Understanding and Evaluating Hallucinations in 3D Visual Language Models

Anonymous ACL submission

Abstract

Recently, 3D-LLMs, which combine point-cloud encoders with large models, have been proposed to tackle complex tasks in embodied intelligence and scene understanding. In addition to showing promising results on 3D tasks, we find that they are significantly affected by hallucinations. For instance, they may generate objects that do not exist in the scene or produce incorrect relationships between objects. To investigate this issue, this work presents the first systematic study of hallucinations in 3D-LLMs. We begin with quickly evaluating hallucinations in several representative 3D-LLMs and reveal that they are all significantly affected by hallucinations. We then define hallucinations in 3D scenes and, through a detailed analysis of datasets, uncover the underlying causes of these hallucinations. We find three main causes: (1) Uneven frequency distribution of objects in the dataset. (2) Strong correlations between objects. (3) Limited diversity in object attributes. Additionally, we propose new evaluation metrics for hallucinations, including Random Point Cloud Pair and Opposite Question Evaluations, to assess whether the model generates responses based on visual information and align it with the text's meaning.

1 Introduction

Large Language Models (LLMs) have achieved impressive results in tasks such as code completion (Kanade et al., 2020; Wang et al., 2021), mathematical reasoning (Jiang et al., 2024; Guo et al., 2024), and dialogue generation (Li et al., 2024; Le et al., 2020). Motivated by their success, researchers have extended multi-modal domains. Vision language models (VLMs) (Wang et al., 2024; Deitke et al., 2024) allow models to process images and text jointly. However, 2D visual data provide limited spatial cues as a result of its single-perspective nature. To overcome this, 3D-LLMs (Hong et al., 2023; Xu et al., 2024; Zhen et al., 2024) incorporate point clouds to better understand spatial relationships. These models typically extract features via a point cloud encoder and align them with LLM token space, enabling performance gains in 3D reasoning tasks.

Despite their potential **hallucinations** (Rohrbach et al., 2018; Li et al., 2023; Hu et al., 2023; Guan et al., 2024)—the generation of plausible yet false information—persist across LLMs and VLMs. This undermines their reliability in critical fields like healthcare and law. Existing benchmarks such as TruthfulQA, HalluQA, CHAIR, and POPE have been proposed to evaluate hallucinations in text and 2D visual outputs. However, hallucinations in 3D-LLMs remain underexplored.

The inclusion of depth and geometry introduces new challenges in defining and evaluating hallucinations in 3D contexts. In this work, we first formalize 3D hallucinations, distinguishing them from their 2D and textual counterparts. We then evaluate state-of-the-art 3D-LLMs and reveal that spatial hallucinations are widespread. Our analysis attributes this to high object co-occurrence bias in training data. Unlike prior work focusing on object presence, we emphasize spatial relationship hallucinations and introduce a benchmark to detect them effectively.

Our contributions are: (1) We provide the first formal definition and taxonomy of 3D hallucinations. (2) We evaluate and analyze hallucination patterns in representative 3D-LLMs. (3) We propose a new dataset and benchmark for detecting spatial relationship hallucinations.

2 Related Work

2.1 3D LLMs

Large Vision Models (LVMs) (Shen et al., 2024; Zhang et al., 2022; Kirillov et al., 2023; Oquab et al., 2023) have achieved strong performance across various tasks, motivating their extension to

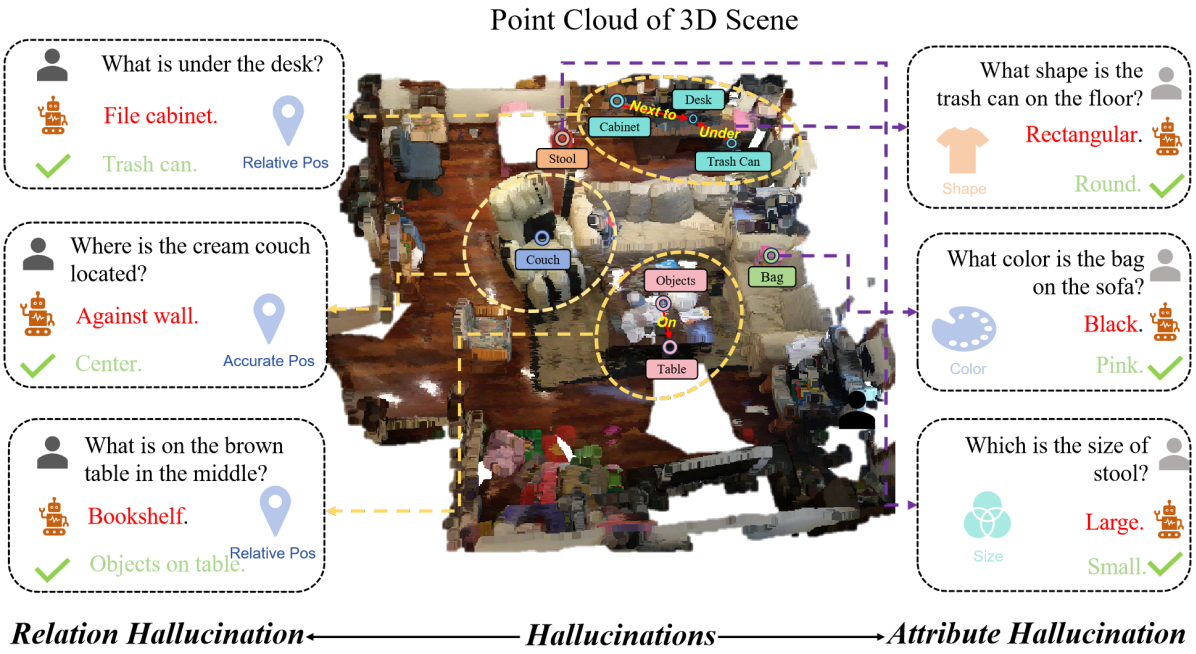

Figure 1: In 3D scenes, the relationships between objects are significantly more complex than those in text or images. The left side of the figure illustrates hallucinations related to relative positional relationships and absolute positional relationships, while the right side demonstrates attribute hallucinations such as color, size, and shape.

other modalities. 3D tasks such as semantic navigation (Zheng et al., 2024; Huang et al., 2023) and embodied intelligence (Jatavallabhula et al., 2023; Hong et al., 2024) has received growing attention due to their real-world relevance, with many approaches leveraging the reasoning capabilities of LLMs.

3D-LLMs (Hong et al., 2023) typically consist of a 3D encoder that maps point clouds into the language space of a pre-trained LLM. Different models vary in their encoding strategies. 3D-LLM (Hong et al., 2023) extracts multi-view 2D features to construct 3D representations using traditional methods. LL3DA (Chen et al., 2024) uses a scene encoder pretrained on ScanNet (Dai et al., 2017) as the point cloud encoder. Leo (Huang et al., 2023) adopts an object-centric approach by encoding each object with a point cloud encoder followed by a spatial transformer. After fine-tuning on downstream 3D tasks, these models exhibit strong spatial reasoning abilities.

These 3D-LLMs have shown promising performance on tasks such as 3D dense captioning, 3D question answering, and scene description.

2.2 Hallucination in Multimodal LLMs

In LLMs, hallucinations refer to outputs that appear plausible but are not faithful to facts or context (Filippova, 2020). These errors undermine the reliability of LLMs in real-world applications. Existing work (Leng et al., 2024; Liu et al., 2023; Yu et al., 2024; Zhai et al., 2023) mitigates hallucinations through model editing, post-training, or contrastive decoding.

As LLMs are increasingly integrated into multimodal systems, hallucinations in LVLMs have become a key research focus. In this context, hallucinations occur when generated text misaligns with visual content (Rohrbach et al., 2018; Li et al., 2023; Hu et al., 2023; You et al., 2023). Previous work mainly targets object-level hallucinations, including those related to object types, attributes, and relationships. Mitigation strategies mirror those of LLMs, including methods at the data level, training level, and decoding level. A large body of research (Hu et al., 2023; Liu et al., 2023) has shown that one significant cause of hallucinations is data bias. The homogeneity of the tasks and the lack of diversity in scenarios limit the model's ability to understand visual information and follow instructions across different environments.

Hallucinations are particularly problematic in 3D tasks such as embodied intelligence and spatial navigation, where accurate spatial understanding is critical. Yet, hallucinations in 3D-LLMs remain unexplored. This work addresses that gap by detecting and analyzing hallucinations in 3D-LLMs.

3 3D Hallucination

In this section, we first validate the existence of significant hallucination issues in the current popular 3D-LLMs on the 3D captioning task using traditional object-centric method which is used in image hallucination evaluation. We then define 3D hallucinations and compare them with the multimodal hallucinations defined in previous works.

3.1 Simple Evaluation Based on Traditional Detection Methods

	Precision	Recall	F1Score	Rouge	Meteor
LL3DA	36.36	16.67	22.86	25.87	14.98
3D-LLM	22.97	8.20	10.92	9.94	4.37
LLaVA-3D	29.44	12.28	15.27	13.72	6.95

Table 1: Evaluate Result of Sota 3D-LLM.Precision reflects the probability that a mentioned object actually exists in the scene — lower precision indicates a higher object hallucination rate. Recall measures how well the description covers the objects present in the scene — higher recall suggests a more comprehensive depiction of the scene.

First, we evaluate whether existing 3D-LLMs suffer from object hallucinations—describing objects not present in the real scene—using the traditional image-text definition. We test this by having 3D-LLMs describe scene point clouds and flag descriptions that include nonexistent objects as hallucinations. We employ precision (Fisher, 1936) and recall to evaluate the probability that the objects described in the generated captions belong to the scene, as well as the coverage of the descriptions over the scene. Formally, we define A as the set of items output by the model, representing $TP + FP$, and B as the set of items present in the real scene,representing $TP + FN$. The evaluation metrics can be defined as:

$$Precision = \frac{|A \cap B|}{|A|} \quad (1)$$

$$Recall = \frac{|A \cap B|}{|B|} \quad (2)$$

To validate that existing 3D models suffer from significant object hallucinations, we selected three representative 3D models : LL3DA , 3D-LLM, and LLaVA-3D for evaluation. We used the metric defined above. The results are presented in Table 1. As we can see, all three models perform badly and exhibit significant hallucination issues in the object description task. To better illustrate the evaluation of hallucinations, we present our evaluation of LL3DA on the description task as a Recall-Precision plot, as shown in Fig. 2. The plot is divided into the bottom-left corner and the top-right corner. The bottom-left corner indicates that the model struggles with hallucinations in the object description task, while the top-right corner demonstrates that the model performs well. It can be observed that most of the samples are concentrated in the lower-left corner of the plot, which reflects the presence of severe hallucinations in the majority of examples produced by the current state-of-the-art models.

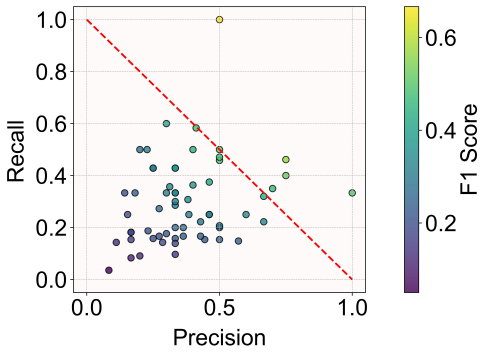

Figure 2: Object hallucination evaluation for 3D LLMs. Precision measures the proportion of described objects that exist in the scene, while recall represents the proportion of scene objects that are described.

3.2 3D Hallucination Definition

3.2.1 Modality Difference

Previous hallucination studies focus on text and image modalities and their interactions. Since 3D-LLMs differ mainly in input modality, we analyze hallucinations from this perspective. As Table **??**Different_modalitytable Different_modalitys, unlike text-based LLMs and text-image LVLMs, 3D-LLMs use text and point clouds, adding depth information.

Model Type	Input Modality			Modality Conflict		
	Text	Vision	Depth	Knowledge Conflict	Text-Image Conflict	Scene Conflict
LLM	✓	✗	✗	✓	✗	✗
LVLM	✓	✓	✗	✓	✓	✗
3D-LLM	✓	✓	✓	✓	✓	✓

Table 2: Modality Difference

The uniqueness of the input modalities leads to differences in the interactions between modalities. In text hallucinations, conflicts only arise between different textual knowledge, i.e., knowledge conflicts, which are also presented in LVLMs

Model Type	Object Hallucination			Relation Hallucination		
	Color	Shape	Size	Abstract	Relative	Accurate
Text Hallucination	✓	✗	✗	✓	✗	✗
Image Hallucination	✓	✓	✗	✓	✓	✗
3D Hallucination	✓	✓	✓	✓	✓	✓

Table 3: Classification of Hallucinations

and 3D-LLMs, as both are built on LLMs. In image hallucinations, conflicts occur between textual and visual information. However, in 3D hallucinations, the depth information leads to conflicts where 3D-LLMs generates fictitious spatial relationships within the scene. We refer to this phenomenon as **scene conflict**.

3.2.2 Hallucination Definition

To define hallucination types appeared in scene conflict more concretely and accurately, we abstract the 3D scene into objects and relationships, thus defining two types of hallucinations: **Object hallucinations** and **Relation hallucinations**. We present the classification in Table 3.

Object hallucinations are primarily related to the attributes of objects, such as color, shape, and size. Among these attributes, **size attribute** requires accurate depth information for proper evaluation, making this a hallucination type unique to 3D scenes. Formally, we use H_{obj} to represent object hallucination, S to represent the attributes set. $Attr^i_{true} \in S$ represents the real object's attribute. $Attr^i_{pred}$ represents the attributes in the prediction of 3D-LLM.

$$H_{obj} = S[Attr^i_{true} \neq Attr^i_{pred}] \tag{3}$$

Relation hallucinations, on the other hand, are primarily concerned with the relationships between objects. Among these relations, **Abstract relationship hallucinations** refer to the functional relationships between objects. **Relative positional relationships** refer to broader postional relationships, such as left-right orientation, which can usually be inferred from a given view. However, because a single view lacks depth information, precise positional relationships, such as "hanging" or "standing on," cannot be determined. In 3D scene, we can deduce **accurate spatial relations** among objects. Formally, we use O_i and O_j to represent two objects, $\xrightarrow{rel}$ to represent relationship between two objects, and $\xrightarrow{pred}$ to represent predicted relationship. The we can define relation hallucination as:

$$O_i \xrightarrow{rel} O_j \neq O_i \xrightarrow{pred} O_j \tag{4}$$

4 Data Bias Intensifies 3D Large Model Hallucinations

In the previous section, we briefly examined the significant hallucinations present in existing 3D large models and provided an analysis and definition of hallucinations in 3D scenes. In this section, we will delve into the underlying causes of this phenomenon. In Section 3 of our study, we evaluated the occurrence of object hallucination in large 3D point cloud models. We found that the model often describes objects that do not exist in the actual scene.We hypothesize that imbalanced object frequencies and object corelation in the dataset contribute heavily to hallucination.

4.1 Imbalanced Frequency Distribution of Objects

We performed statistical analysis on the hallucination rate and occurrence frequency of objects. The hallucination rate(HR) of an object is defined as the ratio of scenes in which the object is incorrectly identified as present, even though it does not actually exist, to the total number of scenes where the object is absent in the test set. The occurrence frequency of an object is defined as the ratio of scenes where the object is present to the total number of scenes. As shown in Figure 3, a represents the object hallucination rate results for 3DLLM, and b represents the object hallucination rate results for LL3DA. From the figures, it can be observed that the curve representing the hallucination rate closely follows the curve representing the occurrence frequency. This suggests that objects with a high frequency of occurrence are more likely to be accurately described by the model, as it tends to repeat the most common elements. In other words, **objects with higher occurrence frequencies are more prone to hallucination**, being more likely to be incorrectly identified as present when they are actually absent.

However,in the Scannet dataset, certain objects such as the floor, wall, and door appear very frequently across many scenes. **Floor** appeared in **1506 out of 1513** scenes. **Wall** appeared in **1473** scenes. **Door** appeared in **1015** scenes. These data demonstrate that scene similarity in ScanNet is high, with the same object appearing repeatedly across multiple scenes. Based on the conclusion that excessively high occurrence frequencies can exacerbate hallucinations, we can infer that **the high overlap of objects across different scenes in**

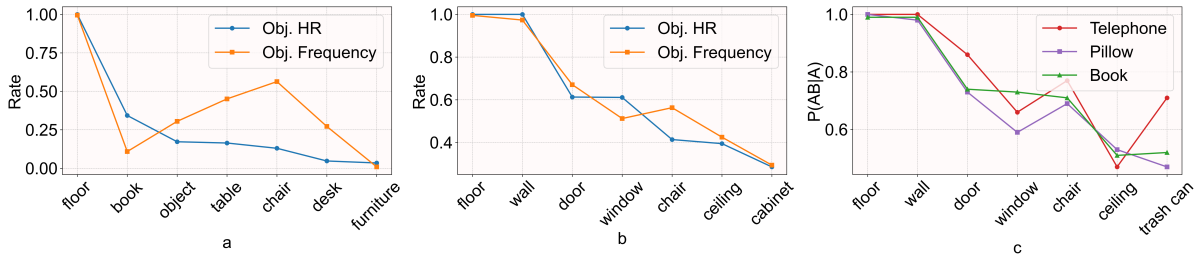

Figure 3: (1) Figures a and b show the relationship between object hallucination rates in 3DLLM and LL3DA and object occurrence frequencies in the dataset. The blue and orange lines represent hallucination rates and object frequencies, respectively. (2) Figure c shows the relationship between strong object correlations and hallucination rates. The y-axis indicates the conditional probability of object occurrence, and the x-axis represents condition A. For example, the red line shows the probability of "telephone" being present given the presence of the object on the y-axis.

the dataset is one of the key factors **contributing to the strong hallucinations** observed in 3D large language models.

4.2 Potential Influence of Object Correlation

In Figure 3, the y-axis represents the conditional probability $P(AB|A)$, where A denotes the presence of object a in the scene and B denotes the presence of object b. A higher value of $P(AB|A)$ indicates a higher likelihood that if object a is present, object b is also likely to be present. The objects b labeled on the x-axis, such as floor, wall, and door, are arranged in descending order of their hallucination rates, and the conditional probabilities also exhibit a downward trend. This suggests that **objects frequently co-occurring with others are more likely to be incorrectly identified as present**, thereby inducing hallucinations. For example, if **chairs** and **tables** often appear together in the same scene, the model might learn an implicit dependency between them. When the chair is present, the model may "hallucinate" the table, even if it isn't present in the actual scene.

ScanNet is an indoor scene dataset containing environments such as bedrooms, bathrooms, and offices. Due to the specific nature of these scenes, they consistently include certain objects—such as toilets, sinks, and toilet paper—always appearing together in bathrooms. This strong correlation between objects in the dataset means that during training, the model may receive rewards for providing answers based on these associations rather than point clouds. As a result, the model may incorrectly associate these objects with one another, leading to hallucinations when detecting one object.

5 Proposed Evaluation Frameworks for 3D Hallucinations

5.1 Inadequacy of Existing Evaluation Frameworks

Existing evaluation frameworks for 2D multimodal models, such as POPE (Li et al., 2023), are insufficient for addressing the challenges in 3D point cloud large language models (LLMs). Since the POPE view uses yes/no questions to evaluate model object hallucinations, which cannot accurately assess the model's understanding of spatial relationships or visual details such as attributes.In Section 3, we assess hallucinations in 3D point cloud models by evaluating object hallucination in description tasks. However, this method has two main limitations: 1) It only detects hallucinations in description tasks, as not all responses involve objects. 2) It doesn't analyze other types of hallucinations, such as attribute or relational errors.

Therefore, we aim to propose a more stable, fair, and flexible evaluation framework for evaluating hallucinations in 3D point clouds.

5.2 Proposed Evaluation Framework

We propose two strategies for detecting hallucinations in 3D point cloud models.

Random Point Cloud Pair Evaluation We select a random point cloud and ask the model the same question on both the original and new point clouds. If the answers are identical, it's considered a hallucination, indicating the model doesn't integrate visual context and just maps the question to a fixed answer.

Opposite Question Evaluation For a fixed point cloud, we ask two Opposite questions (e.g., "What is on the right of the table?" and "What is on the

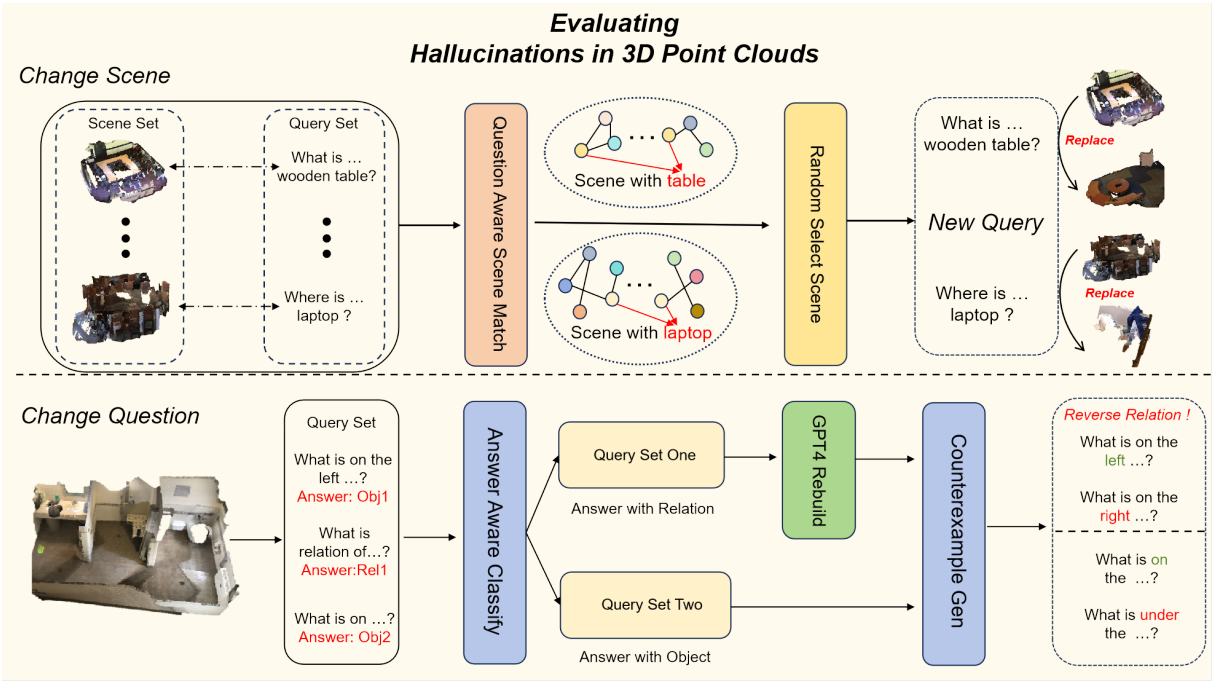

Figure 4: In the evaluation process, we generate new QA pairs by changing the scene while keeping the questions fixed: different scenes are randomly selected to form new QA pairs. Additionally, we modify the questions while keeping the scene fixed: spatial relationship-related questions are selected, and all QA pairs are transformed such that the object A is the focus. Then, the spatial relationship in the questions is inverted, generating new QA pairs.

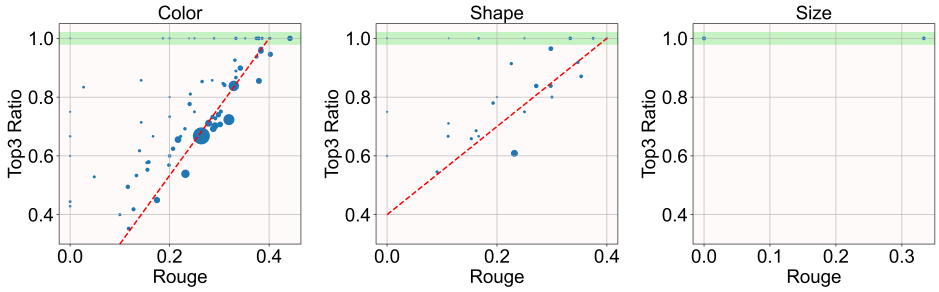

Figure 5: Impact of Attribute Simplicity on Accuracy.ROUGE represents the average quality of question-answer pairs for a specific item, while the Top 3 Ratio is the proportion of the three most common attributes of the item.

left?"). If the model gives the same answer, it's a hallucination, suggesting the model isn't using the spatial information from the point cloud.

By employing these two strategies, we aim to identify cases where the model fails to distinguish between spatially different scenarios or produces inconsistent responses to questions.

5.3 Inadequacy of Existing Evaluation Frameworks

The entire pipeline is illustrated in Figure 4.

Data Generation We first construct a *scene graph* G_i for each scene, where G_i consists of a set of relational triplets in the form of $(object_1, object_2, relation)$. These triplets are used to evaluate *scene similarity* and to verify whether the spatial relationships described in questions are actually present in the scene.

In the **Change Scene** experiment, for each QA-scene triplet (Q_i, A_i, S_i), we randomly select a different scene S_j from the dataset to construct a new data instance: $(Q_i, A_i, \{S_i, S_j\})$. To ensure that S_j does not contain the spatial relation required to answer Q_i, we extract the spatial relation triplets from S_i and S_j, denoted as $T(S_i)$ and $T(S_j)$, respectively, and enforce that:

$$T(S_i) \cap T(S_j) = \emptyset \tag{5}$$

This guarantees that the same question Q_i leads to different answers in S_i and S_j.

Type	LL3DA		3D-LLM		LEO		LLaVA3D	
	ROUGE-L	HR_{ran}%	ROUGE-L	HR_{ran}%	ROUGE-L	HR_{ran}%	ROUGE-L	HR_{ran}%
Direction	31.09	19.89	30.46	39.03	17.02	**42.66**	5.22	21.58
Containment	41.30	25.24	41.78	**49.00**	17.45	38.84	4.4	23.30
Contact	33.46	24.53	35.46	**47.20**	13.82	45.48	2.98	20.94
Distance	31.05	20.85	32.19	**38.10**	12.98	37.02	3.13	19.57
Color	47.72	49.02	52.00	**79.37**	37.03	66.94	0.21	33.18
Shape	42.67	41.58	44.61	**67.02**	34.61	65.98	0.74	52.92
Size	53.50	**77.14**	47.48	68.58	34.00	62.86	0.0	42.86
Comparison	24.75	**63.16**	29.43	**63.16**	10.48	47.37	0.0	52.63
Quantity	51.29	51.60	49.44	70.10	48.79	**81.28**	0.18	39.27
Usage	32.41	36.23	31.80	**55.22**	16.86	40.58	2.10	30.44
Other	35.29	35.18	39.36	**52.83**	11.93	35.19	0.62	18.52

Table 4: Model Performance and Hallucination Rate in Random Scenarios.Accuracy refers to the evaluation result between the model's response and the ground truth. HR_{ran} is defined as the hallucination rate from random scene pairs (see Section 5.2).

In the **Change Question** experiment, we first select questions involving spatial relationships and use GPT-4 to transform each QA pair into a format where the answer is a single object, resulting in the *ScanQA-SR* dataset. For each question Q_i in *ScanQA-SR*, we generate its opposite Q_j (e.g., by reversing the spatial relation) to form the pair: $(\{Q_i, Q_j\}, A_i, S_i)$ which constitutes the *ScanQA-SR-Opposite* dataset.

To ensure that Q_j does not apply to the same answer A_i in scene S_i, we extract the spatial relation triplet implied by (Q_j, A_i) and verify that it does not exist in the scene graph G_i. Formally, we require: $(object_1, object_2, relation) \notin G_i$. This guarantees that Q_i and Q_j yield different answers within the same scene context.

Experiment We evaluate different models using the proposed benchmark.

In **Experiment 1**, given a question q_i, we generate two answers a_{ij} and a_{ik} from two different scenes s_j and s_k, respectively. To measure the semantic similarity between answers, we use BLEU-4 (Papineni et al., 2002) (n-gram precision), ROUGE-L (Lin, 2004) (longest common subsequence), and METEOR (Banerjee and Lavie, 2005) (semantic alignment with synonym matching).

Based on human-verified answers from the ScanQA test set, the average ROUGE-L and METEOR score are 0.71 and 0.49 respectively. Therefore, we consider two answers to be semantically equivalent if ROUGE-L > 0.71 and METEOR > 0.49. The hallucination rate is defined as:

$$HR_{\text{ran}} = \frac{1}{N} \sum_i \mathbf{1} \left(\text{ROUGE-L}(a_{ij}, a_{ik}) > 0.71 \right. \\ \left. \text{and METEOR}(a_{ij}, a_{ik}) > 0.49 \right) \quad (6)$$

In **Experiment 2**, for a fixed scene s_i, we generate answers a_{ji} and a_{ki} for two semantically opposite questions q_j and q_k. The hallucination rate is computed as:

$$HR_{\text{opp}} = \frac{1}{N} \sum_i \mathbf{1} \left(\text{ROUGE-L}(a_{ij}, a_{ik}) > 0.71 \right. \\ \left. \text{and METEOR}(a_{ij}, a_{ik}) > 0.49 \right) \quad (7)$$

6 Evaluation on 3D-LLMs

6.1 Hallucinations in Random Scene Queries

We evaluate four models using the approach above. Table 4 presents the results for random scenes. ROUGE-L measures performance on ScanQA, while HR_{ran} is defined in Section 5.2. The table shows a positive correlation between accuracy and hallucination rate. LL3DA , 3DLLM, LEO and LlaVA3D all exhibit low accuracy and hallucination rates for spatial questions but higher rates for object attributes.

For instance, models with higher ROUGE-L scores often exhibit higher hallucination rates.Specifically, LL3DA achieves the highest accuracy for size-related questions, 3D-LLM for color-related questions, and LEO for quantity-related questions; however, each model also exhibits the highest hallucination rate in its respective category. This pattern suggests that higher accuracy does not necessarily correlate with a deeper understanding of the relationship between the questions and the point clouds. These findings indicate that the models exhibit significant hallucination issues, where it answers questions without considering the visual context, yet its responses appear 'better' or closer

Model	ScanQA		ScanQA-SR		ScanQA-SR-Opposite		
	ROUGE-L	METEOR	ROUGE-L	METEOR	ROUGE-L	METEOR	HR_{opp}%
LL3DA	36.56	26.95	4.12	28.27	50.25	52.94	46.52
3D-LLM	37.46	28.18	15.55	10.28	60.78	56.22	53.80
LEO	22.85	16.08	18.46	12.97	66.52	61.13	**62.12**
LLaVA-3D	3.29	16.71	3.87	28.42	61.03	58.51	56.82

Table 5: This table compares model performance across three tasks: ScanQA, ScanQA-SR (spatial questions), and ScanQA-SR-Opposite. It uses RougeL and Meteor to measure similarity between model responses and ground truth (GT) in ScanQA and ScanQA-SR. For ScanQA-SR-Opposite, higher RougeL and Meteor scores indicate a higher probability of the model generating the same response for opposite spatial questions, reflecting a higher hallucination rate.

to the ground truth. Upon examining the training set, we find that object attributes often align with typical characteristics—for example, tables are usually black, white, or brown, and televisions are typically rectangular. This indicates that the model learns attribute associations due to the homogeneous nature of indoor scenes and the limited diversity of attributes.

6.2 Relationship Between Attribute Uniformity and Answer Accuracy

We plotted Figure 5 to illustrate the relationship between the uniformity of an object's properties and the accuracy of the answers. For instance, chair color is queried 346(N) times, with black (T_1 times), brown (T_2 times), and gray (T_3 times) as the most frequent colors. To quantify attribute uniformity, we introduce the "Top-K Ratio," where the Top-3 Ratio for the chair can be calculated as:

$$\text{Top-3 Ratio} = \frac{T_1 + T_2 + T_3}{N}. \tag{8}$$

The x-axis shows the average ROUGE-L score for questions about a specific object, reflecting how easily its properties can be correctly answered. The three plots (color, shape, size) illustrate that answer accuracy increases with property uniformity—especially for color and shape, where a clear linear trend appears. Many points cluster near a Top-3 Ratio of 1, indicating that the dataset contains objects with highly uniform attributes, which may lead the model to hallucinate correct answers more easily.

6.3 Hallucinations in Opposite-Question Queries

The results for testing with opposite questions within the same scene are presented in Table 5.The ScanQA dataset includes a wide range of QA pairs involving various attributes, spatial relationships, and other data types. In contrast, ScanQA-SR focuses solely on spatial relationships and transforms all QA pairs into those where the answer is the object itself.

By comparing the results from these two datasets, we observe that the ROUGE scores for ScanQA-SR are significantly lower than those for ScanQA. This indicates that the model is more prone to errors when dealing with spatial relationship tasks. To investigate whether the model truly understands the meaning of spatial relationships, we created a dataset of opposite questions specifically for spatial relationships. The goal was to assess the model's ability to handle questions about opposing spatial positions.

However, we found that the hallucination rate for both models exceeded 50%. This suggests that when posed with opposite questions about the same scene, the model has a 50% chance of giving the same answer. This result further supports our earlier observation that the model is prone to errors and hallucinations when handling spatial relationship queries. The results imply that the model may lack a proper visual-semantic understanding of spatial relationships, leading it to answer incorrectly without considering point cloud data.

7 Conclusion

This study categorizes 3D hallucinations and assesses their severity in 3DLLM, LL3DA, LEO and LLaVA3D using description and QA tasks.We find that high object frequency, strong correlations, and attribute uniformity drive hallucinations. Since existing metrics rely on text similarity, we design two experiments to better define hallucinations and investigate whether models truly use and understand visual information when answering correctly. Results show that models often fail to answer contextually accurate questions and struggle with aligning spatial relationships to visual input.

8 Limitations

In this study, we provide a detailed classification of hallucination types specifically for the QA task. Each QA pair is classified to detect corresponding hallucinations. However, for the description task and other long-text tasks, no specific approach is proposed to detect the types of hallucinations present in the generated answers. This limitation means that our evaluation only demonstrates the significant hallucination issues within 3D point cloud models, and uses different types of short QA pairs to explore the following questions: 1) Which types of questions are more likely to induce hallucinations in the model? 2) How does the dataset distribution impact the occurrence of hallucinations in the model?

Furthermore, we identify that models are particularly prone to attribute hallucinations and investigate the relationship between dataset distribution and hallucination rates. Regarding spatial relationship hallucinations, our experiments only reveal that the models lack understanding of spatial relationships, but do not explain why the models perform worse on spatial relationship-related questions compared to other question types.

Third, in our experiments designed to explore whether the models answer based on visual information or rely on textual inputs alone, the results indicate that the current dataset is overly simple and highly regular, which enabling models to neglect point cloud information in favor of answering based on text alone. However, we do not provide insights into why the models do not incorporate point cloud information in their responses from an architectural perspective.

Finally, we utilize GPT-4 to generate a new annotated dataset, which, compared to manual annotation, may contain some minor errors. Although we have discussed hallucination issues in 3D large language models and highlighted the problem of models not responding based on point cloud data, this highlights current limitations, but we believe the field holds strong potential. On the contrary, we aim to identify the reasons behind their suboptimal performance, such as the dataset distribution issues discussed in this paper. We hope that our work can provide new insights and ideas for further improving the performance of 3D large language models.

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. 2024. Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26428–26438.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*.

Katja Filippova. 2020. Controlled hallucinations: Learning to generate faithfully from noisy data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 864–870, Online. Association for Computational Linguistics.

Ronald A Fisher. 1936. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385.

Pei Guo, Wangjie You, Juntao Li, Yan Bowen, and Min Zhang. 2024. Exploring reversal mathematical reasoning ability for large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 13671–13685.

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. *Advances in Neural Information Processing Systems*, 36:20482–20494.

Yining Hong, Zishuo Zheng, Peihao Chen, Yian Wang, Junyan Li, and Chuang Gan. 2024. Multiply: A multisensory object-centric embodied large language model in 3d world. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, pages 26406–26416.

Hongyu Hu, Jiyuan Zhang, Minyi Zhao, and Zhenbang Sun. 2023. Ciem: Contrastive instruction evaluation method for better instruction tuning. *arXiv preprint arXiv:2309.02301*.

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 preprint arXiv:2311.12871*.

Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Alaa Maalouf, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, et al. 2023. Conceptfusion: Open-set multimodal 3d mapping. *arXiv preprint arXiv:2302.07241*.

Weisen Jiang, Han Shi, Longhui Yu, Zhengying Liu, Yu Zhang, Zhenguo Li, and James Kwok. 2024. Forward-backward reasoning in large language models for mathematical verification. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6647–6661.

Aditya Kanade, Petros Maniatis, Gogul Balakrishnan, and Kensen Shi. 2020. Learning and evaluating contextual embedding of source code. In *International conference on machine learning*, pages 5110–5121. PMLR.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026.

Hung Le, Doyen Sahoo, Chenghao Liu, Nancy F Chen, and Steven CH Hoi. 2020. Uniconv: A unified conversational neural architecture for multi-domain task-oriented dialogues. *arXiv preprint arXiv:2004.14307*.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882.

Chuyuan Li, Yuwei Yin, and Giuseppe Carenini. 2024. Dialogue discourse parsing as generation: A sequence-to-sequence llm-based approach. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 1–14.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*.

Yunhang Shen, Chaoyou Fu, Peixian Chen, Mengdan Zhang, Ke Li, Xing Sun, Yunsheng Wu, Shaohui Lin, and Rongrong Ji. 2024. Aligning and prompting everything all at once for universal visual perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13193–13203.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Yue Wang, Weishi Wang, Shafiq Joty, and Steven CH Hoi. 2021. Codet5: Identifier-aware unified pre-trained encoder-decoder models for code understanding and generation. *arXiv preprint arXiv:2109.00859*.

Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. 2024. Pointllm: Empowering large language models to understand point clouds. In *European Conference on Computer Vision*, pages 131–147. Springer.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. 2023. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*.

Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. 2024. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12944–12953.

Bohan Zhai, Shijia Yang, Chenfeng Xu, Sheng Shen, Kurt Keutzer, and Manling Li. 2023. Halle-switch: Controlling object hallucination in large vision language models. *arXiv e-prints*, pages arXiv–2310.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. 2022. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*.

Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 2024. 3d-vla: A 3d vision-language-action generative world model. *arXiv preprint arXiv:2403.09631*.

Duo Zheng, Shijia Huang, Lin Zhao, Yiwu Zhong, and Liwei Wang. 2024. Towards learning a generalist model for embodied navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13624–13634.

9 Appendix

9.1 Introduction of GPT-4o Prompts

In this experiment, we employed GPT-4o to analyze existing textual data and generate new data samples. The corresponding prompts used in each sub-task are as follows:

1. **Object Hallucination:** To extract objects mentioned in the model's responses, we used the following prompt:

   Provide a description to list the items in a room, ensuring the output is in singular form. For example, if the description is: "The room is a well-organized space with a floor, walls, and a ceiling," the output should be: [floor, wall, ceiling]. Just provide the list of items, no explanation is needed. Given the following room description: description

2. **Question Categorization:** To enable a more fine-grained analysis of model performance, we categorized all question-answer pairs using the following prompt:

   Given a question, please determine the type of the question without answering it. Choose the question type from the following options: [Spatial Relationship, Size Comparison, Object's Properties (color, size, shape), Quantity, Usage of an Object, Other]. Please do not provide an answer outside of the listed options. The question is as follows: question.

3. **Opposite Question Evaluation – Question Generation:** To evaluate spatial understanding, we generated new questions by reversing spatial relationships using the following prompt:

   Give a question, such as "What is on the front of the brown table?" and change the spatial relationship in the question to the exact opposite, for example, change it to "What is behind the brown table?" Just provide the modified result without explanation. The question is as follows: question

4. **Opposite Question Evaluation – Triple Extraction:** To verify whether the spatial relationships in generated question-answer pairs exist in the scene, we used semantic scene graphs, which represent relationships in the form of triples (object1, object2, relation). To convert the QA pairs into such triples, we used the following prompt:

   Provide a question and answer list pair related to spatial relationships. Based on the question and answer, abstract a triple (Item 1, Item 2, Relationship). The question and answer list are as follows: question answer list: answer.