Probing the Object Awareness of Current 3D LLMs

Tao Sun, Emily Steiner, and Martin JJ. Bucher

CS468 Final Project Report, Fall 2024

Abstract

This project investigates object awareness in 3D Large Language Models (3D-LLMs) through systematic perturbation experiments inspired by polling-based evaluation methods. Using controlled variations in scene density and frame sampling, we evaluate three current 3D-LLM architectures on their ability to accurately detect object presence. Our experiments reveal that models exhibit *density-induced hallucination* — a systematic bias toward false positive detections as scene complexity increases, with negative query accuracy dropping from 70.5% to 20.7% in high-density scenes. Contrary to expectations, we find that model performance saturates with sparse frame sampling, suggesting that visual confusion rather than input sparsity drives hallucination behavior. This study represents an initial step towards identifying model and data bottlenecks in 3D-LLMs for object reasoning capabilities.

1 Introduction

Although LLMs have shown impressive language understanding and reasoning capabilities over the past few years[19;21;7], their adaptation to 3D environments remains preliminary[17]. The trend thus far has been to fine-tune existing pre-trained LLMs such as LLama[7] with additional modalities such as image features, Point Clouds (PCs), or fused features from different visual modalities. Then, a decoder-style LLM outputs text tokens that serve as *spatially grounded* answers for tasks such as Visual Question Answering (VQA), Embodied Reasoning, 3D Scene Captioning, or 3D Object Detection. These models can be broadly aggregated under the umbrella of *3D LLMs*. We define them as multimodal models that take a text prompt and an additional 'visual modality' for a 3D scene as input — and produce text output. Early work such as 3D-LLM[12] pioneered this area and created a dataset with scene-language pairs based on PCs. Following works are simultaneously investigating different directions to incorporate 3D information with LLMs[12;4;29;25;14;23].

1.1 Benchmarks & Data Bottleneck for 3D LLMs

Several benchmarks and datasets have recently been published on 3D Question Answering (3D-QA). ScanQA[1] and SQA3D[18] have been proposed for question answering and reasoning. Additionally, for 3D visual grounding and localization, ScanRefer[26], Multi3DRefer[2], and Scan2Cap[5] have been introduced. Despite the efforts of several works to produce datasets leveraging LLMs for annotation tasks[3;22;13;30;14;15;10], one of the major bottlenecks of current models is the lack of broad internet-scale data for 3D scenes with high quality and diverse scenes and prompts. Thus, the common approach has been to leverage a pre-trained LLM backbone with additional pre-trained 2D image feature or 3D PC feature extractors while using rather small-scale datasets for instruction-tuning that are crafted from existing annotated 3D indoor scene datasets such as ScanNet[5;26]. Recently, 3D-GRAND[24] has proposed to approach the

data limitations through *sim-to-real* transfer via large-scale grounding datasets to solve the data bottleneck, effectively showing that models trained on a large synthetic dataset can perform well on real-world scenes. However, 3D-GRAND does not directly operate with multimodal data, as it explicitly assumes a preprocessing step that extracts a Scene Graph from the scene.

1.2 Object Hallucinations

Alongside novel methods, datasets and benchmark tasks, several works have further probed the limitations and inherent abilities of current models for visual scene understanding. 3D LLM work is mostly evaluated via text-based question-answering (QA) on 3D scenes. This can introduce inherent biases in the evaluation, mainly if the model relies on semantic information from text rather than actual spatial context. For example, when an object is referred to by a textual description ("What's the color of this TV set?"), it is crucial to understand if and how the model uses spatial context — or if it falls back to pre-trained knowledge (*e.g.*, most TV sets are in dark color). Additionally, rooms often have semantic trends (e.g., kitchen), where a model may be able to infer the scene's contents (eg. fridge, oven, etc.) by focusing on general semantics. Thus, even if models perform strongly on these evaluations, their object awareness remains uncertain. Specifically, object-level hallucination occurs when the model incorrectly generates a response that assumes a particular object is present in the scene.

For 2D VLMs, POPE introduced polling-based query methods to probe and investigate object-specific hallucinations[16]. To systematically evaluate the presence of hallucinations, POPE polls the model with questions asking *"Is there a __ in the image?"*. The objects polled are grouped into categories of random, popular and adversarial settings to vary the difficulty of questions. The more challenging categories, popular and adversarial, are determined by querying the top-k common objects in the dataset and top-k common co-occurrences with objects in the sample in the dataset, respectively. Object hallucination remains a common issue in 3D LLMs where even less data is available for training. 3D-POPE naturally extends POPE to 3D by introducing an object hallucination benchmark using triplets of scenes from ScanNet200 with random, popular, and adversarial queries[24]. However, they do not explore out-of-distribution scenes or further explore the reason for hallucination.

The source of hallucinations or factual errors in VLMs is challenging to isolate a single cause from an extensive list of interdependent factors (encoder architecture, decoder architecture, prompt engineering, training choices, training data diversity, parameters, etc.). Recent work[27] found that 2D-VLMs underperform on image classification due to insufficient training data for the decoder, despite critical information being captured in the latent space.

2 Method

Inspired by the 3D-POPE benchmark, we investigated the task of 3D Object Probing, *i.e.* asking the model to answer if an object exists in the scene or not with 'Yes' or 'No'. We tested synthetic and real-world scenarios using two existing indoor scene datasets with ground-truth semantic annotations. By exploring a customized dataset, we ensure that the proposed data is outside the distribution of typical 3D scene understanding tasks. We introduced varying degrees of modifications to them while keeping the prompting scheme fixed.

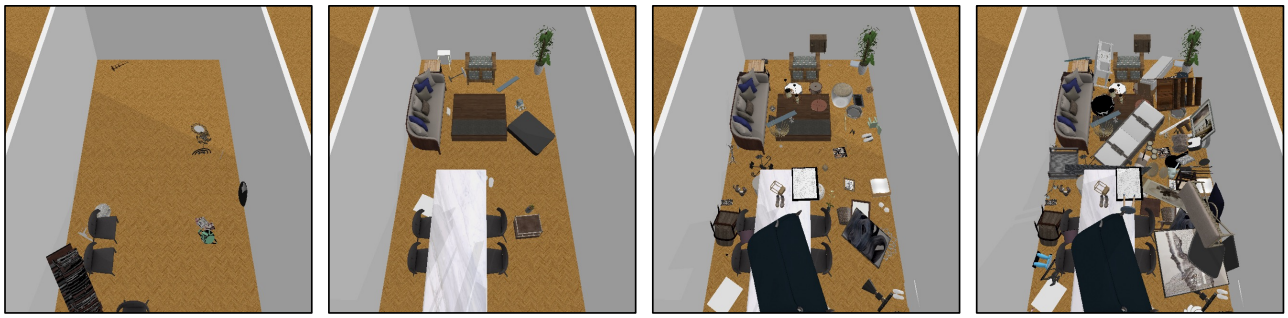

Figure 1: **Scene with varying densities generated from one room of 3D-FRONT**: The caption below each figure shows the density bin (left) and the density value (right).

2.1 Dataset Creation

We control the perturbations to evaluate potential bias with 3D data via varying (i) scene density and (ii) frame density.

Exp 1: Scene Density — For scene density, we vary the number of objects present in the scene to control the amount of visual noise present. This perturbation isolates evaluation to assess a 3D-LLM's ability to handle visual noise or confusion. Data samples are generated from synthetic datasets 3D-FRONT and 3D-FUTURE[8;9]. The initial 3D-FRONT room layouts are populated with objects using a physics-based simulation provided by PyBullet[6]. Objects are added to the scene by being dropped from a random height level of $[0, 15]$ m above the ceiling, and the simulator determines their final stable position. The number of objects follows an exponential search, starting with 16 objects and increasing until a target scene density of at least 70% is reached. Here, scene density is defined as the ratio of the total area covered by the projected bounding boxes of all objects on the floorplan. To further vary scene density in a controlled manner, we sequentially remove objects from the scene based on an object collision graph. Only non-dependent objects (*i.e.*, those not supporting other objects) are removed in each iteration to avoid object collapse. Figure 1 showcases visual examples of a room with different densities, from a relatively empty room to a highly cluttered scene.

Exp 2: Frame Density — To vary frame density, we randomly sample 100 scenes from the ScanNetV2 validation set where we have ground-truth instance-level semantic annotation. We then introduce a range of different frame numbers $N = \{5, 10, 15, 20, 30, 50, 75, 100, 150, 200, 300\}$, sampling $n_i \in N$ frames along with corresponding camera poses and depth maps. For models utilizing multi-view images, this data is sufficient. For instance-level point cloud models, we reproject RGB pixels into 3D using camera poses and depth maps, then match them to the dense reconstruction of object instance point clouds using nearest neighbor matching. This simulates sparser point clouds at varying frame densities.

Prompt Generation — For 100 base scenes, each density modification is made, and five positive and five negative questions are generated per modification. Positive questions are randomly selected from the scene's objects. For negative questions for custom scenes, GPT-4o API suggested absent objects based on the descriptions of existing objects. Negative samples for ScanNetV2 scenes are randomly chosen from absent classes in the dataset.

2.2 Models Evaluated

We tested three different current 3D LLMs. The models are: (i) 3D-LLM[23], which uses projected multi-view 2D image features and an out-of-the-box VLM to encode visual tokens, (ii) LEO[14], which uses PointNet[20] along with a Hierarchical Spatial Transformer[11] as the point cloud encoder*, and (iii) LLaVA-3D[29], which uses multi-view images as the 3D input. All three models, LLaVA-3D, LEO, and 3D-LLM have shown strong performance on various visual-language navigation tasks.

3 Experiments Results

Exp 1: Scene Density — For the first experiments, we present the overall accuracy of two methods, LEO and LLAVA-3D, across different density bins in Figure 2. LEO consistently outperforms LLAVA-3D in all density bins. At the sparsest bin (10-20%), LEO achieves 62.4% accuracy, compared with 55.4% of LLAVA-3D. LEO maintains above 53.5% accuracy even in the most dense scenes (70%+). In contrast, LLAVA-3D struggles with around 50% accuracy (random guess) in scenes with a density larger than 30%. These results highlight the limitations of SOTA models in handling dense 3D scenes and strengthen the need for improved object awareness of current 3D LLMs.

We further examine model performance on positive and negative questions, as shown in Figure 3. Interestingly, model performance on negative questions reveals an unexpected trend: as scene density increases, the accuracy for negative questions in both models falls below the 50% random guessing threshold for scenes with more than 40% density. For scenes of density above 70%, the accuracy of LEO and LLAVA-3D drop to 20.7% and 36.1%, respectively. We refer to this phenomenon as *Density-induced Hallucination*. We hypothesize that when there are significantly more objects in the scene, the models tend to assume that "everything is possible to be in the room." We hope our dataset will enable further research on mitigating this type of hallucination. Notably, addressing samples that fall below the 50% threshold could already gain meaningful performance improvements.

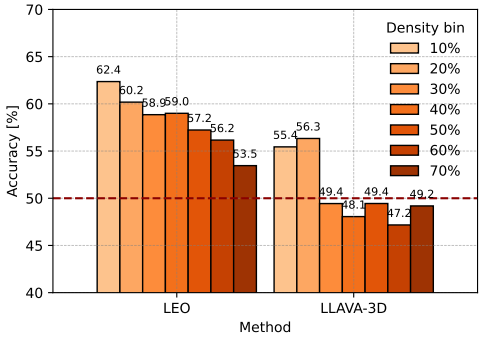

Figure 2: **Accuracy on Different Density Bins**: LEO consistently outperforms LLAVA-3D in all density bins.

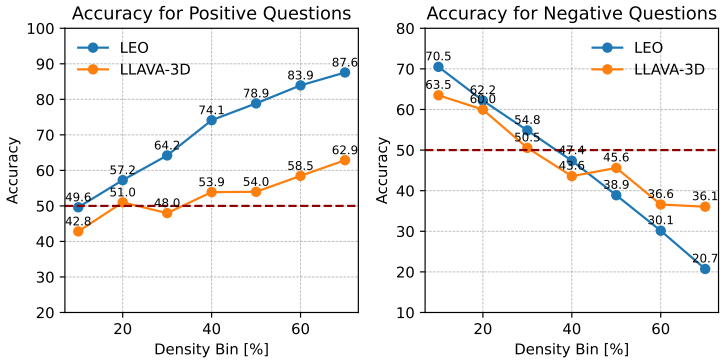

Figure 3: **Accuracy for Positive and Negative Questions**: Both models present a significant decline as scene density increases.

*While LEO also supports an auxiliary 2D image input, we do not use that as part of the input.

Exp 2: Frame Density — For the second experiment, we show our main results in Figure 4, where we plot the accuracy for each modification level against its False Negative Rate (FNR). The underlying hypothesis was that fewer frames might lead to less density in the visual data, and thus, the models might miss more objects, leading to a higher FNR. However, the results reject this hypothesis, as a handful of frames through evenly-spaced sampling — covering a good range of the scene — seems already enough to cover most objects. Thus, performance already saturates quite soon with higher frame density. Additionally, in Figure 5, we aggregate the accuracy delta between 300 and 5 frames for each semantic label, sorted in descending order. The hypothesis was that smaller objects might have a larger delta since they will only be visible from certain angles (and frames). However, we can reject this second hypothesis since no clear trend is visible.

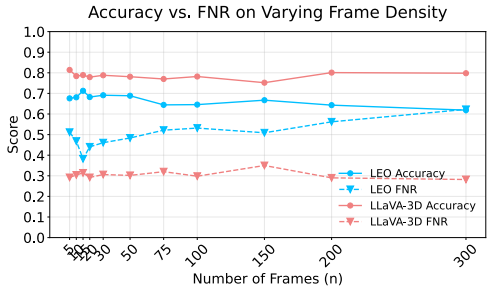

Figure 4: **Accuracy & FNR of different frame densities**: Fewer frames do not lead to much lower accuracy, indicating that performance already saturates quickly.

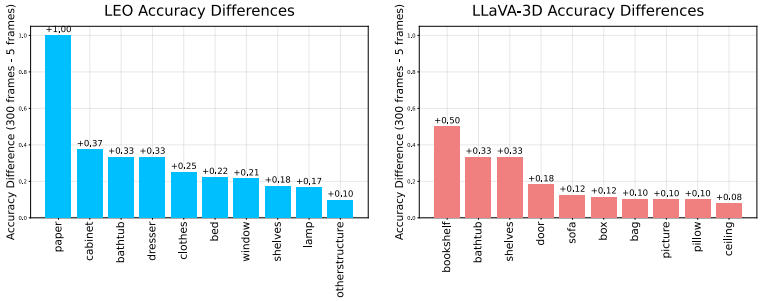

Figure 5: **Accuracy Delta** between 300 Frames and 5 Frames, aggregated by each semantic class and sorted descending by magnitude. No clear trend is visible for object size and higher frame rate.

We adapted 3D-LLM [23] for both tasks but found it consistently responded ‘yes‘ to all prompts due to a lack of yes/no questions in training data. To better assess object awareness, we adjusted prompts to resemble training tasks. For Exp 2, while ScanNet200 was included in the training, the model produced identical descriptions regardless of input density, indicating mode collapse. Additionally, it failed to generate meaningful results for the custom dataset in Exp 1, showing poor generalization to out-of-distribution data and prompts.

4 Conclusions

This project examines recent advances in 3D Large Language Models (LLMs) that reason about 3D indoor environments. A prevalent issue for current models, which are heavily bottlenecked by training data, is object-level hallucination. We hypothesize that these models may not fully engage with visual data, leading to ungrounded responses. Our empirical study on 3D Object Probing for several leading 3D LLMs reveals that as scene density increases, the models show overconfidence in ’Yes’ responses, resulting in a higher false positive rate, indicating visual confusion. However, it’s unclear whether this arises from limitations in visual encoding, insufficient training data for the text decoder, or sensitivities in the prompting style after fine-tuning. Future research will compare more models and employ Linear Probing by training a binary classifier on the first output token’s embeddings. This could help determine if the models possess the right knowledge but struggle to decode the correct answers when mapped to the token vocabulary, as seen in 2D Visual Language Models (VLM) [28].

References

[1] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19129–19139, 2022.

[2] Dave Zhenyu Chen, Angel X. Chang, and Matthias Nießner. ScanRefer: 3D Object Localization in RGB-D Scans using Natural Language, November 2020. arXiv:1912.08830 [cs].

[3] Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. LL3DA: Visual Interactive Instruction Tuning for Omni-3D Understanding, Reasoning, and Planning, November 2023. arXiv:2311.18651 [cs].

[4] Yilun Chen, Shuai Yang, Haifeng Huang, Tai Wang, Ruiyuan Lyu, Runsen Xu, Dahua Lin, and Jiangmiao Pang. Grounded 3d-llm with referent tokens. *arXiv preprint arXiv:2405.10370*, 2024.

[5] Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3193–3203, 2021.

[6] Erwin Coumans and Yunfei Bai. Pybullet, a python module for physics simulation for games, robotics and machine learning. http://pybullet.org, 2016–2021.

[7] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[8] Huan Fu, Bowen Cai, Lin Gao, Lingxiao Zhang, Jiaming Wang Cao Li, Zengqi Xun, Chengyue Sun, Rongfei Jia, Binqiang Zhao, and Hao Zhang. 3D-FRONT: 3D Furnished Rooms with layOuts and semaNTics, May 2021. arXiv:2011.09127 [cs].

[9] Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3D-FUTURE: 3D Furniture shape with TextURE, September 2020. arXiv:2009.09633 [cs].

[10] Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. Scene-LLM: Extending Language Model for 3D Visual Understanding and Reasoning, March 2024. arXiv:2403.11401 [cs].

[11] Wenchong He, Zhe Jiang, Tingsong Xiao, Zelin Xu, Shigang Chen, Ronald Fick, Miles Medina, and Christine Angelini. A hierarchical spatial transformer for massive point samples in continuous space. *Advances in neural information processing systems*, 36:33365–33378, 2023.

[12] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3D-LLM: Injecting the 3D World into Large Language Models, July 2023. arXiv:2307.12981 [cs].

[13] Haifeng Huang, Yilun Chen, Zehan Wang, Rongjie Huang, Runsen Xu, Tai Wang, Luping Liu, Xize Cheng, Yang Zhao, Jiangmiao Pang, and Zhou Zhao. Chat-Scene: Bridging 3D Scene and Large Language Models with Object Identifiers, September 2024. arXiv:2312.08168 [cs].

[14] Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

[15] Baoxiong Jia, Yixin Chen, Huangyue Yu, Yan Wang, Xuesong Niu, Tengyu Liu, Qing Li, and Siyuan Huang. SceneVerse: Scaling 3D Vision-Language Learning for Grounded Scene Understanding, September 2024. arXiv:2401.09340 [cs].

[16] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating Object Hallucination in Large Vision-Language Models, October 2023. arXiv:2305.10355 [cs].

[17] Xianzheng Ma, Yash Bhalgat, Brandon Smart, Shuai Chen, Xinghui Li, Jian Ding, Jindong Gu, Dave Zhenyu Chen, Songyou Peng, Jia-Wang Bian, Philip H. Torr, Marc Pollefeys, Matthias Nießner, Ian D. Reid, Angel X. Chang, Iro Laina, and Victor Adrian Prisacariu. When LLMs step into the 3D World: A Survey and Meta-Analysis of 3D Tasks via Multi-modal Large Language Models, May 2024. arXiv:2405.10255 [cs].

[18] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. *arXiv preprint arXiv:2210.07474*, 2022.

[19] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[20] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *arXiv preprint arXiv:1612.00593*, 2016.

[21] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[22] Zehan Wang, Haifeng Huang, Yang Zhao, Ziang Zhang, and Zhou Zhao. Chat-3D: Data-efficiently Tuning Large Language Model for Universal Dialogue of 3D Scenes, August 2023. arXiv:2308.08769 [cs].

[23] Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. PointLLM: Empowering Large Language Models to Understand Point Clouds, September 2024. arXiv:2308.16911 [cs].

[24] Jianing Yang, Xuweiyi Chen, Nikhil Madaan, Madhavan Iyengar, Shengyi Qian, David F. Fouhey, and Joyce Chai. 3D-GRAND: A Million-Scale Dataset for 3D-LLMs with Better Grounding and Less Hallucination, June 2024. arXiv:2406.05132 [cs].

[25] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[26] Yiming Zhang, ZeMing Gong, and Angel X Chang. Multi3drefer: Grounding text description to multiple 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15225–15236, 2023.

[27] Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. Why are Visually-Grounded Language Models Bad at Image Classification?, May 2024. arXiv:2405.18415 [cs].

[28] Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. Why are visually-grounded language models bad at image classification? *arXiv preprint arXiv:2405.18415*, 2024.

[29] Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. LLaVA-3D: A Simple yet Effective Pathway to Empowering LMMs with 3D-awareness, September 2024. arXiv:2409.18125 [cs].

[30] Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3D-VisTA: Pre-trained Transformer for 3D Vision and Text Alignment, August 2023. arXiv:2308.04352 [cs].