	11.1	-	-
	M3DRef-CLIP [75]	51.9	44.7	42.8	38.4	-	-	-	-	-	-
	3D-VisTA [80]	50.6	45.5	-	-	66.9	34.0	72.9	13.1	48.5	-
	ConcreteNet [48]	50.6	46.5	-	-	-	-	-	-	-	-
	Vote2Cap-DETR++ [9]	-	-	-	-	67.6	**37.1**	-	-	-	-
LLM-based Models	LAMM [66]	-	3.4	-	-	-	-	42.4	5.8	-	-
	Chat-3D [54]	-	-	-	-	-	-	53.2	6.4	-	-
	3D-LLM [21]	30.3	-	-	-	-	-	69.4	12.0	-	-
	LL3DA [6]	-	-	-	-	65.2	36.8	76.8	13.5	-	-
	LEO [23]	-	-	-	-	68.4	36.9	80.0	11.5	-	53.7
	Scene-LLM [18]	-	-	-	-	-	-	80.0	12.0	54.2	-
	Chat-Scene (Ours)	**55.5**	**50.2**	**57.1**	**52.4**	**77.1**	36.3	**87.7**	**14.3**	**54.6**	**57.5**

这个模型比较了之前的SOTA, 结果SQA3D提高了0.4%的准确率。。。

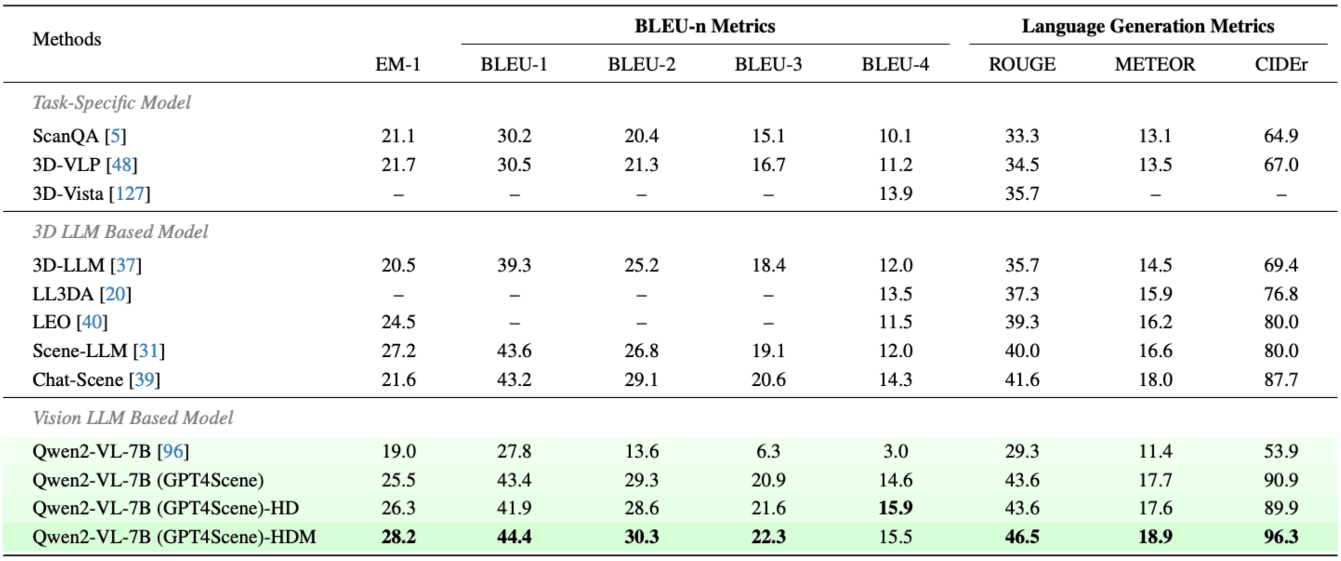
Paper 8:

GPT4Scene: Understand 3D Scenes from Videos with Vision-Language Models (arxiv preprint March 2025)

链接

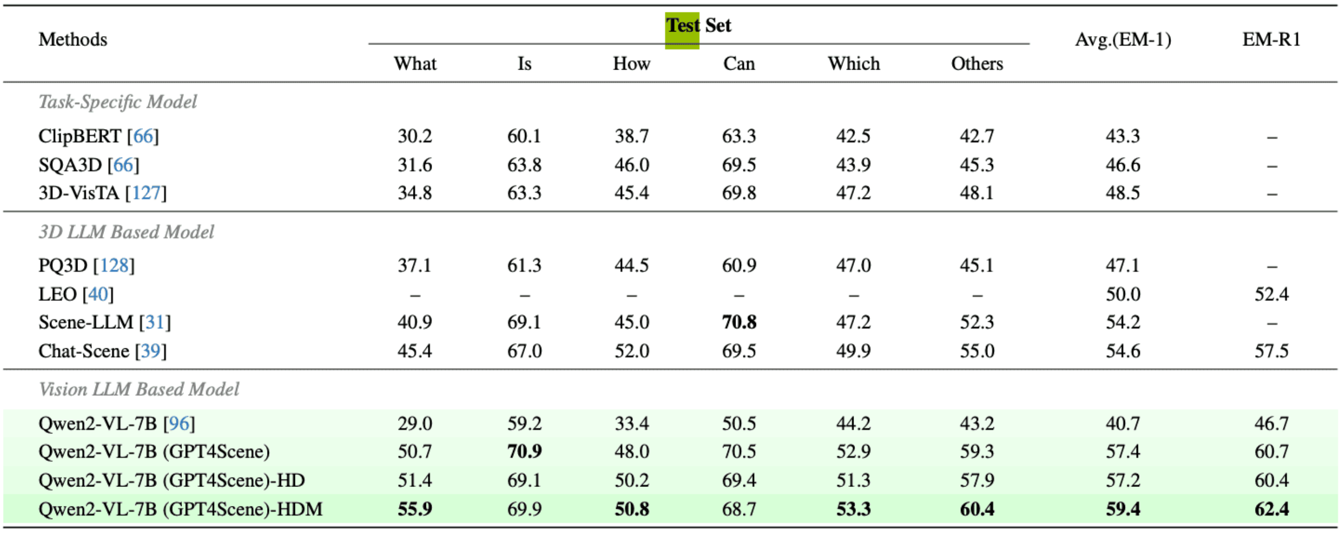
From HKU & 上海AI lab

结果：



Methods		BLEU-n Metrics				Language Generation Metrics		
	EM-1	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	METEOR	CIDEr
Task-Specific Model								
ScanQA [5]	21.1	30.2	20.4	15.1	10.1	33.3	13.1	64.9
3D-VLP [48]	21.7	30.5	21.3	16.7	11.2	34.5	13.5	67.0
3D-Vista [127]	–	–	–	–	13.9	35.7	–	–
3D LLM Based Model								
3D-LLM [37]	20.5	39.3	25.2	18.4	12.0	35.7	14.5	69.4
LL3DA [20]	–	–	–	–	13.5	37.3	15.9	76.8
LEO [40]	24.5	–	–	–	11.5	39.3	16.2	80.0
Scene-LLM [31]	27.2	43.6	26.8	19.1	12.0	40.0	16.6	80.0
Chat-Scene [39]	21.6	43.2	29.1	20.6	14.3	41.6	18.0	87.7
Vision LLM Based Model								
Qwen2-VL-7B [96]	19.0	27.8	13.6	6.3	3.0	29.3	11.4	53.9
Qwen2-VL-7B (GPT4Scene)	25.5	43.4	29.3	20.9	14.6	43.6	17.7	90.9
Qwen2-VL-7B (GPT4Scene)-HD	26.3	41.9	28.6	21.6	**15.9**	43.6	17.6	89.9
Qwen2-VL-7B (GPT4Scene)-HDM	**28.2**	**44.4**	**30.3**	**22.3**	15.5	**46.5**	**18.9**	**96.3**

Table 12. **Full Evaluation of 3D Question Answering on ScanQA [5].**



Methods	Test Set						Avg.(EM-1)	EM-R1
	What	Is	How	Can	Which	Others		
Task-Specific Model								
ClipBERT [66]	30.2	60.1	38.7	63.3	42.5	42.7	43.3	–
SQA3D [66]	31.6	63.8	46.0	69.5	43.9	45.3	46.6	–
3D-VisTA [127]	34.8	63.3	45.4	69.8	47.2	48.1	48.5	–
3D LLM Based Model								
PQ3D [128]	37.1	61.3	44.5	60.9	47.0	45.1	47.1	–
LEO [40]	–	–	–	–	–	–	50.0	52.4
Scene-LLM [31]	40.9	69.1	45.0	**70.8**	47.2	52.3	54.2	–
Chat-Scene [39]	45.4	67.0	52.0	69.5	49.9	55.0	54.6	57.5
Vision LLM Based Model								
Qwen2-VL-7B [96]	29.0	59.2	33.4	50.5	44.2	43.2	40.7	46.7
Qwen2-VL-7B (GPT4Scene)	50.7	**70.9**	48.0	70.5	52.9	59.3	57.4	60.7
Qwen2-VL-7B (GPT4Scene)-HD	51.4	69.1	50.2	69.4	51.3	57.9	57.2	60.4
Qwen2-VL-7B (GPT4Scene)-HDM	**55.9**	69.9	**50.8**	68.7	**53.3**	**60.4**	**59.4**	**62.4**

Table 13. **Full Evaluation of 3D Question Answering on SQA3D [66].**

这个模型在SQA3D上做到了62.4%的准确率

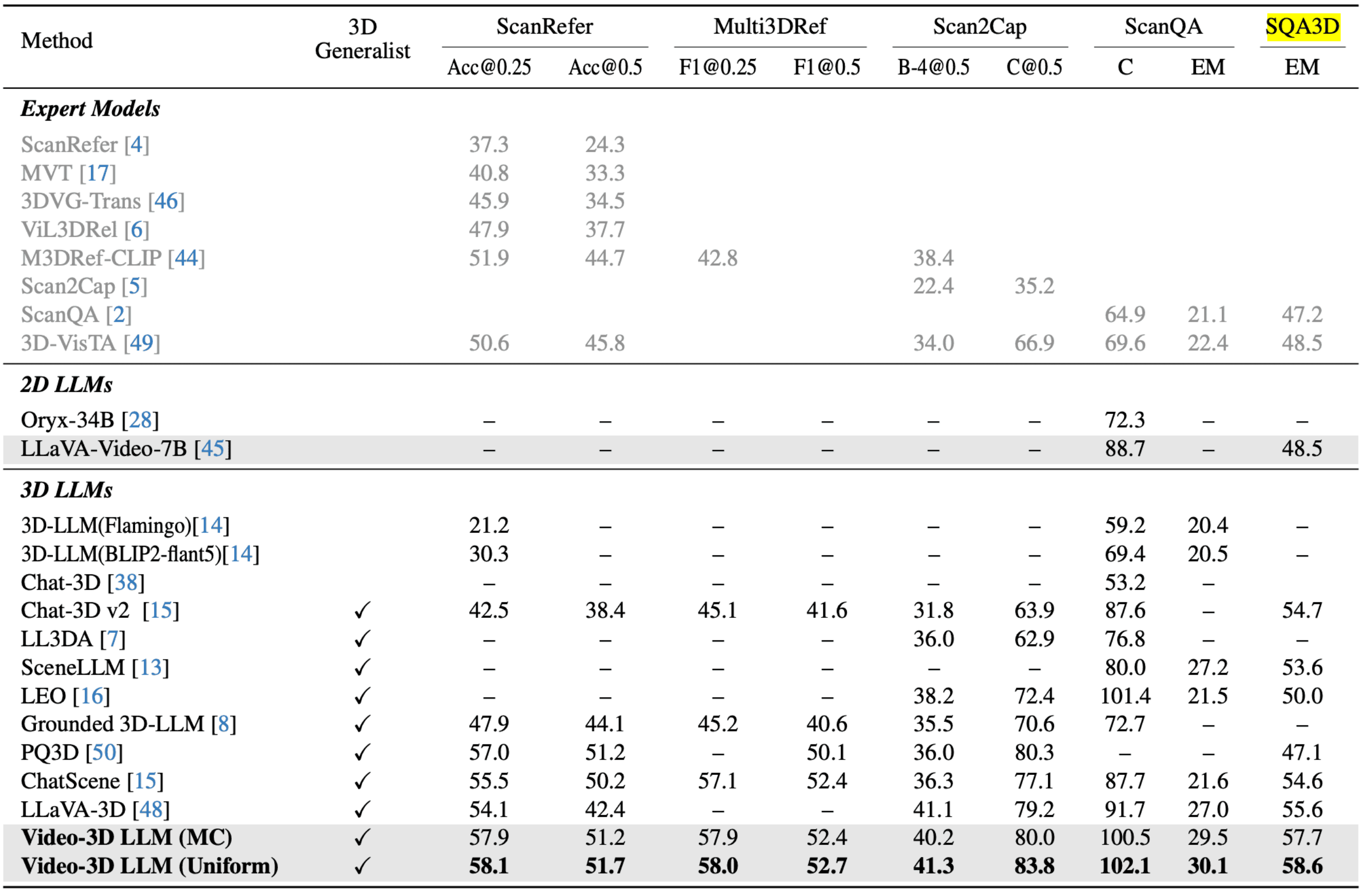
Paper 9:

Video-3D LLM: Learning Position-Aware Video Representation for 3D Scene Understanding (CVPR 2025)

链接

From The Chinese University of Hong Kong

结果



Method	3D Generalist	ScanRefer		Multi3DRef		Scan2Cap		ScanQA		SQA3D
		Acc@0.25	Acc@0.5	F1@0.25	F1@0.5	B-4@0.5	C@0.5	C	EM	EM
Expert Models										
ScanRefer [4]		37.3	24.3							
MVT [17]		40.8	33.3							
3DVG-Trans [46]		45.9	34.5							
ViL3DRel [6]		47.9	37.7							
M3DRef-CLIP [44]		51.9	44.7	42.8		38.4				
Scan2Cap [5]						22.4	35.2			
ScanQA [2]								64.9	21.1	47.2
3D-VisTA [49]		50.6	45.8			34.0	66.9	69.6	22.4	48.5
2D LLMs										
Oryx-34B [28]		–	–	–	–	–	–	72.3	–	–
LLaVA-Video-7B [45]		–	–	–	–	–	–	88.7	–	48.5
3D LLMs										
3D-LLM(Flamingo)[14]		21.2	–	–	–	–	–	59.2	20.4	–
3D-LLM(BLIP2-flant5)[14]		30.3	–	–	–	–	–	69.4	20.5	–
Chat-3D [38]		–	–	–	–	–	–	53.2	–	
Chat-3D v2 [15]	✓	42.5	38.4	45.1	41.6	31.8	63.9	87.6	–	54.7
LL3DA [7]	✓	–	–	–	–	36.0	62.9	76.8	–	–
SceneLLM [13]	✓	–	–	–	–	–	–	80.0	27.2	53.6
LEO [16]	✓	–	–	–	–	38.2	72.4	101.4	21.5	50.0
Grounded 3D-LLM [8]	✓	47.9	44.1	45.2	40.6	35.5	70.6	72.7	–	–
PQ3D [50]	✓	57.0	51.2	–	50.1	36.0	80.3	–	–	47.1
ChatScene [15]	✓	55.5	50.2	57.1	52.4	36.3	77.1	87.7	21.6	54.6
LLaVA-3D [48]	✓	54.1	42.4	–	–	41.1	79.2	91.7	27.0	55.6
Video-3D LLM (MC)	✓	57.9	51.2	57.9	52.4	40.2	80.0	100.5	29.5	57.7
Video-3D LLM (Uniform)	✓	**58.1**	**51.7**	**58.0**	**52.7**	**41.3**	**83.8**	**102.1**	**30.1**	**58.6**

这个模型在ScanQA的CIDEr上做到了SOTA

Paper 10 (To be continued):

3D-R1: Enhancing Reasoning in 3D VLMs for Unified Scene Understanding (arxiv preprint July 2025)

From Shanghai University of Engineering Science & PKU

链接

结果：

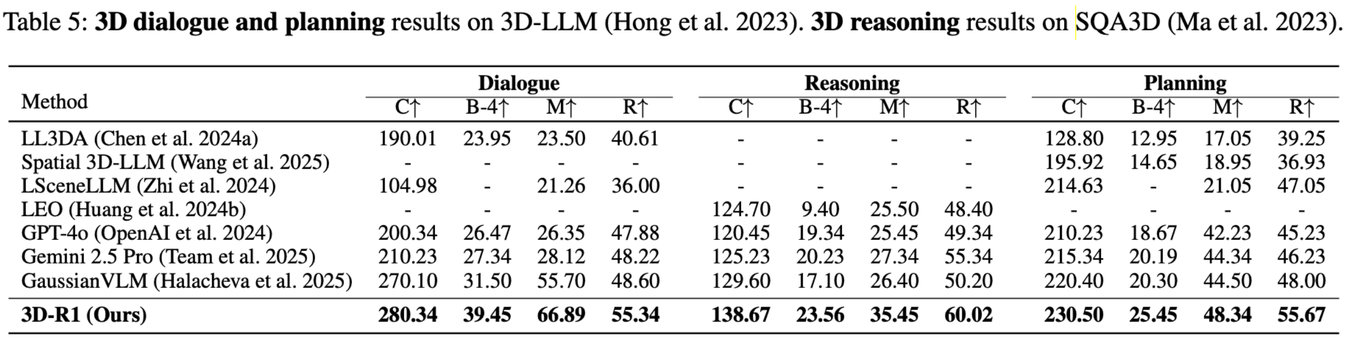


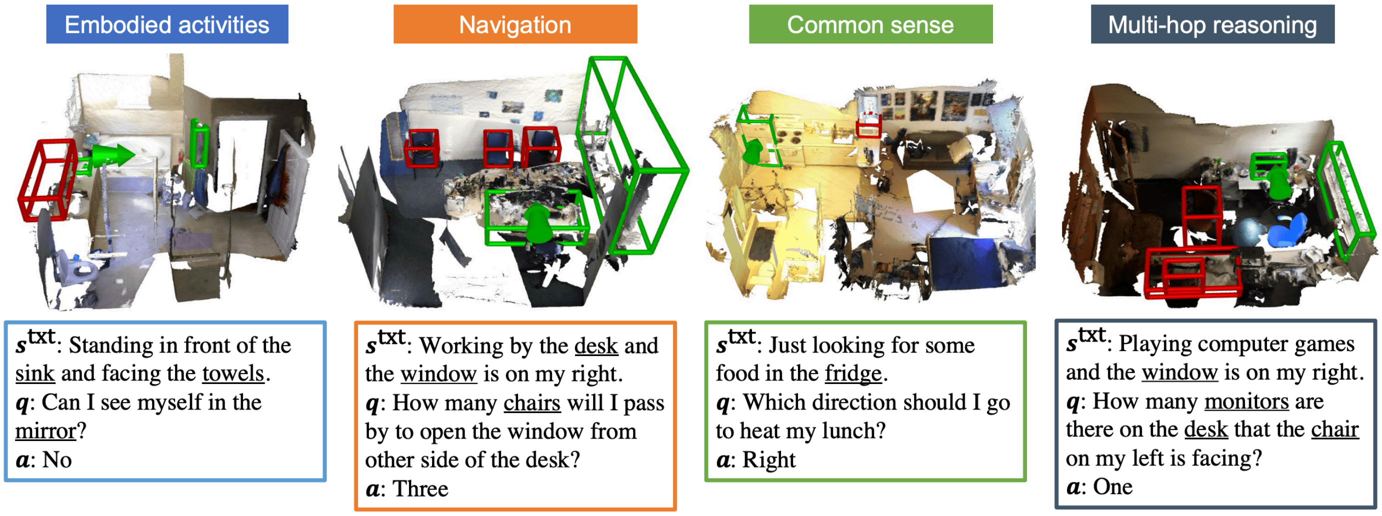
Table 5: **3D dialogue and planning** results on 3D-LLM (Hong et al. 2023). **3D reasoning** results on SQA3D (Ma et al. 2023).

Method	Dialogue				Reasoning				Planning			
	C↑	B-4↑	M↑	R↑	C↑	B-4↑	M↑	R↑	C↑	B-4↑	M↑	R↑
LL3DA (Chen et al. 2024a)	190.01	23.95	23.50	40.61	-	-	-	-	128.80	12.95	17.05	39.25
Spatial 3D-LLM (Wang et al. 2025)	-	-	-	-	-	-	-	-	195.92	14.65	18.95	36.93
LSceneLLM (Zhi et al. 2024)	104.98	-	21.26	36.00	-	-	-	-	214.63	-	21.05	47.05
LEO (Huang et al. 2024b)	-	-	-	-	124.70	9.40	25.50	48.40	-	-	-	-
GPT-4o (OpenAI et al. 2024)	200.34	26.47	26.35	47.88	120.45	19.34	25.45	49.34	210.23	18.67	42.23	45.23
Gemini 2.5 Pro (Team et al. 2025)	210.23	27.34	28.12	48.22	125.23	20.23	27.34	55.34	215.34	20.19	44.34	46.23
GaussianVLM (Halacheva et al. 2025)	270.10	31.50	55.70	48.60	129.60	17.10	26.40	50.20	220.40	20.30	44.50	48.00
3D-R1 (Ours)	**280.34**	**39.45**	**66.89**	**55.34**	**138.67**	**23.56**	**35.45**	**60.02**	**230.50**	**25.45**	**48.34**	**55.67**

Report了非常奇怪的结果，需要进一步细看

Benchmark 1: SQA3D

SQA3D



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

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

上图是一个询问的例子

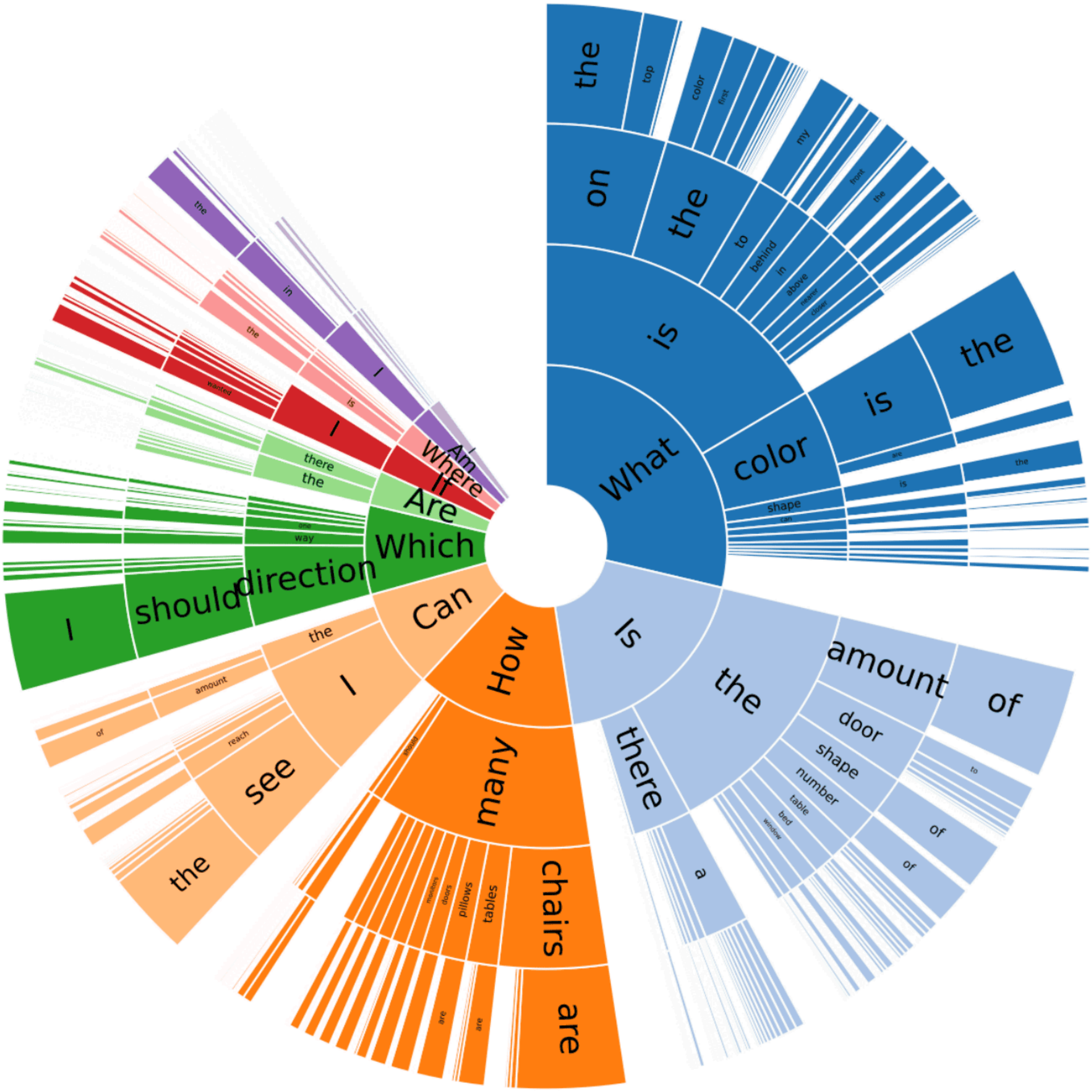
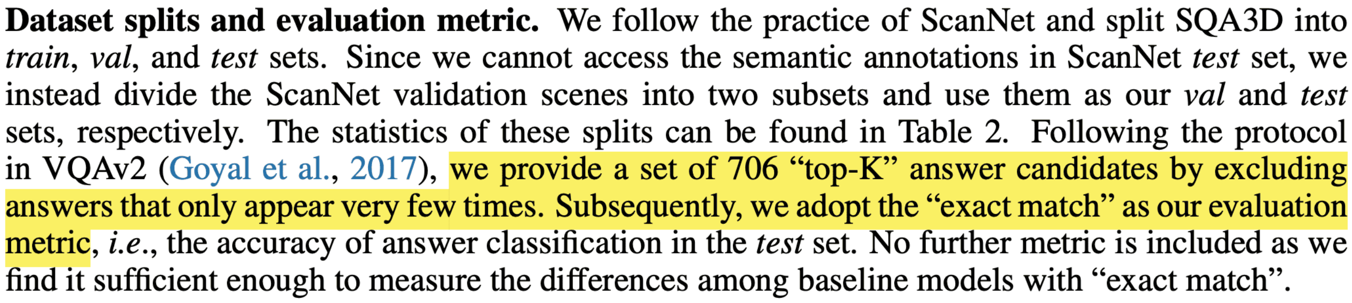


Figure 5: Question distribution in SQA3D

这个图解释了为什么有些方法会用 "Which" "How" 这些来分类。 总结来说，以这几个词开头的问句较多，且方向不同

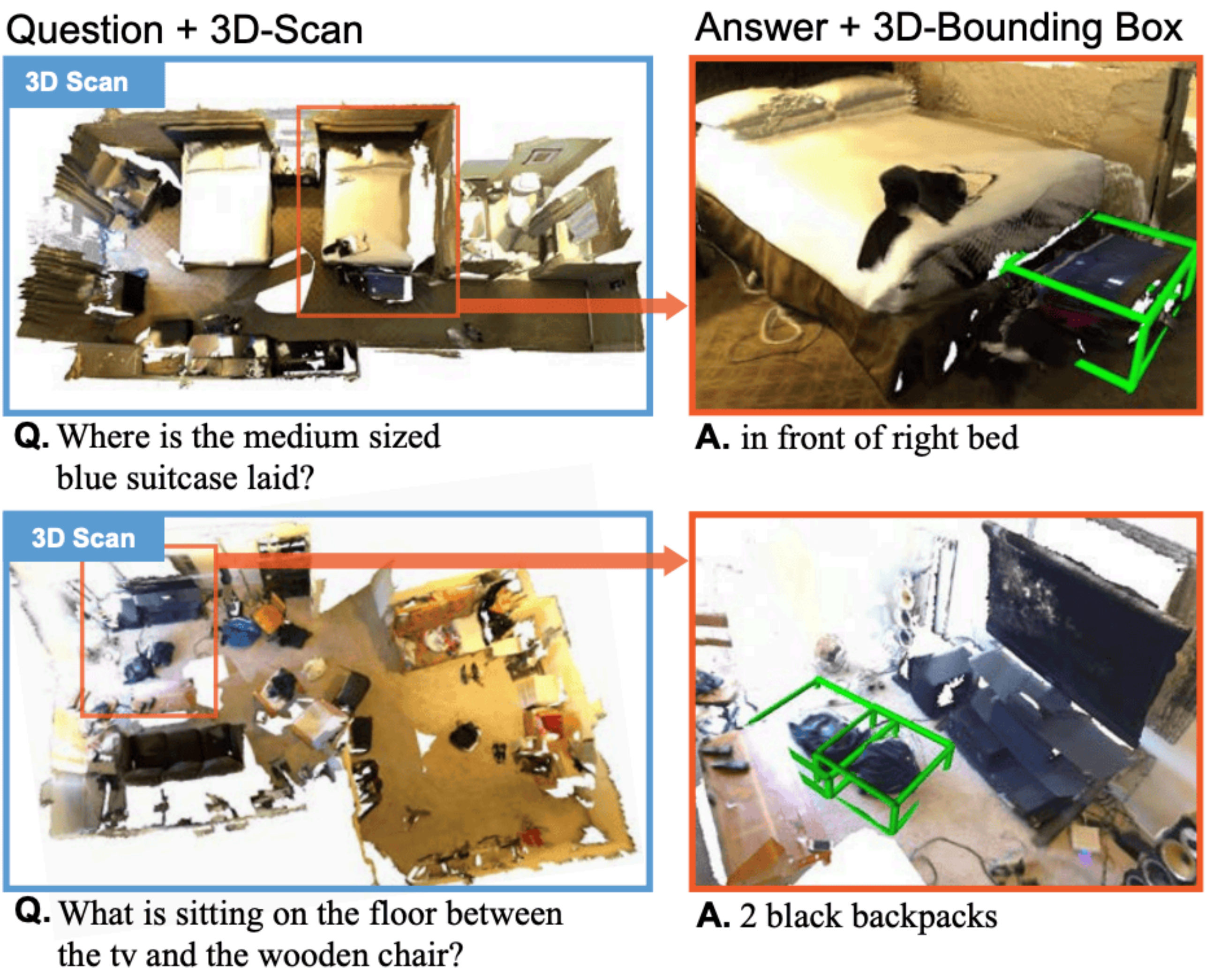


Dataset splits and evaluation metric. We follow the practice of ScanNet and split SQA3D into *train*, *val*, and *test* sets. Since we cannot access the semantic annotations in ScanNet *test* set, we instead divide the ScanNet validation scenes into two subsets and use them as our *val* and *test* sets, respectively. The statistics of these splits can be found in Table 2. Following the protocol in VQAv2 (Goyal et al., 2017), we provide a set of 706 “top-K” answer candidates by excluding answers that only appear very few times. Subsequently, we adopt the “exact match” as our evaluation metric, *i.e.*, the accuracy of answer classification in the *test* set. No further metric is included as we find it sufficient enough to measure the differences among baseline models with “exact match”.

最后统计的是准确率(按照文章中的说法，居然是做一个706维度的分类？）需要和Ground Truth完全一致

Benchmark 2: ScanQA

ScanQA



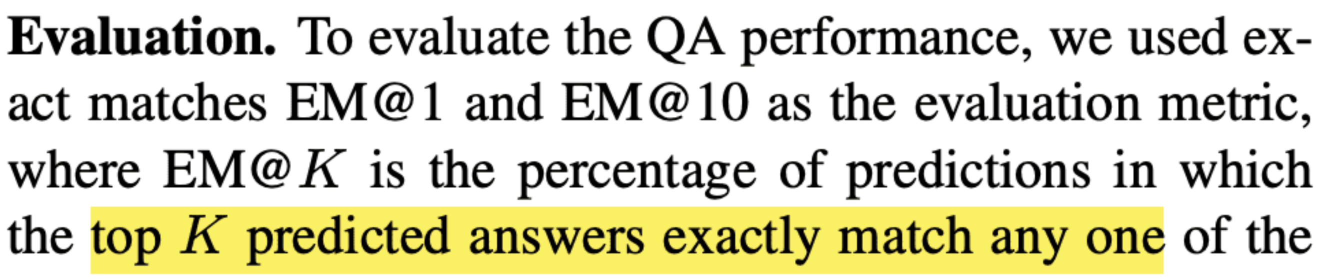
Q. Where is the medium sized blue suitcase laid?

A. in front of right bed

Q. What is sitting on the floor between the tv and the wooden chair?

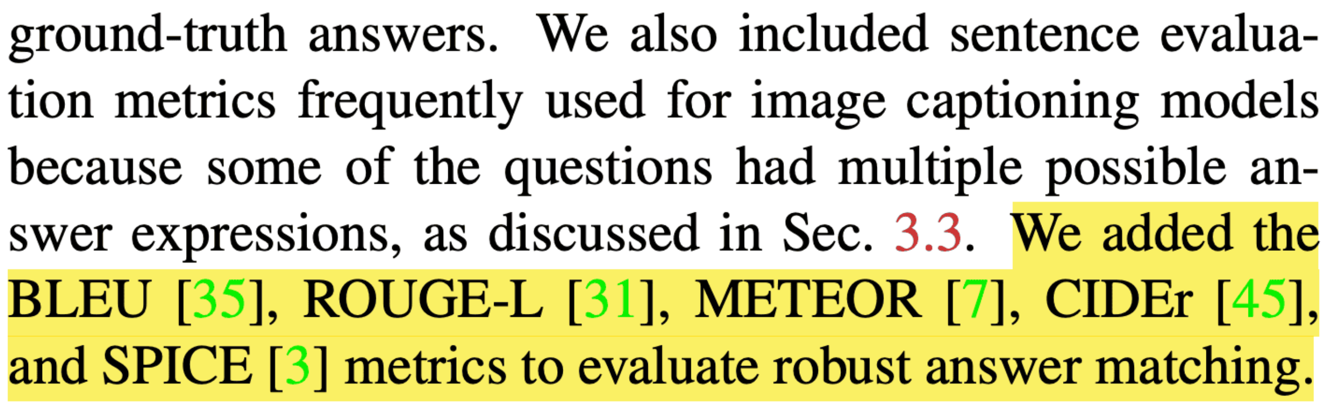
A. 2 black backpacks

上图是一个询问的例子



Evaluation. To evaluate the QA performance, we used exact matches EM@1 and EM@10 as the evaluation metric, where EM@K is the percentage of predictions in which the top K predicted answers exactly match any one of the

同样是准确率 需要和Ground Truth完全一致



ground-truth answers. We also included sentence evaluation metrics frequently used for image captioning models because some of the questions had multiple possible answer expressions, as discussed in Sec. 3.3. We added the BLEU [35], ROUGE-L [31], METEOR [7], CIDEr [45], and SPICE [3] metrics to evaluate robust answer matching.

和上面不同，这里也加入了其他metric去算答案和Ground Truth的相似性