

3D Large Language Models: Are We on the Right Path?

Siyuan Huang Beijing Institute for General Artificial Intelligence (BIGAI)

Figure generated by ChatGPT 40

Multi-modal LLMs

State-of-the-art multi-modal LLMs can communicate with humans in natural language with a sophisticated understanding of images and videos

Can they understand 3D scenes like humans?



From 2D Images to 3D Scenes



2D image

Pixel grid

E.g., ImageNet: 1.2M images -> 150G

5~10 objects

Few explicit spatial relations

Limited occlusion reasoning



3D scene

Mesh, points, voxels...

E.g., ScanNet: 1.5k scenes -> over 1T

20~100 objects

Hundreds of spatial relations

Complex occlusion and clutter

Challenges of 3D scene understanding



Intricate geometry and semantics





Complicated spatial relations

Embodied reasoning with long contexts

3D Vision-Language Tasks



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.

3D Question answering: ScanQA (CVPR 2022)



3D Object grounding: ScanRefer (ECCV 2020)



Situated and embodied reasoning: SQA3D (ICLR 2023), OpenEQA (CVPR 2024)

Can 2D MLLMs Accomplish these Tasks?

VSI-Bench (CVPR 2025)



How to train a 3D LLM that performs better?

3D-VL Datasets: Grounding and QA

3D Visual Grounding and Scaling Up



ScanRefer (ECCV 2020): 3D object grounding with natural language descriptions



ReferIt3D (ECCV 2020): 3D object grounding with finegrained spatial reasoning



Multi3DRefer (ICCV 2023): 3D multi-object grounding



the tv and the wooden chair? ScanQA (CVPR 2022): question answering in 3D scenes

A. 2 black backpacks

What is sitting on the floor between

3D Question-Answering



 $\label{eq:Question q: Can I go straight to the coffee table in front of me?}$

Scene context *S*: 3D scan, egocentric video, birdeye view (BEV) picture, etc.



SQA3D (ICLR 2023): situated reasoning in 3D scenes



SceneVerse (ECCV 2024): scaling up 3D-VL grounding



3D-GRAND (CVPR 2025): dense 3D-VL grounding and mitigating hallucination



MSR3D (NeurIPS 2024): multi-modal situated reasoning and one-step navigation in 3D scenes

Training 3D-LLMs with 3D VL Datasets



3D-VisTA (ICCV 2023): unified 3D-VL pretraining with object-centric representation



LLM-Grounder (ICRA 2024): LLM agent for open-vocabulary zero-shot 3D object grounding

Models focused on 3D object grounding



3D-LLM (NeurIPS 2023): 3D MLLM leveraging multi-view 2D features for 3D scene understanding



LL3DA (CVPR 2024): 3D language assistant with scene-level point cloud perception

Models with improved 3D scene embedding

Embodied Generalist Agent in 3D World Robotic Manipulatio 2 P = [0.1, -0.2, 0]R = [0, 0, 0, 1] 3D Question EO Answering Navigation EO: Standing on the desi Large Language Mode 3D Dialogue EO: Of course! What spe ally do you need help with LEO: A 3D model of a Southern Whit EO: I'm sorry, but the e Make the hed

LEO (ICML 2024): first VLA generalist agent in 3D scenes with unified instruction tuning



Assistant: ... The living area is furnished with an L-shaped couch (<OBJ032>) facing a brown entertainment center housing a black TV (<OBJ015>). A gray ottoman coffice table (<OBJ034>) is placed in front of the couch. A wooden shelf (<OBJ056>) enclosing two sections of books adds a coxy touch to the space...

Chat-Scene (NeurIPS 2024): 3D-VL generalist model integrating object-centric 2D-3D features

Models with object-centric representation

Can 3D-LLMs Generalize?



Despite increasing performances on existing 3D-VL benchmarks, 3D-LLMs still cannot generalize well and fail in numerous simple scenarios

Do we train and evaluate 3D-LLMs correctly?

Mist over 3D-VL Understanding: Flawed Test Data

Grounding: ScanRefer, Nr3D

Humans achieve 69% accuracy on ScanRefer and 62% accuracy on ScanQA



View-dependent: "left", "right"

and the same hit is a second se



Ambiguous Question Q: Where is the <u>white cabinet</u> in relation to the <u>couch</u>? GT: [to <u>left</u> of black couch ; to <u>right</u> of it] Multiple referred objects: "cabinet", "couch"



Multiple referred objects: "monitor"

Ambiguous Question Q: How many shelfs are on <u>the book shelf</u>? GT: [<u>3</u>; <u>seventeen</u>]

Multiple referred objects: "book shelf"



Language quality: confusing target object

Incomplete Answers Q: What color is the surface of the table? GT: [green ; <u>black</u>]

Missing answers: "brown"



Confusing description: "short long"

Incomplete Answers Q: What is next to the organizer shelf? GT: [coat rack]

Missing answers: "bed", "desk"...

QA: ScanQA

Mist over 3D-VL Understanding: Insufficient Evaluation Metrics for Model Pitfalls

Model pitfall 1: Visual ignorance

- Scenario: models perform VL tasks without the need for visual input
- Evidence: blind LLMs achieve good results on SQA3D
- Lessons: test data quality; metrics to identify these shortcuts

Blind LLM	OPT-1.3B	Gemma2-2B	Vicuna-7B	LLaMA3-3B	LLaVA-3D ^{\dagger}
EM-1	43.9	48.8	49.4	50.0	55.6



Mist over 3D-VL Understanding: Insufficient Evaluation Metrics for Model Pitfalls

Model pitfall 2: Weak language robustness

- Scenario: models are susceptible to language variations
- > Evidence: rephrasing input texts elicits different results
- Lessons: we need robust metrics that incorporate language variations



Mist over 3D-VL Understanding: Isolation of Grounding and QA

Why do models fail in 3D QA task? Failure in grounding or answering?

Evaluating the coherence between grounding and QA



grounding-QA coherence

Grounding-QA Chains

Beacon3D for 3D-VL Grounding and QA Tasks

- Limitations of existing 3D-VL benchmarks
- 📚 Flawed test data
- Insufficient evaluation metrics
- \oslash Isolation of grounding and QA tasks

- Highlights of the Beacon3D benchmark
- 📚 High-quality test data
- **Object-centric evaluation metrics**
- Ø Grounding Chain and Grounding-QA Chain

Unveiling the Mist over 3D Vision-Language Understanding: Object-centric Evaluation with Chain-of-Analysis CVPR 2025

Jiangyong Huang* Baoxiong Jia* Yan Wang Ziyu Zhu Xiongkun Linghu Qing Li Song-Chun Zhu Siyuan Huang



https://beacon-3d.github.io





Object-centric Evaluation with Chain-of-Analysis





Ø Grounding-QA Chain



High-quality test data



- > 800+ objects across ScanNet, 3RScan, and MultiScan
- Diverse knowledge types: class, appearance, geometry, spatial, and existence



Evaluation Metrics: from Per-Case to Per-Object

- ◆ Object-centric metrics elicit a significant performance drops, particularly in QA
- Current 3D-VL models lack a comprehensive understanding of objects and are fragile to language variations

					<u></u>	· = = = 1
	ł	Knowled	Ove	erall		
	Class	App.	Geo.	Spa.	Case	Obj.
w/o LLM					1	i
ViL3DRel	61.8	66.9	46.5	59.5	61.8	39.8
3D-VisTA	71.0	64.6	56.3	68.9	71.0	50.9
PQ3D	76.1	71.2	66.0	74.5	76.1	57.2
SceneVerse	73.4	64.9	64.6	71.9	73.5	52.1
LLM-based					1	
LEO-multi	14.3	10.9	15.3	15.1	14.3	2.8
LEO-curricular	22.0	22.2	20.8	15.4	22.0	3.8
PQ3D-LLM	70.3	66.2	53.5	68.3	70.2	47.4
Chat-Scene	62.7	57.3	56.3	57.8	62.7	44.3
						. <u> </u>

						<u> </u>	
		Know	Ove	erall			
	Class	App.	Geo.	Spa.	Exi.	Case	Obj.
w/o LLM						1	
3D-VisTA	20.5	33.5	52.1	33.8	36.5	35.3	8.1
PQ3D	36.4	28.0	27.8	11.9	45.5	27.8	3.5
SceneVerse	35.6	41.7	48.9	41.9	35.7	40.3	6.6
LLM-based							
GPT-40 [†]	33.3	49.9	54.9	52.1	73.8	57.1	20.2
LEO-multi	25.8	37.7	52.8	46.2	37.4	41.1	3.5
LEO-curricular	17.4	41.0	53.2	48.7	39.7	43.2	7.8
PQ3D-LLM	28.0	30.8	35.2	25.2	26.2	27.9	2.3
Chat-Scene	36.4	39.8	56.7	47.6	48.8	45.8	7.8

Grounding

Chain-of-Analysis: from Grounding to QA

> Weak grounding-QA coherence (both correct $\leq 20\%$)

- \succ $R_1 \sim 50\%$: half of QA failure stems from <u>bad answering skills</u>
- \triangleright R₂ ~ 50%: half of QA success stems from shortcut or overfitting



Effects of LLMs on 3D-VL Tasks

Incorporating LLMs weakens grounding

- LLM directly used for grounding: Chat-Scene (blue boxes)
- LLM not directly used for grounding: PQ3D-LLM (purple boxes)

	ł	Knowled	lge type		Overall		
	Class	App.	Geo.	Spa.	Case	Obj.	
w/o LLM							
ViL3DRel	61.8	66.9	46.5	59.5	61.8	39.8	
3D-VisTA	71.0	64.6	56.3	68.9	71.0	50.9	
PQ3D	76.1	71.2	66.0	74.5	76.1	57.2	
SceneVerse	73.4	64.9	64.6	71.9	73.5	52.1	
LLM-based							
LEO-multi	14.3	10.9	15.3	15.1	14.3	2.8	
LEO-curricular	22.0	22.2	20.8	15.4	22.0	3.8	
PQ3D-LLM	70.3	66.2	53.5	68.3	70.2	47.4	
Chat-Scene	62.7	57.3	56.3	57.8	62.7	44.3	

Effects of LLMs on 3D-VL Tasks

Incorporating LLMs does not fundamentally enhance QA

- ➢ Results: w/o LLM ≈ LLM-based (e.g., PQ3D vs. PQ3D-LLM)
- Why LLMs may not help QA
 - Current 3D QA task only involves basic language modeling
 - Bottleneck lies in perception and VL alignment

		Knov		Overall			
	Class	App.	Geo.	Spa.	Exi.	Case	Obj.
w/o LLM							
3D-VisTA	20.5	33.5	52.1	33.8	36.5	35.3	8.1
PQ3D	36.4	28.0	27.8	11.9	45.5	27.8	3.5
SceneVerse	35.6	41.7	48.9	41.9	35.7	40.3	6.6
LLM-based							
GPT-40 [†]	33.3	49.9	54.9	52.1	73.8	57.1	20.2
LEO-multi	25.8	37.7	52.8	46.2	37.4	41.1	3.5
LEO-curricular	17.4	41.0	53.2	48.7	39.7	43.2	7.8
PQ3D-LLM	28.0	30.8	35.2	25.2	26.2	27.9	2.3
Chat-Scene	36.4	39.8	56.7	47.6	48.8	45.8	7.8

Object-centric metrics: w/o LLM ≈ LLM-based



Grounding-QA coherence: w/o LLM ≥ LLM-based

Beacon3D Takeaways

Beacon3D: a novel benchmark and evaluation protocol for 3D-VL grounding and QA tasks

- High-quality test data and robust object-centric metrics
- Chained analysis reveals task intrinsic gap and prominent overfitting risk

How to improve current 3D-VL models

- Perception: endow 3D-LLMs with stronger perception capability
- ➢ Reasoning: properly unleash the potential of LLMs for 3D-VL tasks (e.g., CoT reasoning)



Endow 3D-VL models with stronger perception capability



Properly unleash the potential of LLMs for 3D-VL tasks (*e.g.*, CoT reasoning)

Scaling Up with Efficient Representation



LEO-VL: Towards 3D Vision-Language Generalists via Data Scaling with Efficient Representation

Jiangyong Huang Xiaojian Ma Xiongkun Linghu Yue Fan Junchao He Wenxin Tan Qing Li Song-Chun Zhu Yixin Chen Baoxiong Jia Siyuan Huang







Efficient Representation Bridging 2D-3D Perception



LEO-VL: bridge 2D-3D perception with improved efficiency, unlocking data scalability

SOTA Performance with Higher Efficiency

Table 2: **Comparison with state-of-the-art models on 3D-VL benchmarks.** Benchmarks are colorized according to scene domains: ScanNet, 3RScan, and ARKitScenes. "C" stands for "CIDEr", "B-4" for BLEU-4, "EM" for top-1 exact match, "EM-R" for refined top-1 exact match [32], and "Obj." for object-centric metrics [33]. Gray figures indicate extra inductive bias in object order. [†] indicates the original results in the paper.

M - 1-1	Saana (#talvana)	ScanQA (val)		SQA3D (test)		MSQA (test)	Beacon3D		MSQA (test)	MSQA (test)	
Model	Scene (#tokens)	C	B-4	EM	EM-R	GPT-Score	Case	Obj.	GPT-Score	GPT-Score	
ScanQA [3]	Query (256)	64.9	10.1	47.2	-	-	-	-	-	-	
3D-LLM [27]	Query (32)	74.5	12.9	49.8	-	-	-	-	-	-	
PQ3D [113]	Query (80)	-	-	47.1	-	-	35.9	4.2	-	-	
DSPNet [49]	Query (256)	-	-	50.4	-	-	-	-	-	-	
3D-LLaVA [19]	Query (100)	92.6	17.1	54.5	56.6	-	-	-	-	-	
3D-VisTA [112]	Object (80)	69.6	10.4	48.5	-	-	35.3 [†]	8.1^{+}	-	-	
LEO [32]	Object (60)	101.4	13.2	50.0	52.4	56.7 [†]	45.2	7.5	50.4^{\dagger}	59.7^{\dagger}	
SceneVerse [37]	Object (80)	-	-	49.9	-	-	40.5	4.7	-	-	
Chat-Scene [30]	Object (200)	87.7	14.3	54.6	57.5	-	45.8^{+}	7.8^{\dagger}	-	-	
Inst3D-LMM [95]	Object (200)	88.6	14.9	-	-	-	-	-	-	-	
Scene-LLM [24]	Voxel (0.18m)	80.0	11.7	53.6	-	-	-	-	-	-	
LLaVA-3D [110]	Voxel (3096)	91.7	14.5	55.6	57.6	-	-	-	-	-	
Video-3D LLM [106]	Video (32 frames)	100.5	16.3	57.7	-	-	-	-	-	-	
GPT4Scene [68]	Video (32 frames)	96.3	15.5	59.4	62.4	-	-	-	-	-	
LEO-VL	Grid (750)	97.6	14.8	59.7	62.6	58.2	59.5	18.2	51.4	64.2	





Instruction: Detail the object at the selected spot.

Response: The nightstand is a wooden furniture piece with a rectangular shape and a brown color. It has two drawers and a shelf, and can be used as a bedside table. The texture of the nightstand is wooden, and it has a simple and classic design.



Instruction: *I am inspecting the trash can. Draft a concise portrayal of this 3D scene.*

Response: To my left, in the middle distance, there is a black and silver stove. It's made of metal and glass, used for cooking with a smooth and glossy texture. The stove has a flat top with burners and a control panel, and it's currently off. In the middle distance to my right, there is a kitchen cabinet. This room appears to be a wellequipped kitchen or dining area, with various appliances, storage solutions, and seating options.

Instruction: Look at the organizer shelf next to the desk, is it taller or shorter than the desk?

Response: Taller.



Instruction: *Provide a directive for the task: organize and tidy up the living area.*

Response: 1. Dust and clean the surfaces of the furniture, such as the sofa, table, and cabinet. 2. Organize items on the table, ensuring that they are neatly arranged and not cluttered.

3. Clean the TV screen and ensure it is turned off when not in use.

4. Sweep or vacuum the floor to remove any dirt or debris.

5. Adjust the position of the sofa if needed for better seating arrangement.

6. Check and adjust the temperature of the heater if necessary.

7. Remove any unnecessary items or clutter from the living area to create a more organized space.



Response: The remote control is placed within the area of the TV stand.

Instruction: Where can I find the monitor in the room?

Response: The monitor is resting on the desk.

Instruction: Thanks for the detailed information.

Response: You're welcome! If you have any more questions, feel free to ask.



CoT for 3D Scene Understanding

SCENECOT

Eliciting Chain-of-Thought Reasoning in 3D Scenes







Jiangyong Huang



Ziyu Zhu



Baoxiong Jia



Siyuan Huang

https://scenecot.github.io/



Limitations of existing 3D-LLMs:

- Lack of explicit visual grounding procedures
- Lack of explainability for complex spatial reasoning

Our Goal:

- Implement explicit visual grounding for spatial reasoning
- Build an explainable step-by-step reasoning framework in 3D scenes

Existing 3D LLMs Black **Model Input SceneCOT** I am facing a desk. There is a brown chair on my left. What is the color of the <think_type>This is a question of attribute.</think type> bike on my 2 o'clock? <grd rgn> <think rgn>Now I need to list all the objects on my 2 o'clock.</think rgn> <think grd>Ground the object: The bike on my 2 o'clock.</think grd> [OBJ] <think task>Now I need to retrieve the image of the target object. <highlight_obj> <img_start> <think sum>Now answer the question based on the object image.</think sum>

<answer>Silver</answer>

SceneCOT: Step-by-step Reasoning in 3D Scenes



SceneCOT: Training and Inference

Training

Inference



SceneCOT: SOTA Performance in 3D-VL Reasoning

- SceneCOT demonstrates its strength in situated reasoning (SQA3D and MSQA) and object-centric reasoning (Beacon3D)
- Particularly strong performance on counting and spatial relations

Methods	Scene Input	MSQA							Beacon3D		SQA3D	
		Count.	Exist.	Attr.	Spatial	Navi.	Others	Overall	Case	Obj.	EM	EM-R
GPT-4o*‡	TXT	32.3	79.3	79.0	37.0	31.7	91.6	52.3	57.1	20.2	_	_
LEO‡	PCD	32.5	88.5	58.7	44.2	39.6	81.4	54.8	43.2	7.8	50.0	52.4
MSR3D‡	PCD	32.3	93.1	50.0	46.5	54.1	75.6	54.2	_	_	_	52.9
PQ3D	PCD+IMG	_	_	_	_	_	_	_	27.8	3.5	47.1	_
SceneVerse	PCD	_	_	_	_	_	_	_	40.3	6.6	49.9	_
ChatScene	PCD+IMG	_	_	_	_	_	_	_	45.8	7.8	54.6	57 5
SplatTalk	IMG	19.6	60.3	44.0	35.8	35.5	61.8	41.8	_	_	47.6	49.9
SCENECOT (S)	PCD+IMG	_	_	_	_	_	_	_	_	_	37.7	41.0
SCENECOT (M+G)	PCD+IMG	47.9	82.1	49.6	47.2	51.6	80.3	55.6	58.9	23.2	_	_
SCENECOT (M+S+G)	PCD+IMG	47.0	84.2	43.8	47.9	47.9	78.7	54.0	60.8	22.6	56.5	59.3

Summary: What is the Right Path?



Perception: LEO-VL

Reasoning: SceneCOT

- > 3D LLMs need more rigorous evaluations and more high-quality training data.
- > Aligned VL representation and unleashing reasoning capability of LLMs are the most promising paths.
- > Post-training is underexplored yet promising towards a reliable and trustworthy model.