3DBench: A Scalable 3D Benchmark and Instruction-Tuning Dataset

Junjie Zhang[1] , **Tianci Hu**[1] , **Xiaoshui Huang**[2*] , **Yongshun Gong**[3] and **Dan Zeng**[1]

[1]Shanghai University
[2]Shanghai AI Laboratory
[3]Shandong University
huangxiaoshui@pjlab.org.cn

Abstract

Evaluating the performance of Multi-modal Large Language Models (MLLMs), integrating both point cloud and language, presents significant challenges. The lack of a comprehensive assessment hampers determining whether these models truly represent advancements, thereby impeding further progress in the field. Current evaluations heavily rely on classification and caption tasks, falling short in providing a thorough assessment of MLLMs. A pressing need exists for a more sophisticated evaluation method capable of thoroughly analyzing the spatial understanding and expressive capabilities of these models. To address these issues, we introduce a scalable 3D benchmark, accompanied by a large-scale instruction-tuning dataset known as 3DBench, providing an extensible platform for a comprehensive evaluation of MLLMs. Specifically, we establish the benchmark that spans a wide range of spatial and semantic scales, from object-level to scene-level, addressing both perception and planning tasks. Furthermore, we present a rigorous pipeline for automatically constructing scalable 3D instruction-tuning datasets, covering 10 diverse multi-modal tasks with more than 0.23 million QA pairs generated in total. Thorough experiments evaluating trending MLLMs, comparisons against existing datasets, and variations of training protocols demonstrate the superiority of 3DBench, offering valuable insights into current limitations and potential research directions. Codes are available at https://github.com/Inshsang/3DBench.

1 Introduction

Recently, there have been significant advancements in multi-modal large language models (MLLMs) [Li *et al.*, 2023; Zhu *et al.*, 2023; Radford *et al.*, 2021], catalyzing a profound revolution across various tasks. Large-scale instruction-tuning data is essential to harness the capabilities of MLLMs. To facilitate the integration of large models into the 3D domain, our goal is to establish a scalable evaluation benchmark
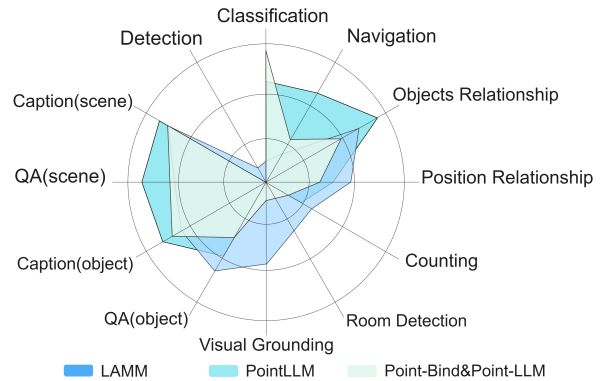

Figure 1: Zero-shot evaluation of three state-of-the-art 3D-LLMs on proposed 3DBench with ten multi-modal tasks.

specifically designed for assessing 3D-LLMs. Additionally, we elaborate on the detailed development of a large-scale dataset to address the scarcity of instruction-tuning datasets in the 3D domain.

Existing instruction-tuning datasets for 3D-LLMs originate from publicly open datasets. PointLLM [Xu *et al.*, 2023] utilizes Cap3D [Luo *et al.*, 2023], a 3D object captioning dataset derived from Objaverse [Deitke *et al.*, 2023]. Point-Bind & Point-LLM [Guo *et al.*, 2023] is trained using Ulip [Xue *et al.*, 2023], constructed based on ShapeNet [Chang *et al.*, 2015]. LlaVa [Wang *et al.*, 2023] efficiently incorporates the image modality into large models by prompting GPT-4 [Fu *et al.*, 2022] with captions from the COCO dataset [Lin *et al.*, 2014]. They also conduct experiments to validate the effectiveness of using GPT-generated [OpenAI, 2023] image-text pairs as training data.

However, existing **benchmarking** efforts primarily focus on single-object classification and captioning, neglecting an effective evaluation of the spatial understanding capabilities of large models for complex and scene-oriented point clouds. Benchmarks for point cloud scenes, such as Visual Grounding (VG), Detection, and VQA tasks outlined by LAMM [Yin *et al.*, 2023], can assess a model's interpretation of object positions in the given scene and its general knowledge inferred from language model. However, these tasks still do not sufficiently address the comprehension of multi-object relationships and the room-area perception capabilities, commonly

---
*Corresponding author

referred as the scene reasoning ability of large models.

Additionally, high-quality **instruction-tuning datasets** in the 3D domain remain limited [Wei *et al.*, 2021]. Current datasets are mainly collected to introduce large models and are generated based on open 3D datasets. This approach introduces a potential risk of data leakage in pre-training models [Dai *et al.*, 2024]. Significant distinctiveness in task categories exists among datasets, with few benchmarks for each one and a lack of uniformity, making it challenging to accurately and comprehensively assess the capabilities of 3D large models.

To address above issues, We propose a scalable benchmark comprising ten multi-modal tasks and three evaluation metrics. As shown in Fig. 2, the tasks include common ones such as classification, VG, detection, and counting. Building on this foundation, we extend VG to include a room detection task. Additionally, we expand object understanding from 2D to 3D, requiring reasoning about the relative positions and attribute relationships of multiple objects. To evaluate the quality of large model dialogues and long-text generation capabilities, we include QA and caption tasks. Finally, we design a navigation benchmark to evaluate the spatial planning capabilities of 3D-LLMs. Evaluation metrics consist of three types: (1) For tasks with textual outputs, like relationship reasoning, we adopt a heuristic approach to instruct ChatGPT to score the predictions. (2) Traditional metrics such as precision and mAP are employed to evaluate detection and VG tasks. Furthermore, we introduce a path loss to assess navigation task. (3) For questions with simple-structured answers like classification and counting, we generate 300 to 2000 multiple-choice questions [Liu *et al.*, 2023] and calculate the accuracy of output options.

The process of acquiring a large-scale instruction-tuning dataset consists of two main steps. During the initial step, we extract comprehensive metadata from the Procthor simulation framework [Deitke *et al.*, 2022]. This metadata comprises depth maps and corresponding diverse types of ground-truth for both objects and scenes. Utilizing the depth maps, we manage to reconstruct a variety of 3D objects and scenes. In the second step, we utilize the ground-truth to inspire GPT to generate knowledge for textual tasks. By incorporating the ground-truth into diverse dialogue templates, we obtain a instruction-tuning dataset for various fine-grained tasks. In summary, our contributions can be summarized in three main aspects:

- **Evaluation Benchmark:** We develop a benchmark that spans from the object to the scene scale, covering dimensions of perceptual and reasoning abilities. It includes ten diverse multi-modal tasks, assessed with three types of customized evaluation metrics. The benchmark is enhanced and extended beyond existing standards through the incorporation of our proposed tasks and metrics.

- **Instruction-tuning Dataset:** We design an approach for the automatic acquisition of a large-scale 3D instruction-tuning dataset, resulting in 34,000 point clouds of everyday objects and 30,000 indoor scenes containing them. The dataset comprises over 0.23 million QA pairs, encompassing ten fine-grained tasks.
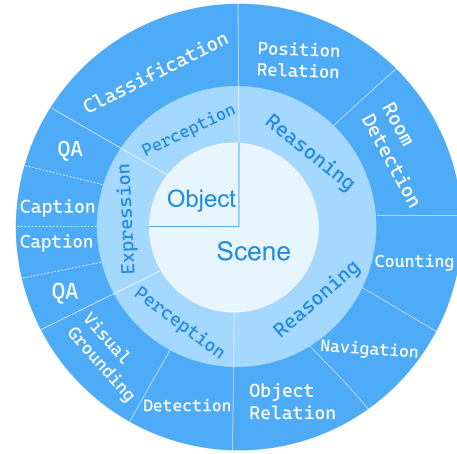

Figure 2: The current task overview of 3DBench. 3DBench comprehensively addresses the complexity of both spatial and logical aspects, categorizing ten individual tasks into three levels. Future tasks can seamlessly integrate into this framework.

- **Experiments and Observations:** Experimental validation substantiates the efficacy of our proposed dataset. As shown in Fig. 1, we conduct a thorough quantitative evaluation of existing 3D-LLMs using 3DBench. Observations cover the quality of the generated dataset and variations in training protocols, offering insights for future explorations.

2 Related Work

The proposed 3DBench comprises two main contributions: benchmark and dataset. To assess the distinctiveness of existing work and our contributions, we review related works from two perspectives.

2.1 Multi-modal Evaluation Benchmark

Several existing benchmarks evaluate MLLMs across a range of tasks [Hao *et al.*, 2022]. MMBench [Liu *et al.*, 2023], with a focus on perception and reasoning, designs 20 fine-grained tasks that cover the three-tier capacity dimensions of MLLMs. MME [Fu *et al.*, 2023] curates all instruction-answer pairs manually to prevent data leakage and utilizes 14 sub-tasks to assess the perceptual and cognitive capabilities of VLLMs. LAMM [Yin *et al.*, 2023] extends research to incorporate point clouds, creating three visual task benchmarks for 3D tasks to facilitate the assessment of scene-level perceptual abilities. To assess the text quality of large model outputs, AlpacaFarm [Dubois *et al.*, 2023] introduces a swift and reliable GPT-4-based automated benchmark. Additionally, traditional metrics such as Bleu [Papineni *et al.*, 2001], ROUGE [Lin, 2004], and METEOR [Banerjee and Lavie, 2005] produce ineffective evaluation results due to answer bias, unlike GPT. Although current benchmarks have made noteworthy contributions to evaluating MLLMs, there remains a deficiency in comprehensive quantitative assessment for 3D-LLMs.
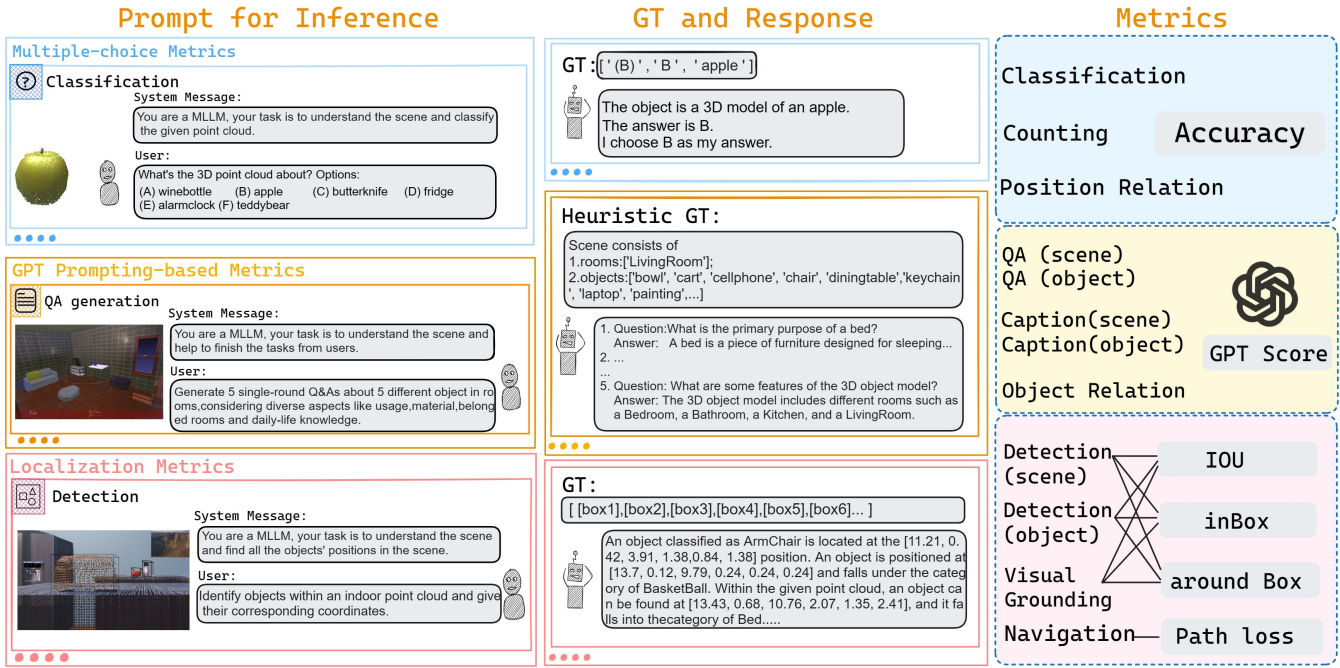
Figure 3: Overview of the 3DBench benchmark, encompassing ten 3D computer vision tasks and metrics from three perspectives, including traditional accuracy, IOU metric, GPT scores, and the novel path loss metric introduced by us.

2.2 Instruction-tuning Dataset

Instruction-tuning datasets consist of paired samples with instructions and their corresponding answers. Public multimodal datasets such as 2D COCO Caption and 3D ShapeNet offer objects and respective captions. LlaVa prompts GPT to generate detailed image descriptions with concise subtitles and object positions. Subsequently, it introduces new modalities into LLMs through textual information associated with the images. LAMM goes a step further by embedding ground truth from traditional computer vision tasks into dialogue templates, incorporating rich metadata from the dataset into instruction fine-tuning data to enhance the MLLM's ability to handle diverse tasks. In the 3D domain, ULIP-2[Xue *et al.*, 2023] and Cap3D prompt MLLMs to automatically generate captions for images rendered from point clouds. Point-LLM[Guo *et al.*, 2023] uses prompts from ChatCaptioner to generate high-quality Objaverse descriptions, while PointLLM[Xu *et al.*, 2023] leverages captions generated by Cap3D to prompt GPT for more detailed point cloud descriptions in a conversational format. However, existing 3D instruction-tuning datasets are relatively scarce, and their sources are often limited to publicly available datasets. The availability of high-quality instruction-tuning datasets for 3D tasks is indeed a critical need.

3 3DBench

3.1 Overview

3DBench, as illustrated in Fig. 2 and 3, stands out from existing 3D benchmarks for two key reasons. Firstly, 3DBench includes ten diverse tasks to thoroughly assess the capabilities of 3D models in addressing aforementioned challenges. Secondly, 3DBench is built upon a unified instruction-tuning dataset that is extensive and hierarchically rich. It is grounded in human cognition of real-world tasks, illustrating a three-tiered assessment. The top-level dimension, denoted as L-1, signifies the most apparent object-scene scale perception dimension. Beyond this, we consider the characteristics of large models and the logical reasoning process, allowing us to categorize tasks into three distinct categories: perception, reasoning, and expression. This introduces the L-2 competency dimension, covering three dimensions: 1) single or multiple-instance perception; 2) multi-instance relational reasoning; 3) conversational and descriptive abilities. Additionally, we derive the L-3 tasks dimension from the L-2 competency dimension. The benchmark includes ten evaluation tasks and three types of evaluation metrics.

3.2 Evaluation Tasks

Expansion Tasks. We systematically compile tasks from the 3D domain and language models. After excluding tasks beyond the capability of existing large models, we choose four tasks for our evaluation: classification, VG, detection, and counting. Building on prior works, we extend the scope of these existing tasks. Additionally, we expand the detection task to encompass the scene scale, instructing the model to predict bounding boxes for all rooms to assess the large model's overall perception capability in complex 3D scenes. Recognizing the expressive capacity of 3D-LLMs, we also introduce two textual generation tasks: generating detailed descriptions for long texts and generating dialogues for chat.

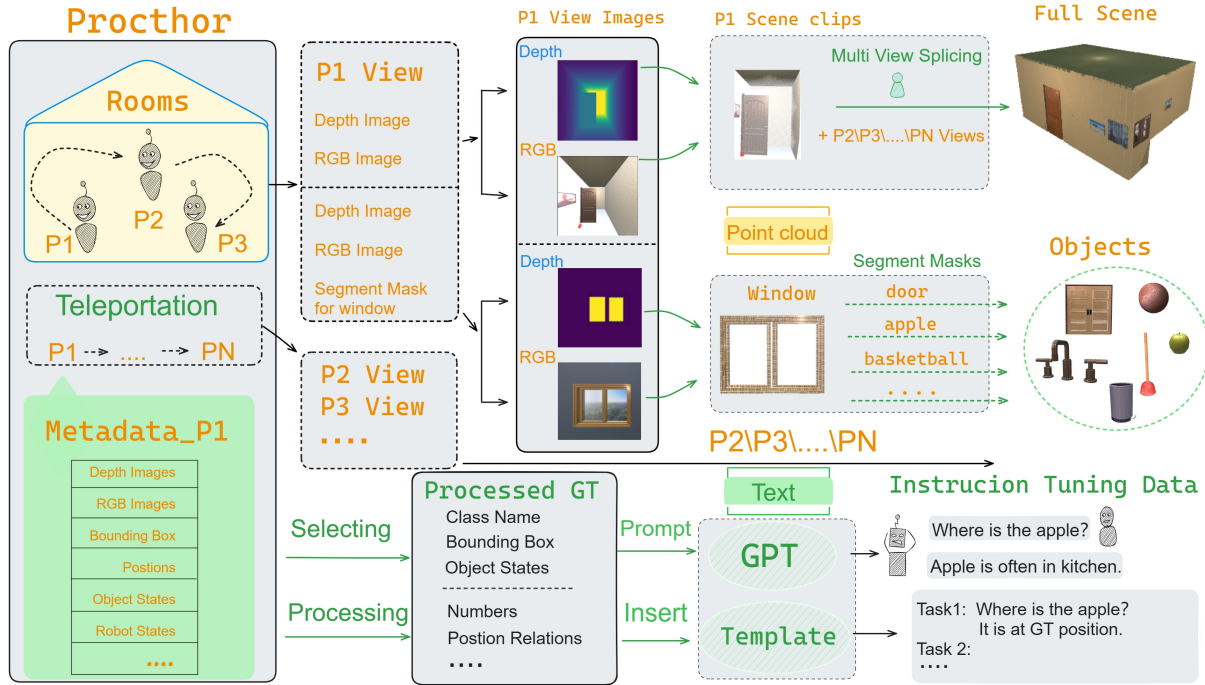**Novel Tasks.** Drawing inspiration from 2D multi-instance

Figure 4: Pipeline 3for generating the dataset. We are able to automatically collect instruction-tuning data for all detailed tasks in 3DBench. Each data sample comprises a point cloud (scene or object) along with the corresponding task dialogue.

tasks, we introduce tasks that require considering relationships and positional reasoning among multiple objects in real-world scenes, specifically, the multi-instance task within a given scene. Additionally, anticipating the capabilities of future large models in scene navigation and planning, we establish a navigation benchmark to assess the intricate scene perception and path planning abilities of 3D models.

3.3 Evaluation Metrics

GPT Prompting-based Metrics. As depicted in Fig. 3, we present a heuristic prompt tailored to address text generation tasks spanning different scales. When utilizing LLMs as the text evaluator, we provide GPT with information about all actual objects in the scene to avoid inflated scores caused by illusions. Moreover, this approach enables GPT to assess the authenticity of subtitles without depending solely on manual annotations. Furthermore, manual annotations may not necessarily represent the only correct solution for text generation. For evaluating text generation quality aligned with human habits, we recommend employing GPT for increased robustness.

Localization Metrics. We replace the conventional IOU metric, known for yielding low scores in detection and visual grounding (VG) tasks, with an innovative approach. We evaluate whether the predicted object's center falls within the bounding box, resulting in the 'in box' metric. Furthermore, we relax the criteria, permitting the predicted center to be within a one-meter radius of the ground-truth center, leading to the 'around box' metric. Additionally, we propose a novel path loss metric for the navigation task as shown in Fig. 5. It selects the longer trajectory between the prediction and
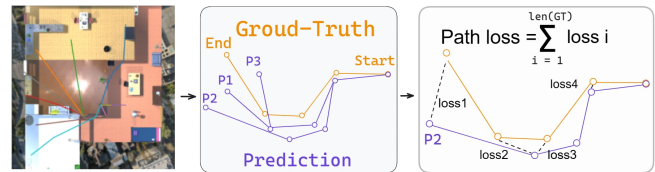

Figure 5: The illustration of path loss. It is the distance accumulation of each endpoint on the longer trajectory (between GT and prediction) and its nearest neighbors on the other one.

ground-truth, measuring the distance between each endpoint of this trajectory and its nearest neighbor from the other one. The accumulated distance is compared against a pre-defined threshold to determine the success of the navigation.

Multiple-choice Metrics. We generate numerous multiple-choice questions for tasks with straightforward ground truth, encompassing classification, counting, and spatial relationships. To improve the ability of weaker 3D-LLMs to choose the correct answer, we direct all models to consider the correct answer as the only option. If a question is difficult to answer, we recommend all models to generate a random answer to avoid nonsensical responses.

3.4 Dataset Construction

Pipeline. The objective is to create large-scale 3D instruction-tuning datasets. To accomplish this, we design a scalable data construction pipeline. The pipeline comprises two main steps, as shown in Figure 4. In the initial step, we collect data for 30,000 houses using Procthor[Deitke *et al.*, 2022] and extract depth images and metadata ground-truth

Task	Train	Test	All
Detection	30k	500	30k
QA(scene)	400	100	0.5k
Caption(scene)	400	100	0.5k
Classification	33k	1k	34k
Visual Grounding	60k	1k	61k
Counting	15k	500	15.5k
Room Detection	30k	500	30.5k
Position Relationship	5k	500	5.5k
Objects Relationship	5k	500	5.5k
Navigation	45k	6.7k	52k
Ten tasks	**224k**	**11k**	**235k**

Table 1: Statistics on the distribution of 3DBench Dataset.

Group Idx	Model	Training Split	Test Split	Task
E1	LAMM	ShapeNet & 3RScan	3DBench	Detection ; VG ; VQA
E2	LAMM	3DBench	ScanNet	Detection ; VG ; VQA
E3	LAMM	Variations of 3DBench	3DBench	Detection ; VG ; Classification
E4	LAMM	3DBench	3DBench	Ten Tasks
E5	All Three	None	3DBench	Ten Tasks

Table 2: Details of five groups of experiment settings.

- **E1 & E2:** We initially assess the zero-shot performance of the LAMM model on the 3DBench test split (referred to as E1). Subsequently, employing the LAMM training framework, we exclusively train a model on our training split and evaluate its zero-shot performance on LAMM-Bench (noted as E2). The cross-validation results of three LAMM tasks (detection, VG, and VQA) for E1 & E2 are compared to analyze differences between our dataset and publicly available ones.

- **E3:** We systematically categorize the 3DBench dataset into various scales and employ the same framework to train multiple LAMM models. The objective is to investigate how the expanded dataset scale influences the enhancement of LLM performance across three tasks: detection, VG, and classification.

- **E4:** We conduct a re-training of the LAMM model using the complete version of our 3DBench. Subsequently, we compare the performance of the re-trained model with its original version (assessing zero-shot ability) on our test split, encompassing ten tasks.

- **E5:** We assess the zero-shot performance of two additional 3D-LLMs using the complete version of 3DBench to evaluate their capabilities across ten tasks.

LAMM Settings: To ensure a fair comparison of the point cloud understanding abilities among three models, we conduct tests using the 7B versions for all models. During the re-training experiment with LAMM, we identify biases in various evaluation metrics related to output text lengths. As a result, we adjust the target length for different tasks, aiming to reveal the optimal performance of each model on the respective dataset.

PointLLM & Point-LLM Settings: We maintain the default model parameter settings for both models. Following their guidelines, we uniformly sample point clouds to a fixed quantity. The evaluation encompass all ten tasks in 3DBench for PointLLM and Point-LLM. Due to the limitation of training data obtained from public sources, which is confined to the object level, and the potential vulnerability of inference results from untrained tasks to illusions, we focus on the evaluation of scene-level tasks. This aims to scrutinize the performance of large models when confronted with unfamiliar tasks and data.

from the embodied AI simulation framework Ai2thor[Kolve *et al.*, 2017]. In the second step, we reconstruct point clouds and derive instruction-tuning datasets for all tasks based on the ground-truth.

Specifically, we reconstruct scene point clouds from color and depth images of the scene and objects with instance segmentation. Prothor can automatically generate random complete indoor scenes and is built on a simulation robot training framework, enabling us to easily obtain metadata from diverse perspectives. Depending on the scene's size, we teleport the robot to different positions to capture complete scene images. Through this automated process, we acquire 30,000 completely random indoor scene point clouds and over 34,000 object point clouds representing 93 categories of everyday items. We use ground-truth to prompt GPT to acquire rich world knowledge, serving as our training data for tasks such as object relation, QA, and caption. To enhance the ability to handle diverse visual tasks, we process the original ground-truth to obtain results for tasks like counting and room detection. Additionally, we generate diverse dialogue templates tailored for different tasks, embedding the results into conversations to create an augmented instruction-tuning dataset.

Data Statistics. We construct a dataset of more than 0.23 million instruction-tuning samples, utilizing 224,000 for training and 8,000 for evaluation. For the training sets related to visual tasks, we extract over 20 samples from each scene, ensuring that the fine-tuning sample size for each task exceeds 10,000. For text generation tasks, we provide approximately 1,000 training samples for fine-tuning models. The distribution of our instruction-tuning dataset is depicted in Table 1.

4 Experiments

4.1 Experiment Settings

The experimental settings are detailed in Table 2. Five groups of validation experiments are conducted to assess the effectiveness of our benchmark and dataset. As LAMM establishes benchmarks for detection, VQ, and VQA tasks, we use it as a robust baseline. It is originally trained on ShapeNet and 3RScan datasets, and then evaluated on the ScanNet dataset. Our experiments consist of five groups:

4.2 Cross-set Validation (E1 & E2)

Considering LAMM's questionable performance under conventional IOU evaluation, and subsequent corrections revealing a success rate below 1%, we introduce "in box" and "around box" metrics as alternatives to the IOU metric, detailed in Section 3.3. In Fig. 6, it is evident that models trained on our dataset outperformed those trained on
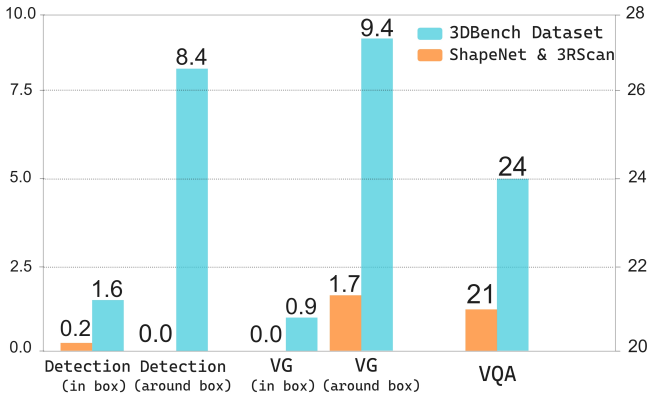
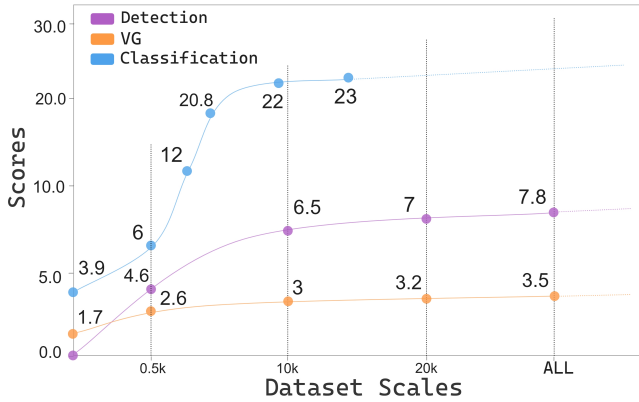Figure 6: Zero-shot results on cross-set validation of group E1 & E2.

Figure 7: Results on the impact of training set scale on 3D-LLMs performance.

	Re-training	Zero-shot	Δ
Detection	7.8	Failed	7.8 ↑
QA (scene)	59.5	62.5	3 ↓
Caption (scene)	71.7	75.9	4.2 ↓
QA (object)	81	66.5	14.5 ↑
Caption (object)	89	70.2	18.8 ↑
Classification	23	3.9	19.1 ↑
Visual Grounding	2.6	1.7	0.9 ↑
Counting	40	16.7	23.3 ↑
Room Detection	7.4	3.3	4.1 ↑
Position Relationship	25	31	6 ↓
Objections Relationship	73.5	63.1	10.4 ↑
Navigation	24.4	6.2	18.2 ↑

Table 3: Results of LAMM on 3DBench, including both zero-shot test results and re-trained results.

4.4 Results after Re-training (E4)

Table 3 displays the performance results of the LAMM on the 3DBench dataset before (to evaluate zero-shot ability) and after re-training, revealing a significant overall improvement in almost all tasks. Particularly notable is the approximately 20% enhancement in classification and counting tasks, suggesting that the features of object point clouds and scene point clouds in the 3DBench dataset are readily learned by the large model. However, there is a discernible decline in performance for partial text generation and position relation tasks. This might be attributed to the use of GPT-3.5 for acquiring world knowledge, potentially causing a disparity in text quality compared to GPT-4, which is employed by LAMM. The experiments, considering comparisons with public datasets and the effectiveness of intrinsic features, collectively affirm the efficacy of the large-scale dataset obtained from our pipeline.

4.5 Comparisons of 3D-LLMs (E5)

In Fig. 9, PointLLM and Point-LLM exhibit significantly better performance in object-level tasks compared to LAMM. PointLLM benefits from its point cloud encoder utilizing point cloud color, and Point-LLM possesses a powerful multimodal feature extractor, allowing them to gain more information compared to LAMM. However, in scene-level tasks, both models achieve inference results close to failure due to the training data only including objects. We plan to further assess their spatial understanding capabilities after re-training both 3D large models using the 3DBench dataset.

5 Observations & Analysis

We summarize a series of observations to assess the performance of 3D-LLMs across diverse tasks, aiming to provide valuable insights to the academic community.

Challenges in Incorporating Additional Features. LAMM demonstrates limited zero-shot classification capabilities, yet a notable improvement is evident when incorporating color information from point clouds as in PointLLM and Point-LLM. Furthermore, we observe a general improvement in classification performance with an increased number of input point clouds, though it tends to reach a plateau rapidly.

ShapeNet and 3RScan in zero-shot results on the LAMM benchmark within the same training framework. This enhancement is attributed to the richer content and broader numerical distribution inherent in our dataset.

4.3 Comparisons of Training Scales (E3)

In Fig. 7, we present the performance of models re-trained with datasets of varying scales on 3DBench. This analysis serves to validate the effectiveness of dataset expansion and discern the maximum limit of performance enhancement with increased data volume. Initially, we assess the impact of dataset size by selecting 33k objects from the first 500 scenes, forming an object-level instruction-following dataset. Subsequently, we extract multiple objects from 30k scenes to create a scene-level dataset, re-training the large model with datasets of different sizes. We further explore the upper limit of performance improvement by expanding the object dataset, as illustrated by the solid line. The results reveal that, as the training dataset size increases, the performance of the LAMM model reaches a plateau around 20k. This suggests that, owing to the incapacity to learn additional features, the simple structure of an encoder plus LLM is insufficient for addressing multi-modal tasks in the 3D domain seamlessly.
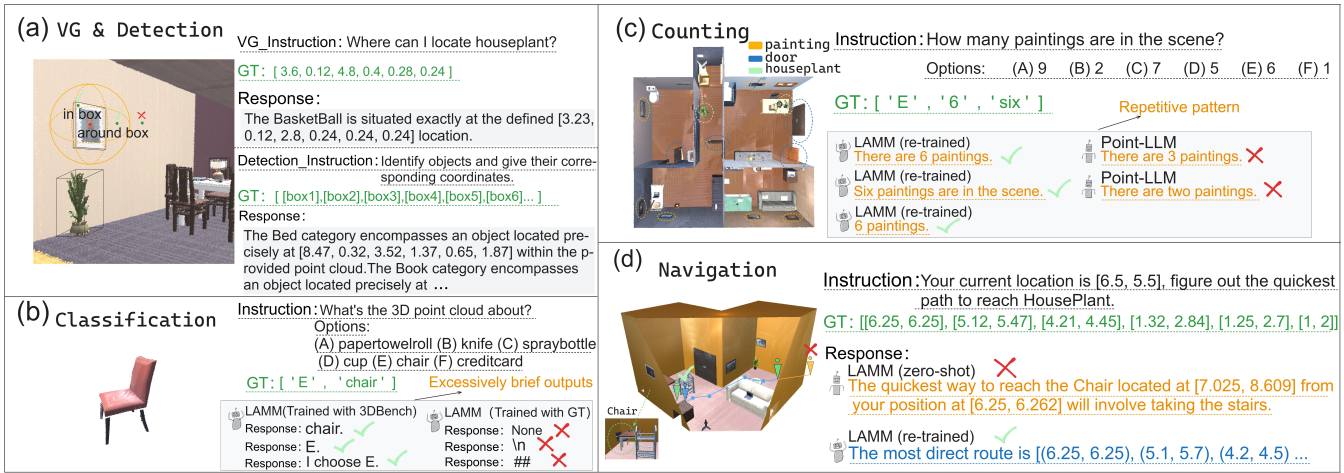
Figure 8: The representative outcomes of 3D-LLMs are showcased as follows: (a) Visualization results for VG and detection tasks. (b) Constant answers in counting. (c) Null output in classification. (d) Divergent answers in navigation before and after re-training.
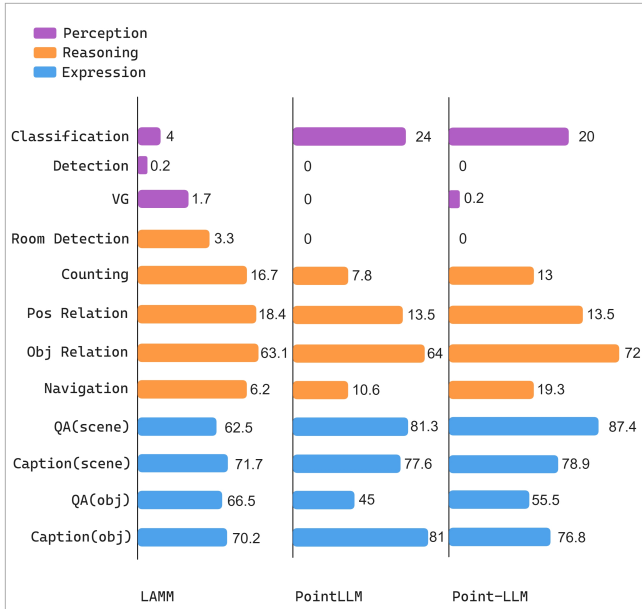

Figure 9: Zero-shot evaluation results for existing 3D-LLMs.

We hypothesize that integrating more efficient feature extraction structures can enhance the performance of traditional vision tasks for 3D-LLMs.

Limitations in Spatial Understanding Capability. 3D-LLMs originally demonstrate subpar performance in tasks involving positional relationships, VG, and object detection. This suggests that existing large models struggle to adeptly capture positional information within scene point clouds. By expanding the training set size and adopting more reasonable evaluation metrics, as depicted in Fig. 8(a), reveals a gradual alignment of inference results with the ground-truth, achieving accuracy several times higher than completely random outcomes. The spatial understanding capability of 3D-LLMs holds considerable potential for enhancement.

Structure of Task Templates. Training 3D-LLMs directly with ground-truth as responses may yield excessively brief outputs, resulting in empty inference results for certain tasks, as demonstrated in the case of classification in Fig. 8(b). Meanwhile, monotonous conversation templates could compromise the diversity and richness of outputs from 3D-LLMs. Upon examining the responses to Point-LLM counting tasks (zero-shot inference on 3DBench) in Fig. 8(c), we notice a repetitive pattern such as "There are N objects" and "The image shows N objects." Therefore, it is advisable to employ GPT for generating diverse predefined conversation templates and integrating ground-truth into these templates to construct an instruction-tuning dataset.

Navigation Benchmark. Fig. 8(d) illustrates results for the navigation task. Navigation places heightened demands on the spatial perception and planning capabilities of 3D-LLMs, particularly in localization tasks. Our evaluation strategy yields improved scores for suboptimal output results, yet still leaves rooms for further improvements.

6 Conclusion

In this paper, we introduce 3DBench, a scalable benchmark designed for evaluating 3D-LLMs covering ten diverse multi-modal tasks with three types of evaluation metrics. Furthermore, we present a pipeline for automatically acquiring high-quality instruction-tuning datasets. Through extensive experiments, we validate the effectiveness of our dataset by cross-validate 3D-LLMs trained with various protocols. Our findings suggest that existing 3D-LLMs have considerable potential for further improvements in point cloud understanding and reasoning. We anticipate that our research will aid the research community in optimizing their models, and inspire the development of more efficient large models and high-quality instruction-tuning datasets.

Acknowledgments

References

[Banerjee and Lavie, 2005] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. *Meeting of the Association for Computational Linguistics,Meeting of the Association for Computational Linguistics*, Jun 2005.

[Chang *et al.*, 2015] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[Dai *et al.*, 2024] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024.

[Deitke *et al.*, 2022] Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Kiana Ehsani, Jordi Salvador, Winson Han, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. Procthor: Large-scale embodied ai using procedural generation. *Advances in Neural Information Processing Systems*, 35:5982–5994, 2022.

[Deitke *et al.*, 2023] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023.

[Dubois *et al.*, 2023] Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback. *arXiv preprint arXiv:2305.14387*, 2023.

[Fu *et al.*, 2022] Yao Fu, Hao Peng, and Tushar Khot. How does gpt obtain its ability? tracing emergent abilities of language models to their sources. *Yao Fu's Notion*, 2022.

[Fu *et al.*, 2023] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

[Guo *et al.*, 2023] Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming Han, Kexin Chen, Peng Gao, Xianzhi Li, Hongsheng Li, et al. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. *arXiv preprint arXiv:2309.00615*, 2023.

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

[Kolve *et al.*, 2017] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. AI2-THOR: An Interactive 3D Environment for Visual AI. *arXiv*, 2017.

[Li *et al.*, 2023] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[Lin, 2004] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. *Meeting of the Association for Computational Linguistics,Meeting of the Association for Computational Linguistics*, Jul 2004.

[Liu *et al.*, 2023] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.

[Luo *et al.*, 2023] Tiange Luo, Chris Rockwell, Honglak Lee, and Justin Johnson. Scalable 3d captioning with pretrained models. *arXiv preprint arXiv:2306.07279*, 2023.

[OpenAI, 2023] R OpenAI. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2:13, 2023.

[Papineni *et al.*, 2001] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, Jan 2001.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[Wang *et al.*, 2023] Bin Wang, Fan Wu, Xiao Han, Jiahui Peng, Huaping Zhong, Pan Zhang, Xiaoyi Dong, Weijia Li, Wei Li, Jiaqi Wang, et al. Vigc: Visual instruction generation and correction. *arXiv preprint arXiv:2308.12714*, 2023.

[Wei *et al.*, 2021] Jason Wei, Maarten Bosma, VincentY. Zhao, Kelvin Guu, AdamsWei Yu, Brian Lester, Nan Du, AndrewM. Dai, and QuocV. Le. Finetuned language models are zero-shot learners. *Learning,Learning*, Sep 2021.

[Xu *et al.*, 2023] Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. *arXiv preprint arXiv:2308.16911*, 2023.

[Xue *et al.*, 2023] Le Xue, Ning Yu, Shu Zhang, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran

Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip-2: Towards scalable multimodal pre-training for 3d understanding. *arXiv preprint arXiv:2305.08275*, 2023.

[Yin *et al.*, 2023] Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Lu Sheng, Lei Bai, Xiaoshui Huang, Zhiyong Wang, et al. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *arXiv preprint arXiv:2306.06687*, 2023.

[Zhu *et al.*, 2023] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.