Revisiting 3D LLM Benchmarks: Are We Really Testing 3D Capabilities?

Anonymous ACL submission

Abstract

In this work, we identify the "2D-Cheating" problem in 3D LLM evaluation, where these tasks might be easily solved by VLMs with rendered images of point clouds, exposing ineffective evaluation of 3D LLMs' unique 3D capabilities. We test VLM performance across multiple 3D LLM benchmarks and, using this as a reference, propose principles for better assessing genuine 3D understanding. We also advocate explicitly separating 3D abilities from 1D or 2D aspects when evaluating 3D LLMs.[1]

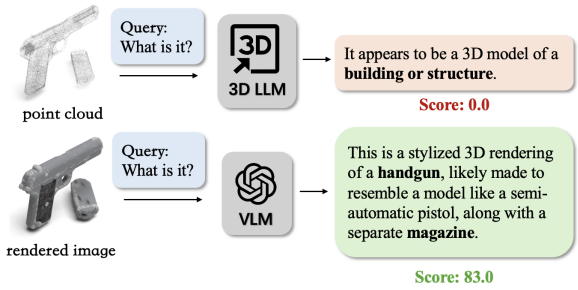


Figure 1: Example of 2D-Cheating. With rendered images of the point cloud, VLMs could easily solve some 3D tasks, and even outperform 3D LLMs.

1 Introduction

Recent advances in Large Language Models (OpenAI, 2022, 2023) have led to the development of Vision-Language Models (VLMs) (Liu et al., 2023; OpenAI, 2024). To overcome their lack of grounding in the real 3D physical world (Hong et al., 2023), researchers developed 3D LLMs (Qi et al., 2024; Xu et al., 2023) for 3D processing.

Given the extreme scarcity of 3D training data (Hong et al., 2023; Zhu et al., 2024), many approaches leverage existing LLMs and VLMs to generate annotations for 3D data (Hong et al., 2023; Qi et al., 2023; Xu et al., 2023). This reliance on 2D and language (1D) priors raises a fundamental question: *What capabilities do 3D LLMs possess that truly differentiate them from 2D VLMs?*

To explore this, we revisit some benchmarks used for 3D LLM evaluation. Current 3D LLMs are primarily evaluated on Q&A or captioning tasks (Ma et al., 2024), rather than specific downstream tasks like object detection (Caesar et al., 2020; Wang et al., 2021), as LLMs provide a general-purpose language interface (Hao et al., 2022). As shown in Figure 1, we find some tasks can be easily solved by 2D VLMs, indicating that they fail to effectively assess the unique capabilities of 3D models. We refer to this phenomenon as **2D-Cheating**.

Specifically, we render point clouds into images to test the performance of VLMs on 3D LLM benchmarks. Experimental results indicate that VLMs can significantly exceed the state-of-the-art performance of 3D models on certain benchmarks. However, on other benchmarks, even with enhanced information, they still fail to achieve the best performance.

We argue that tasks where VLMs perform poorly compared to 3D LLMs are those that truly involve unique 3D capabilities. Based on analysis, we (1) propose several principles for designing benchmarks that effectively evaluate true 3D capabilities, and (2) advocate for separating the evaluation of 3D capabilities from 2D and 1D aspects when assessing 3D LLMs.

2 Method

2.1 VLM3D

As shown in Figure 2, we propose **VLM3D**, a simple yet general pipeline that adapts VLMs to 3D tasks. Specifically, it first renders point clouds into images and augments queries with few-shot, then feeds them into a VLM.

[1]Code and data are available at anonymous github

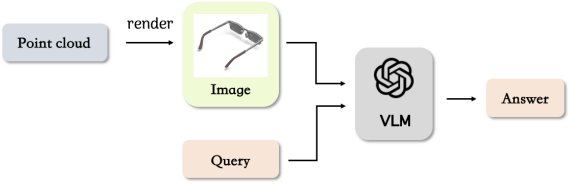


Figure 2: The VLM3D pipeline. For a given task, we render the point cloud into images and feed both the images and the query into a VLM.

2.2 Viewpoint Selection

Viewpoint of image rendering significantly impacts VLM's input information about the 3D asset. We hypothesize that a key limitation of 2D models in 3D understanding stems from the **viewpoint-specific nature of images**, whereas point clouds inherently provide a holistic 3D representation. The challenges posed by viewpoint dependency include: (1) blind spots in areas outside the selected viewpoint; (2) occlusion and overlap of objects; and (3) single-surface capture that lacks multifaceted geometry. Based on this, we set different viewpoint rendering configurations for object and scene point cloud benchmarks.

Object Object point clouds are relatively simple, so we use a fixed viewpoint for each benchmark, typically the front or side view of most objects.

Scene Scene point clouds present greater complexity. To systematically investigate viewpoint impacts, we conducted the following experimental settings, as all viewpoints illustrated in Figure 3.

2.2.1 Single View

In this setting, we follow the common practice (Ma et al., 2022) to render images from BEV.

2.2.2 Multi View

In this setting, we render images from four fixed viewpoints (East, South, West, and North) and combine them into a multi-view image.

2.2.3 Oracle View

In this setting, we conducted experiments on the validation set to explore the upper limits of VLM3D's capabilities. We first rendered images from all five viewpoints shown in Figure 3, and then used the **Best-of-N** (BoN) method to select the optimal view for each question.

Since LLMs are probabilistic, simply taking the highest evaluation score risks inflating the likelihood of randomly guessing the correct answer. To address this, we sampled multiple responses for each viewpoint and calculated the average score. The highest average score among all viewpoints was then selected as the oracle view score.

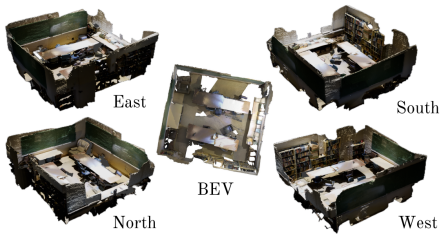


Figure 3: All five viewpoints of an example scene.

Benchmark	Metric	3D SOTA	VLM3D (GPT-4o)
3D MM-Vet	LLM-eval	43.2	**58.1** (+14.9)
ObjaverseXL-LVIS Caption	BLEU-1	32.2	**36.2** (+4.0)
	ROUGE-L	35.5	**36.8** (+1.3)
	CIDEr	78.0	**79.3** (+1.3)

Table 1: Results on object benchmarks. The 3D SOTA models are ShapeLLM-13B and GPT4Point, respectively.

2.3 Other Rendering Factors

While other factors like rendering style may affect image quality, viewpoint selection remains the trickiest due to the difficulty of a unified optimal setup. For each benchmark, we applied a consistent set of other rendering configurations across all point clouds to ensure high-quality results.

3 Experiments

We conducted experiments on the benchmarks in Figure 4, which covered major benchmarks used for evaluating 3D LLMs.

3.1 Object Point Cloud Benchmark

For object point cloud benchmarks, we use **3D MM-Vet** (Qi et al., 2024) for Q&A and **ObjaverseXL-LVIS Caption** (Deitke et al., 2022) for captioning, with results presented in Table 1.

3.2 Scene Point Cloud Benchmark

We conducted experiments on two widely used scene point cloud benchmarks, **ScanQA** (Azuma et al., 2021) and **SQA3D** (Ma et al., 2022). ScanQA evaluates general Q&A, while SQA3D further assesses situation understanding. We chose the 3D Baseline as the best-performing model at the time of the benchmark's release.

3.2.1 Single View Evaluation

We first conducted the single view evaluation mentioned in Section 2.2.1, with results in Table 2.

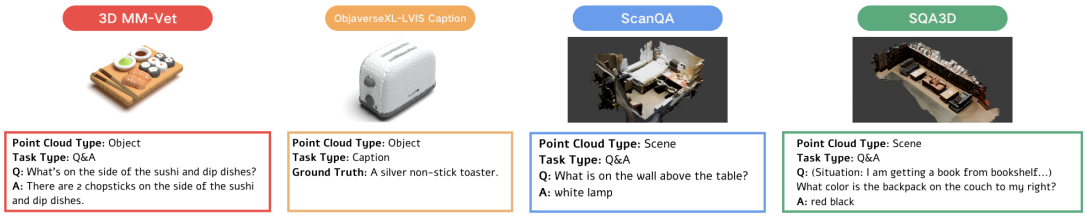


Figure 4: 3D LLM benchmarks tested in our experiments, see detailed introduction in Appendix A.

Benchmark	Metric	VLM3D (Qwen2-VL)	VLM3D (GPT-4o)	3D Baseline	3D SOTA
ScanQA(val)	METEOR	12.8 (-0.3)	14.7 (+1.6)	13.1	20.0 (+6.9)
	ROUGH	27.9 (-5.4)	25.8 (-7.5)	33.3	49.2 (+15.9)
	CIDEr	51.2 (-13.7)	47.2 (-17.7)	64.9	101.4 (+36.5)
SQA3D(test)	EM	42.2 (-5.0)	44.8 (-2.4)	47.2	52.6 (+5.4)

Table 2: Single-view results: For ScanQA, baseline is ScanQA, SOTA is LEO (Huang et al., 2023). For SQA3D, baseline is SQA3D, SOTA is SIG3D (Man et al., 2024). Red and green indicate decrease and increase relative to the 3D baseline, respectively.

We observed that these benchmarks often focus on specific objects within intricate scenes, with single view images frequently missing required information. To further explore the upper bound of VLMs, we conducted experiments using different viewpoints on Qwen2-VL-72B (Wang et al., 2024).

3.2.2 Multi View Evaluation

We conducted multi-view evaluation based on the settings in Section 2.2.2, with results in Table 3.

Benchmark	Metric	SV	MV	3D Baseline	3D SOTA
ScanQA(val)	METEOR	12.8 (-0.3)	13.1 (+0.0)	13.1	20.0 (+6.9)
	ROUGH	27.9 (-5.4)	29.1 (-4.2)	33.3	49.2 (+15.9)
	CIDEr	51.2 (-13.7)	54.0 (-10.9)	64.9	101.4 (+36.5)
SQA3D(test)	Overall	42.2 (-5.0)	43.2 (-4.0)	47.2	52.6 (+5.4)

Table 3: The result of multi-view. Baseline and SOTA models are the same in Table 2. SV refers to single-view, and MV refers to multi-view.

3.2.3 Oracle View Evaluation

We performed oracle view evaluation using the Best-of-N method in Section 2.2.3. For each question, we rendered images from k viewpoints and sampled n responses per viewpoint. As shown in Figure 5, the result decreases as n increases and converges around $n = 20$, indicating that randomness from selecting the maximum score is effectively mitigated. The scores are detailed in Table 4.

Besides the Best-of-N, we explored another method called **Human-Intuition-Selection** (HIS). We calculate the centroids of all relevant objects of each question and use a heuristic algorithm to select the best viewpoint that captures the centroids effectively. Details and results are in Appendix B.1.

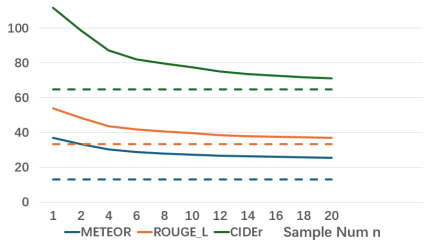


Figure 5: Result of BoN on ScanQA(val) with $k = 5$. The dashed line represents the 3D baseline ScanQA, and the solid line shows VLM3D's performance.

Metric	SV	BoN(k=4)	BoN(k=5)	3D Baseline	3D SOTA
METEOR	12.8 (-0.3)	27.0 (+13.9)	28.2 (+15.1)	13.1	20.0 (+6.9)
ROUGH	27.9 (-5.4)	35.6 (+2.3)	36.9 (+3.6)	33.3	49.2 (+15.9)
CIDEr	51.2 (-13.7)	68.4 (+3.5)	71.2 (+16.3)	64.9	101.4 (+36.5)

Table 4: Result of oracle view on ScanQA(val). the sample number of BoN is $n = 20$.

4 Analysis

Results reveal that VLM outperforms 3D LLMs using only single view images on these object benchmarks. This suggests that these tasks can be easily solved without specialized 3D representations, indicating a lack of effective 3D capability assessment. The limitation may arise from (1) ● the inherent simplicity of these object point clouds, and (2) ● these tasks primarily require only surface-level information that could be equally extracted from a simple rendered image. See more task examples in Appendix A.1 and Appendix A.2.

4.1 Scene Point Cloud

4.1.1 Poor Performance with Single View

In these scene benchmarks, even the most advanced VLMs consistently underperform top 3D models, ● revealing their limitations in handling complex 3D scenes.

Besides, ● many benchmarks rely on text similarity metrics for evaluation, which is overly rigid given the flexible nature of natural language. See case study in Appendix C.1.

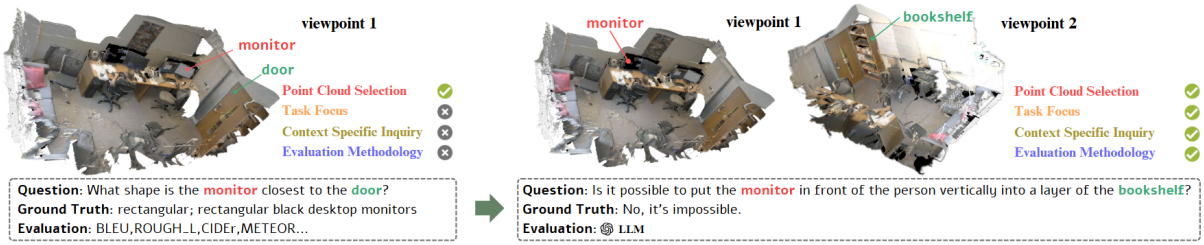


Figure 6: An example of task redesign, with the original task on the left and the new task on the right.

4.1.2 Challenges of Multi View Understanding

As shown in Table 3 & 4, providing multi-view images only brings a slight improvement. In contrast, oracle view leads to a significant boost. Notably, when setting $k = 4$, all candidate viewpoints of Best-of-N are included in the multi-view inputs, theoretically providing sufficient information. However, ● results indicate that VLMs struggle to form a unified understanding of a scene from multi-view.

4.1.3 Enhancement with Oracle View

As shown in Table 4, oracle view performance significantly surpasses the 3D baseline, confirming that providing good viewpoints greatly enhances VLM's potential to solve these tasks through 2D-Cheating. However, dynamically providing the oracle viewpoint for each question considerably simplifies the task by identifying and presenting the key information from the complete scene.

Besides, the performance of BoN still lags behind the 3D SOTA. This suggests that, ● even provided with a favorable view, VLMs still struggle to match the best performance of 3D LLMs in some tasks.

4.1.4 Failure of HIS Method

Table 5 shows that HIS performs significantly worse than Best-of-N. Through case studies, we identify three key factors: (1) ● Selecting the best viewpoint for detail items in complex scenes with intuition is challenging, often leading to occlusions. (2) ● Limited ground truth reveals a clear preference, as shown in Appendix C.2. While multiple viewpoints all provide sufficient information, BoN selects the one that aligns most closely with the ground truth. (3) ● Many questions rely on common sense, not the specific details of the point cloud, as shown in Appendix C.3. Specific information in HIS viewpoints might even lead to a worse score, compared to answering leveraging world knowledge with an unrelated viewpoint in BoN.

5 Principles For Effective 3D Evaluation

Based on the analysis, we propose the following principles for effective 3D capabilities evaluation of 3D LLMs.[2] We applied these principles to redesign a task, as shown in Figure 6.

● **Point Cloud Selection** A complex point cloud should be selected, such as a point cloud of scenes or objects with more intricate structures.

● **Task Focus** Tasks should go beyond general surface-level information, diving into the intricate details of 3D assets. This includes the structural specifics of individual objects and the finer items within complex scenes. Furthermore, questions should prioritize aspects that cannot be easily answered with images from any viewpoints.

● **Context Specific Inquiry** Avoid overly general questions, ask the unique aspects of the current 3D asset, and try to "violate common sense" to test genuine understanding of 3D input.

● **Evaluation Methodology** For problems with multiple reasonable answers, include various possible answers to accommodate more reasonable results and varying detail levels. Besides, avoid text similarity metrics and instead rely on LLMs for more flexible assessments.

6 Conclusion

This paper identifies the 2D-Cheating problem in current 3D LLM evaluation, compares VLM performance with 3D LLMs across various benchmarks, and provides principles for assessing real 3D capabilities in 3D LLMs. However, this does not imply that 3D LLMs can ignore the ability to solve tasks vulnerable to 2D-Cheating. Therefore, we emphasize the need to deliberately decouple the evaluation of 3D capabilities from 1D and 2D capabilities to ensure that core 3D evaluations are not confused or overlooked.

[2]The analyses leading to each principle are marked with corresponding colored circular markers in the preceding text.

7 Limitations

Our evaluation may not fully capture the capabilities of VLMs. First, the 3D LLMs were trained on the benchmark's training set or similar tasks, whereas we only adapted VLMs to these benchmarks through few-shot prompting. Additionally, rendering point clouds into images inherently loses information compared to using real photographs. These factors could limit the assessment of VLMs' real capability and may have weakened the observed 2D-Cheating issue.

Furthermore, our experiments were conducted solely on the Qwen2-VL and GPT-4o, which may not fully represent the performance of all available VLMs. The number of benchmarks we tested was limited, meaning the results might not generalize across a broader set of tasks. Ideally, a new benchmark that aligns with these principles would be developed, but due to constraints in space, resources, and other practical considerations, we leave this for future work.

8 Ethics Statement

This research on 3D LLM benchmarks adheres to ethical principles by exclusively utilizing publicly available data and maintaining transparency through open-source code. Our study does not involve any human annotation or privacy concerns. We do not anticipate any potential risks from this methodological study.

References

Daich Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. 2021. Scanqa: 3d question answering for spatial scene understanding. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19107–19117.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. *Preprint*, arXiv:1702.04405.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. 2022. Objaverse: A Universe of Annotated 3D Objects. *arXiv e-prints*, arXiv:2212.08051.

Agrim Gupta, Piotr Dollár, and Ross Girshick. 2019. LVIS: A Dataset for Large Vocabulary Instance Segmentation. *arXiv e-prints*, arXiv:1908.03195.

Yaru Hao, Haoyu Song, Li Dong, Shaohan Huang, Zewen Chi, Wenhui Wang, Shuming Ma, and Furu Wei. 2022. Language models are general-purpose interfaces. *arXiv preprint arXiv:2206.06336*.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 2023. 3D-LLM: Injecting the 3D World into Large Language Models. *arXiv e-prints*, arXiv:2307.12981.

Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. 2023. An embodied generalist agent in 3d world. *ArXiv*, abs/2311.12871.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.

Xianzheng Ma, Yash Bhalgat, Brandon Smart, Shuai Chen, Xinghui Li, Jian Ding, Jindong Gu, Dave Zhenyu Chen, Songyou Peng, Jia-Wang Bian, et al. 2024. When llms step into the 3d world: A survey and meta-analysis of 3d tasks via multimodal large language models. *arXiv preprint arXiv:2405.10255*.

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. 2022. SQA3D: Situated Question Answering in 3D Scenes. *arXiv e-prints*, arXiv:2210.07474.

Yunze Man, Liang-Yan Gui, and Yu-Xiong Wang. 2024. Situational Awareness Matters in 3D Vision Language Reasoning. *arXiv e-prints*, arXiv:2406.07544.

OpenAI. 2022. Introducing chatgpt. openai.com.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI. 2024. Hello gpt-4o. openai.com.

Zekun Qi, Runpei Dong, Shaochen Zhang, Haoran Geng, Chunrui Han, Zheng Ge, Li Yi, and Kaisheng Ma. 2024. ShapeLLM: Universal 3D Object Understanding for Embodied Interaction. *arXiv e-prints*, arXiv:2402.17766.

Zhangyang Qi, Ye Fang, Zeyi Sun, Xiaoyang Wu, Tong Wu, Jiaqi Wang, Dahua Lin, and Hengshuang Zhao. 2023. GPT4Point: A Unified Framework for Point-Language Understanding and Generation. *arXiv e-prints*, arXiv:2312.02980.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang

Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. 2021. Fcos3d: Fully convolutional one-stage monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 913–922.

Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. 2023. PointLLM: Empowering Large Language Models to Understand Point Clouds. *arXiv e-prints*, arXiv:2308.16911.

Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. 2024. LLaVA-3D: A Simple yet Effective Pathway to Empowering LMMs with 3D-awareness. *arXiv e-prints*, arXiv:2409.18125.

A Benchmarks in Experiments

To the best of our knowledge, given the nascent development of 3D LLMs, there is currently no dedicated and representative benchmark for their evaluation. We select representative benchmarks for 3D LLM evaluation: 3D MM-Vet and ObjaverseXL-LVIS Caption were proposed by model creators (Qi et al., 2024, 2023), while ScanQA and SQA3D are previously established 3D question-answering benchmarks (Azuma et al., 2021; Ma et al., 2022) that have been widely adopted in 3D LLM assessments.

A.1 3D MM-Vet

3D MM-Vet (Qi et al., 2024) is a benchmark designed to evaluate multi-modal large language models in the context of 3D comprehension, particularly for embodied scenarios. The benchmark includes the following categories: **General Recognition**, **Knowledge**, **Language Generation**, **Spatial Awareness**, and **Embodied Interaction**.

Examples of each task type are shown below, using the same point cloud input depicted in Figure 7.

Figure 7: Example rendered image of a point cloud in 3D MM-Vet

- **General Visual Recognition**:
  Question: What is this?
  Answer: There's a board, on which are different kinds of sushi, one dish of Soy sauce, one dish of wasabi and chopsticks.

- **Knowledge**:
  Question: How many sushi will there be if 2 more sushi are added?
  Answer: There will be 8 sushi if 2 are added.

- **Language Generation**:
  Question: Describe its cuisine culture in detail.
  Answer: Sushi is a kind of Japanese cuisine, the main material is vinegar, sugar, salt seasoning, fat, and slightly sweet Japanese pearl rice cooked into vinegar rice, after cooling with fish, seafood, vegetables, eggs, or other meat and other ingredients.

- **Spatial Recognition**:
  Question: Which sushi is closer to the chopsticks, The white sushi or the black sushi?
  Answer: The black sushi is closer to the chopsticks.

- **Embodied Interaction**:
  Question: Give me steps to feed me eat this dish.
  Answer: Step 1: Pick up chopsticks. Step 2: Pick up the sushi with chopsticks. Step 3: Dip in soy sauce and mustard. Step 4: Move the chopsticks to get the sushi into your mouth.

A.2 ObjaverseXL-LVIS Caption

GPT4Point (Qi et al., 2023) utilizes the Objaverse dataset, aligning it with LVIS categories (Gupta et al., 2019), to create the Objaverse-LVIS dataset. In this dataset, scenes featuring complex environments such as indoor houses or outdoor parks are excluded, focusing instead on scenarios involving individual objects or combinations of multiple objects.

The ObjaverseXL-LVIS benchmark consists of two tasks: **3D object point cloud captioning** and **3D point cloud question answering**. Since only the captioning portion of the dataset is open-source, our experiments were conducted solely on this part. Figure 8 and 9 show two examples of captions.

Figure 8: Caption: A red and blue gradient teapot with three spouts.

Figure 9: Caption: A Mexican food cart decorated with traditional Mexican cuisine, topped with a sombrero hat.

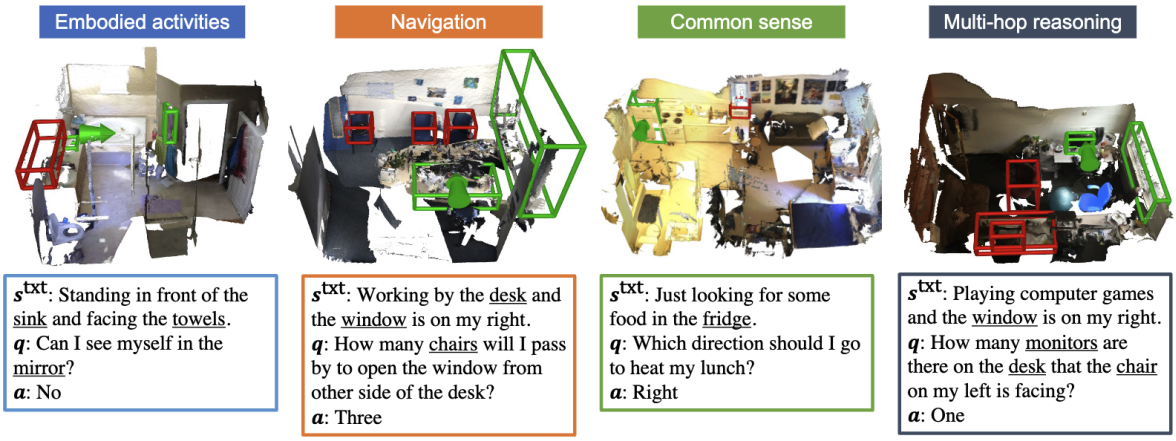


Figure 10: Examples from SQA3D. *Image from Ma et al., "SQA3D: Situated Question Answering in 3D Scenes", ICLR 2023, used under CC BY 4.0 license.*

A.3 ScanQA

ScanQA (Azuma et al., 2021) was created for the 3D-QA task using RGB-D scans of indoor scenes and annotations from the ScanNet dataset. The dataset includes both question-answer pairs and 3D object localization annotations, making it one of the largest datasets for specifying object properties in 3D scenes via question-answering. The benchmark includes the following task categories: **Place**, **Number**, **Color**, **Object nature**, **Object** and **Other**.

Examples of each task type are shown below, using the same point cloud input depicted in Figure 11.

Figure 11: Example rendered image of a point cloud of the scene in ScanQA

- **Place**:
  Question: Where is the couch located?
  Answer: behind tables/ near small table/ against wall/ behind small round tables

- **Number**:
  Question: How many wooden tables are on either side of the blue couch?
  Answer: 3

- **Color**:
  Question: What color is the couch?
  Answer: blue/ dark blue

- **Object nature**:
  Question: What type of chair is next to the round brown table?
  Answer: dark blue couch / red ottoman chair
  Question: What kind of table is right behind a long black chair?
  Answer: round table / light brown round table
  Question: What shape is the wooden table?
  Answer: round/ circular round shape

- **Object**:
  Question: What is the red squared chair under?
  Answer: table/ round table in middle

A.4 SQA3D

SQA3D (Ma et al., 2022) is a benchmark that tests scene understanding of embodied agents. Given a 3D scene and a description of its position, the agent must first understand its situation and then reason and answer the question. The benchmark examines a wide spectrum of reasoning capabilities for an intelligent agent, ranging from spatial relation comprehension to commonsense understanding, navigation, and multi-hop reasoning, with examples in Figure 10.

B Human-Intuition-Selection Method For Oracle View Evaluation

B.1 Method Introduction

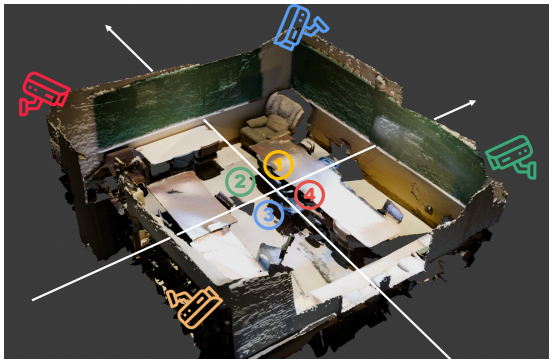


Figure 12: Illustration of the Human-Intuition-Selection Oracle Viewpoint Method: When the centroid of all relevant objects falls within a colored region, the viewpoint corresponding to the same color camera is selected as the oracle view.

ScanQA provides the `object_id` in ScanNet (Dai et al., 2017) for all relevant objects of each question. Based on this information, the HIS method retrieves the bounding box and center coordinates for each relevant object and calculates the centroid of all relevant objects. As shown in Figure 12, the scene is divided into four regions based on the camera angles (East, South, West, and North) for viewpoint number $k = 4$. When the overall centroid falls within a specific region, the opposite viewpoint is selected as the oracle view. For viewpoint number $k = 5$, we introduce an additional central region, and if the centroid falls within this area, the BEV viewpoint is chosen as the oracle view.

B.2 Experimental Result

Metric	Single View	Oracle View (HIS)	Oracle View (BoN)
METEOR	12.8	13.2	30.8
ROUGH	27.9	29.1	39.7
CIDEr	51.2	53.9	77.7

Table 5: Results of Different Oracle View Methods on ScanQA(val). HIS refers to Human-Intuition-Selection.

C Case Study

C.1 Limitations of Text Similarity Metrics

Using text similarity metrics for evaluation is too strict and often struggles to fully capture the nuances of model responses due to the openness and flexibility of natural language. Following are some examples from SQA3D and ScanQA, which use different text similarity metrics for evaluation:

- Task1 (From SQA3D)
  - Question: "*Is the table in front or at my back?*"
  - Ground truth: "*back*"
  - VLM3D's answer: "*at my back*"
  - Score: 0 (EM)
- Task2 (From ScanQA)
  - Question: "*What shape is the window behind the table?*"
  - Ground truth: "*rectangular window*", "*rectangular*"
  - VLM3D's answer: "*rectangle*"
  - Score: 0 (BLUE_1), 0 (METEOR), 0 (CIDEr), 0 (ROUGE_L)

Besides, benchmarks like ObjaverseXL-LVIS and Scan2Cap also employ text similarity metrics that rigidly evaluate syntactic patterns rather than semantic correctness. Such metrics fail to accommodate the open-domain nature of LLM responses, particularly for VLMs untrained on task-specific datasets to align with prescribed response styles.

C.2 Ground Truth Preference

C.2.1 3D MM-Vet

Figure 13 presents an example from 3D MM-Vet, where the task is to determine if a rusty barrel is perfectly round. The correct answer is "No", as the barrel is not perfectly round. Both ShapeLLM and GPT-4o reached similar conclusions, with ShapeLLM providing a brief answer that aligns more closely with the ground truth. However, VLM3D(GPT-4o) offered a more detailed explanation, resulting in a significantly lower score. It's worth noting that this benchmark already uses large models to evaluate the answers.

C.2.2 ScanQA

Here is an example where the Best-of-N method achieved a significantly better score due to its alignment with ground truth preference: the question

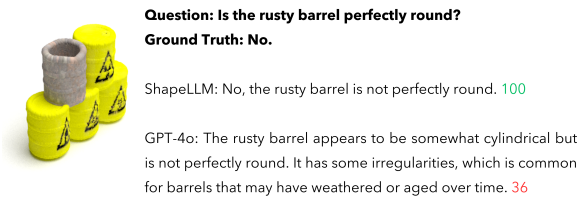
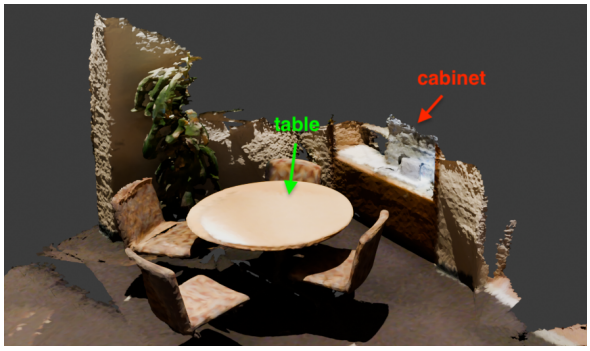


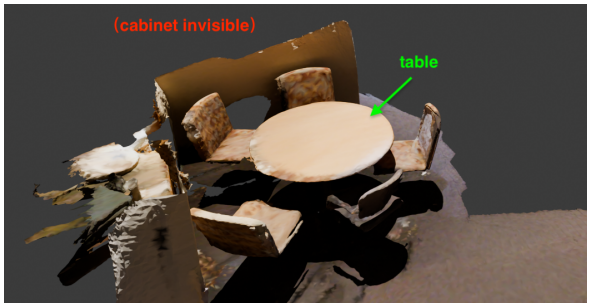
Figure 13: A case of ground truth preference from 3D MM-Vet.

is *"Where is the brown chair located?"* and the ground truths are *"under the circular table"* and *"next to the left table"*. The answer provided by the Best-of-N method is *"on the right side of the table"*, based on the input image shown in Figure 14 (a). The Human-Intuition-Selection method gives the answer *"on the right side of the table"*, using the input image shown in Figure 14 (b).

It is evident that there are many reasonable answers to this question; however, the ground truth prefers responses related to the **table**. While both viewpoints depict the chair, the answer from the Human-Intuition-Selection method is influenced by the nearby **cabinet** in its chosen viewpoint, resulting in a lower score.



(a) The viewpoint selected by Best-of-N.

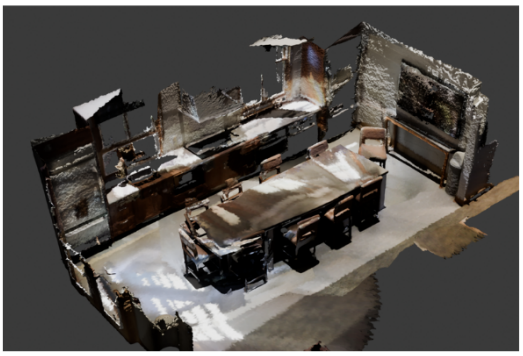


(b) The viewpoint selected by Human-Intuition-Selection.

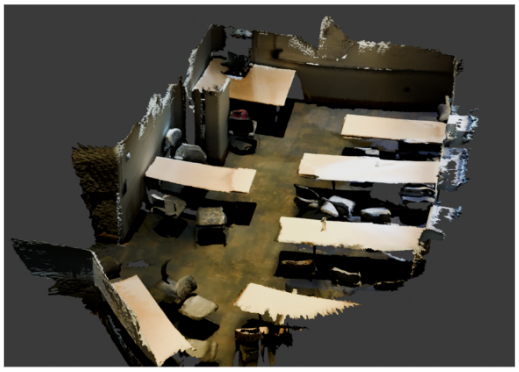
Figure 14: Images used for each oracle view method. To better present the details, some blank areas in the images have been cropped.

C.3 General Knowledge Over Specific Details

Figure 15 presents several examples from the case study. It is evident that the questions can be resolved through general world knowledge, without relying on the unique characteristics of the scenes involved.



Question:
"What shape is the refrigerator on the left of a kitchen counter?"
Ground Truth:
"rectangular"
"tall rectangular shape"



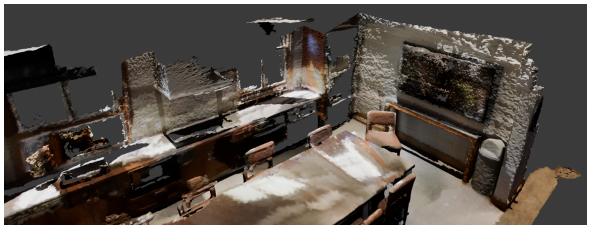
Question:
"What type of table has a wood look?"
Ground Truth:
"wood table"
"long and brown tables"

Figure 15: Example tasks from ScanQA.

Besides, here is an example where the Best-of-N method answers the question using an irrelevant viewpoint and still achieves better results. The reason is that the question is too general and can be answered solely based on world knowledge; however, the Human-Intuition-Selection method provides a more specific answer. The question is *"What is under the long kitchen counter?"* and the ground truths are *"brown rectangular kitchen cabinets"* and *"brown kitchen cabinets"*. The answer

provided by the Best-of-N method is *"kitchen cabinet"*, based on the input image shown in Figure 16 (a). The Human-Intuition-Selection method gives the answer *"black dishwasher"*, using the input image shown in Figure 16 (b).

(a) The viewpoint selected by Best-of-N.



(b) The viewpoint selected by Human-Intuition-Selection.

Figure 16: Images used for each oracle view method. To better present the details, some blank areas in the images have been cropped.

In the image used by the Best-of-N method, neither the kitchen counter nor kitchen cabinets are explicitly shown. However, in a typical kitchen layout, the area beneath the kitchen counter is commonly referred to as 'kitchen cabinets'. This can be answered using general world knowledge, which allows the method to successfully address the vague question. In contrast, the Human-Intuition-Selection method employs a viewpoint that clearly presents the objects related to the question, providing a more specific answer: 'black dishwasher'. However, the question is intended to assess the general area beneath the counter as a whole, which results in this more specific answer receiving a lower score.

D Experimental Details

D.1 Model Size

We use Qwen2-VL with 72B parameters in all experiments.

D.2 Rendering

For the two object point cloud benchmarks, we used the Mitsuba renderer due to its efficiency and lightweight nature. For the two scene point cloud benchmarks, which utilize mesh-based point clouds, we used the Blender rendering engine, consistent with the methodology used in the original work (Azuma et al., 2021; Ma et al., 2022).

We optimized the rendered images by adjusting key parameters such as lighting configurations and point cloud particle size. For more specific details, please refer to our code repository.

D.3 Evaluation

We replaced the GPT-3.5 model originally used for evaluation in the 3D-MM-Vet benchmark with GPT-4o-mini.