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

Jiangyong Huang[1,2,*]    Baoxiong Jia[1,*]    Yan Wang[1]    Ziyu Zhu[1,3]    Xiongkun Linghu[1]
Qing Li[1]    Song-Chun Zhu[1,2,3]    Siyuan Huang[1]

[1]State Key Laboratory of General Artificial Intelligence, BIGAI
[2]Peking University, [3]Tsinghua University
https://beacon-3d.github.io

Figure 1. **An overview of BEACON3D, a novel benchmark for 3D grounding and question answering (QA) tasks.** BEACON3D features an object-centric evaluation framework, with Grounding-Chains (G-Chains) and Grounding-QA-Chains (GQA-Chains) for each object. The evaluation adopts object-centric metrics to ensure robustness and utilizes chain-of-analysis for studies in task coherence. We also involve the study of various knowledge types such as class, appearance ("App."), spatial ("Spa."), and geometry ("Geo.").

## Abstract

*Existing 3D vision-language (3D-VL) benchmarks fall short in evaluating 3D-VL models, creating a "mist" that obscures rigorous insights into model capabilities and 3D-VL tasks. This mist persists due to three key limitations. First, flawed test data, like ambiguous referential text in the grounding task, can yield incorrect and unreliable test results. Second, oversimplified metrics such as simply averaging accuracy per question answering (QA) pair, cannot reveal true model capability due to their vulnerability to language variations. Third, existing benchmarks isolate the grounding and QA tasks, disregarding the underlying coherence that QA should be based on solid grounding capabilities. To unveil the "mist", we propose BEACON3D, a benchmark for 3D-VL grounding and QA tasks, delivering a perspective shift in the evaluation of 3D-VL understanding. BEACON3D features (i) high-quality test data with precise and natural language, (ii) object-centric evaluation with multiple tests per object to ensure robustness, and (iii) a novel chain-of-analysis paradigm to address language robustness and model performance coherence across grounding and QA. Our evaluation of state-of-the-art 3D-VL models on BEACON3D reveals that (i) object-centric evaluation elicits true model performance and particularly weak generalization in QA; (ii) grounding-QA coherence remains fragile in current 3D-VL models, and (iii) incorporating large language models (LLMs) to 3D-VL models, though as a prevalent practice, hinders grounding capabilities and has yet to elevate QA capabilities. We hope BEACON3D and our comprehensive analysis could benefit the 3D-VL community towards faithful developments.*

*Equal contribution.

## 1. Introduction

The ability to understand 3D scenes is an essential facet of human-level intelligence [9, 30, 57, 64, 97]. Recent 3D vision-language (3D-VL) models have achieved notable progress in language-grounded 3D scene understanding [7, 8, 22, 25, 27, 29, 35, 51, 98, 99], and various benchmarks have been established for 3D-VL tasks like object grounding [2, 5, 35, 78, 81, 91] and question answering (QA) [4, 24, 50, 52]. Despite the improving performance on these benchmarks, a critical question remains to be addressed:

*How effective are these benchmarks for 3D-VL understanding; are the progress and results on these benchmarks reliable enough to guide the development of 3D-VL models?*

We raise considerable concerns on this question, observing several key limitations in existing 3D-VL benchmarks:

- First, we observe notable flaws in the test data, which may undermine the reliability of evaluations. For example, referential text in the grounding task can be *ambiguous* or *unnatural*, leading to ill-posed tests; *ambiguous questions* in QA data may mislead to divergent answers; *incomplete answer labels* can misrepresent model performance by penalizing correct predictions. Our human studies highlight these flaws in ScanRefer [5] and ScanQA [4], as validated by the limited human performance. Additionally, we show that addressing the flaws in ScanRefer can lead to a more accurate evaluation of model performance.

- Second, the evaluation metrics in current 3D-VL benchmarks fall short in accurately capturing model capability. Oversimplified metrics, such as averaging accuracy over individual QA pairs, are vulnerable to model pitfalls like *visual ignorance* (*i.e.*, predictions determined solely by texts) and *weak language robustness* (*i.e.*, predictions susceptible to varied texts). We demonstrate their vulnerability by showing that blind LLMs can achieve unexpectedly high accuracy on SQA3D [50], and even minor language rephrasing can significantly affect QA accuracy. This suggests the need for more robust evaluation metrics through language variations and multiple tests for each object.

- Third, current 3D-VL benchmarks isolate grounding and QA tasks, exposing QA in the risk of shortcuts. To address this gap, we design Grounding-QA-Chains (GQA-Chains) to assess model performance coherence between grounding and QA. These chains ensure that the contents of QA are covered by corresponding grounding texts. Our study on GQA-Chains reveals two types of broken coherence: (i) *model correctly grounds the object but fails in QA*, showing poor QA skills; and (ii) *model fails in grounding but succeeds in QA*, suggesting shortcuts in QA. Specifically, on a state-of-the-art 3D-VL model PQ3D [99], we observe that half of QA errors are associated with correct grounding predictions, while one-quarter of correct answers result from shortcuts. This implies the potentially fragile grounding-QA coherence in 3D-VL models.

Motivated by our analyses, we construct BEACON3D, a novel benchmark for 3D-VL grounding and QA tasks, providing a new perspective in 3D-VL evaluation. The benchmark is built on 30 meticulously selected high-quality scenes from ScanNet [14], 3RScan [75], and MultiScan [55]. We exhaustively annotate objects in each scene and introduce object-level evaluation with three cases per object for both grounding and QA. This yields more robust and reliable object-centric metrics, reflecting the true model capabilities. Additionally, we propose Grounding-Chains (G-Chains) for the grounding task, spanning grounding texts from coarse (*e.g.*, "chair") to fine-grained (*e.g.*, "gray chair next to the corner table") descriptions. To address the isolation of grounding and QA tasks, we further construct GQA-Chains associated with G-Chains to assess model performance coherence across grounding and QA tasks. BEACON3D comprises a total of 837 objects, 2511 G-Chains and 2511 GQA-Chains, with all annotations manually crafted for language clarity and naturalness. We employ object-centric evaluation metrics that require accurate predictions across all three tests per object for grounding and QA, helping to better manifest model pitfalls. The G-Chains and GQA-Chains also enable a novel chain-of-analysis evaluation paradigm in BEACON3D, providing a holistic assessment of 3D-VL model capabilities.

We apply BEACON3D to evaluate state-of-the-art 3D-VL models. Compared to conventional per-case averages, object-centric metrics elicit a significant model performance drop in both grounding and QA. This highlights that models are prone to language variations and exhibit a limited object-level understanding. Analyses on G-Chains show that models struggle when the granularity of grounding texts increases. And analyses on GQA-Chains reveal a fragile grounding-QA coherence in 3D-VL models, underscoring the gap between grounding and QA skills, and the prevalence of shortcuts in 3D QA. Furthermore, contrary to existing practices [8, 25, 29, 65], our results show that incorporating LLMs for 3D-VL models hinders grounding and has yet to improve QA performance on BEACON3D, offering new insights into the learning of grounding and QA tasks.

We summarize our contributions as follows:

1. We present detailed investigations into limitations of existing 3D-VL benchmarks and expose fragile performance coherence across grounding and QA in 3D-VL models.

2. We propose BEACON3D, a benchmark for 3D grounding and QA that shifts the evaluation paradigm to object-centric evaluation with chain-of-analysis on grounding and grounding-QA chains, providing a high-quality, faithful, and holistic tool for evaluating 3D-VL models.

3. We present a comprehensive analysis of state-of-the-art 3D-VL models on BEACON3D, highlighting common model pitfalls like grounding-QA incoherence and incomplete object understanding, along with the unexpected hindrance of LLM for 3D-VL tasks.

Figure 2. **Various types of test data flaws in** ScanRefer, Nr3D, ScanQA. Underlined texts indicate explicit flaws. (1) The top row shows grounding data with the target object highlighted. **Ambiguous text** includes viewpoint-dependent expressions like "left" and "right", or lacks information to uniquely specify the target object. **Unnatural descriptions** are hard to understand by humans for being too tedious or grammatically invalid. **Incorrect annotation** refers to the mismatch between text and target object. (2) The bottom row shows QA data with ground truth (GT) shown in square brackets. **Ambiguous question** lacks context to clarify the queried object, potentially leading to contradictory answers. **Incomplete answers** may forbid alternative correct answers.

## 2. Related Work

**3D vision-language models.** Fueled by the advancement of vision-language models (VLMs) [21, 28, 38, 39, 60, 66, 88] and reconstruction techniques [10, 34, 45–47, 58, 59, 76, 77, 85], the capability of 3D scene understanding has been greatly improved. Key contributions in this area include 3D perception techniques [1, 7, 31, 47, 62, 63, 69, 79, 93], 2D-3D feature integration [23, 33, 37, 61, 83, 99], and 3D-VL pretraining [18, 35, 78, 80, 95, 98]. On the other hand, the rapid development of large vision-language models (LVLMs) [15, 40, 43] drives 3D-VL models to evolve from task-specific architectures to generalist frameworks [8, 13, 20, 25, 27, 29, 36, 82, 92, 96]. While these 3D LVLMs demonstrate impressive capabilities, there is also a pressing demand for advanced benchmarks to comprehensively evaluate these models, and address underexplored questions, *e.g.*, generalizability and the effect of LLMs.

**3D vision-language datasets and benchmarks.** Early research in 3D-VL learning has produced initial task-specific benchmarks for grounding [2, 5, 91] and QA [4, 24, 50, 84], akin to the early stage of 2D vision-language (2D-VL) benchmarks [3, 32, 54, 56, 70, 86]. As recent LVLMs evolve to be more powerful and intricate, 2D vision-language (VL) benchmarks have advanced towards meticulously designed evalua-

tion or detailed analysis [6, 19, 41, 43, 44, 68, 74, 87, 89, 90]. In contrast, recent 3D-VL works mainly focus on large-scale learning [29, 35, 42, 48, 49, 78, 98] while adhering to conventional evaluation criteria [2, 4, 5, 50]. On the other hand, recent advance in the evaluation of 3D-VL models [52, 53, 71–73, 94] provides suites for analyzing issues such as hallucination and robustness [16, 36, 81]. Nonetheless, prior works have not established an evaluation criterion with reliable metrics and in-depth analysis of 3D grounding and QA tasks, which is the exact goal of this paper.

## 3. An Investigation into 3D-VL Benchmarks

### 3.1. Flawed Test Data

When examining existing 3D-VL benchmarks, we identified flaws in the test data as a significant issue for evaluating model performance. We provide justifications from both quantitative and qualitative aspects as follows:

**Qualitative analysis.** We analyze the test data quality from prevalent 3D-VL benchmarks: ScanRefer [5] and Nr3D [2] for grounding, and ScanQA [4] for 3D-QA. We identify common data flaws, shown in Fig. 2. Key grounding issues include: (i) *ambiguous referential text*, which lacks information to uniquely identify the target object; and (ii) *unnatural descriptions*, being excessively complex, that are difficult

Table 1. **Human study on ScanRefer val set.** We report clarity and naturalness scores (1∼5) of the referential text, as well as human and model prediction accuracy. We use PQ3D [99] for model evaluation.

| Data Source | Clarity | Naturalness | Human Accuracy | Model Accuracy |
|---|---|---|---|---|
| ScanRefer | 3.70 | 4.23 | 69% | 63% |
| Refined | 4.59 | 4.34 | **100%** | **70%** |

Table 2. **Human study on ScanQA (val) and SQA3D (val and test).** Quality scores range from 1 to 5. Human accuracy is evaluated using answer labels as the ground truth.

| Data Source | Question Quality | Answer Quality | Human Accuracy |
|---|---|---|---|
| ScanQA | 3.44 | 3.60 | 62% |
| SQA3D | 4.64 | 4.46 | 80% |



Figure 3. **Illustrative examples on visual ignorance.** The model predicts answers directly from questions, ignoring scene information (*e.g.*, chair color).



Figure 4. **Illustrative examples on language robustness.** Rephrased and more detailed questions of the same concept can easily lead to wrong model predictions.

to identify the target object. For 3D-QA, we observe that (i) *ambiguous questions* with no clear targeting object easily leads to contradictory answers, and (ii) questions with *incomplete answers* can undermine evaluation reliability by forbidding alternative valid answers predicted by the models.

**Quantitative analysis.** We provide quantitative measurements of data flaws and their impacts. For grounding, we sample a subset of 100 grounding texts from the ScanRefer validation set and instruct human evaluators to re-predict the target object based on the referential text and score the *clarity* and *naturalness* of each text (scored from 1 to 5). As shown in Tab. 1, a large portion (31%) of the test data leads to incorrect human predictions. We test a recent state-of-the-art 3D-VL model, PQ3D [99], before and after manually refining these texts. We observe a significant model performance improvement (7%) without model-side adjustments.

For QA, we also randomly sample 100 QA pairs from ScanQA and SQA3D [50]. We instruct human evaluators to re-answer the questions and rate the quality of the QA text. As shown in Tab. 2, the low human prediction accuracy (62% on ScanQA) highlights that *the flaws in QA data pose a tangible upper bound on model performance*. These analyses on existing grounding and QA benchmark underscore the need for rigorous quality control in 3D-VL benchmarks.

### 3.2. Insufficient Evaluation Metrics

In this section, we show that simple metrics like average accuracy over all test instances in existing 3D-VL benchmarks are insufficient to reveal true model pitfalls including *visual ignorance* and poor *language robustness*:

- **Visual ignorance** refers to the scenario where models can perform tasks without the need for visual input, as illustrated in Fig. 3. As an example, we show in Tab. 3

Table 3. **Blind LLMs finetuned with LoRA on SQA3D.** [†] indicates the performance of state-of-the-art 3D-VL model [96].

| Blind LLM | OPT-1.3B | Gemma2-2B | Vicuna-7B | LLaMA3-3B | LLaVA-3D[†] |
|---|---|---|---|---|---|
| EM-1 | 43.9 | 48.8 | 49.4 | 50.0 | 55.6 |

that fine-tuning "blind" LLMs yields a comparable result on SQA3D metrics compared to state-of-the-art 3D-VL models. This indicates a deficiency in SQA3D's metrics for evaluating the visual capability of 3D-VL models.

- **Language robustness** refers to a model's susceptibility to language variations. For example, in QA (see Fig. 4), models often struggle with *rephrased* or *more detailed* questions about the same object concept (*e.g.*, chairs). We demonstrate this by rephrasing good questions sampled in Sec. 3.2 and comparing PQ3D's performance on the rephrased sets versus the original sets. The results in Fig. 5(b,c) show model sensitivity to language variations do exist, especially on SQA3D where 16% of predictions switch from correct to incorrect. However, such a problem is overlooked with current 3D-VL benchmarks treating these variations as separate instances during evaluation.

To prevent lingual shortcuts arising from *visual ignorance*, we need careful data curation to avoid scene-irrelevant questions and introduce vision-oriented metrics to assess models' visual capability. To better evaluate *language robustness* of models, we need robust evaluation frameworks that incorporate language variations and multiple evaluation instances per object. Thus, we argue that 3D-VL benchmarks must evolve to better visualize these crucial dimensions of 3D-VL model performance.

### 3.3. Grounding-QA Coherence

During our exploration, one critical question we identified, yet has been overlooked by existing benchmarks, is: *Why do*

Figure 5. **(a) Illustration of GQA-Chains.** The questions derive from the grounding text and query a specific feature of the target object. We define two broken types for grounding-QA coherence: (Type 1) correct grounding and incorrect QA, indicating a lack of QA skills; (Type 2) incorrect grounding and correct QA, suggesting shortcuts in QA. **(b) The effect of rephrasing ScanRefer texts on the performance of PQ3D. (c) The effect of rephrasing SQA3D questions on the performance of PQ3D. (d) Results of PQ3D on GQA-Chains.** We observe over half of QA failures (24% out of 46%) stem from insufficient QA skills while nearly a quarter of correct QA predictions (14% out of 54%) are achieved via shortcuts.

*models fail in 3D-QA tasks; is it due to language complexity or inadequate scene understanding capabilities?* Believing that accurate QA predictions should be grounded in strong scene understanding, we propose a novel Grounding-QA-Chain (GQA-Chain) that connects grounding and QA evaluations to provide detailed analyses of model performance coherence across tasks. The core idea behind GQA-Chains is to align questions with referential descriptions, ensuring the queried content is directly present in the descriptive texts. For example, in Fig. 5(a), the questions ask about the appearance, geometry, and spatial relationships of the target object, all of which are explicitly described in the referential texts.

With the expectation that strong 3D-VL should exhibit consistent performance across grounding-QA pairs in GQA-Chains, we generate GQA-Chains based on the refined ScanRefer subset from Sec. 3.1 as a preliminary experiment. We evaluate PQ3D on both datasets and visualize the results in Fig. 5(d). We observe that **over half of QA failures stem from insufficient QA skills while nearly a quarter of correct QA predictions are achieved via shortcuts**. These findings suggest the prevalence of broken grounding-QA coherence in 3D VL models, as well as the demand for benchmarks to systematically evaluate grounding-QA coherence.

## 4. The BEACON3D Benchmark

In this section, we introduce BEACON3D, a novel benchmark for 3D-VL grounding and QA tasks that addresses key evaluation limitations identified in Sec. 3. We propose the formats of Grounding-Chain (G-Chain) and Grounding-QA-Chain (GQA-Chain) for organizing grounding and QA data, along with an object-centric chain-of-analysis paradigm that evaluates models' performance coherence under language

variations and across tasks using object-centric metrics.

### 4.1. Benchmark Design

**Data Design** We consider two tasks in BEACON3D: (i) 3D grounding, where models are required to predict the target object's 3D bounding box given the scene point cloud and object referential texts; and (ii) 3D-QA, where models are required answer a question about a target object based on the scene point cloud. The data for these two tasks consists of:

- **Grounding:** we create G-Chain that consists of a series of referential texts, ranging from coarse to fine. At its finest level, the primary grounding text uniquely identifies the target object. It is then rephrased into progressively coarser texts at each subsequent level, referred to as simplified grounding texts (see in Fig. 1). This relaxation in object descriptions expands the set of correct objects for simplified ground texts at each level, requiring model predictions to fall within its set for correctness evaluation.

- **Question Answering:** As in Sec. 3.3, we construct GQA-Chains by designing QA pairs based on the primary grounding texts in G-Chains. Each answer in a GQA-Chain question is explicitly present in the corresponding primary grounding text. To provide a holistic evaluation, similar to other benchmarks, and accommodate questions that require commonsense knowledge, we also curate a set of questions with queried content not explicitly found in the primary grounding texts. We tag these questions with an *"extra knowledge"* flag and exclude them from the coherence analysis.

In addition, we tag each grounding and QA data with its required knowledge types: `class` (semantic category), `appearance` (color, material, texture, *etc*.), `geometry` (shape, size, *etc*.), and `spatial-relation`. An extra

Figure 6. **Human study on grounding data.**



Figure 7. **Human study on QA data.**



Figure 8. **Data statistics in** BEACON3D.

knowledge type `exist` is added to QA for the questions about whether something exists. Each QA data is assigned a single knowledge type according to its *queried content*.

**Data Collection** We begin data collection by selecting high-quality scenes from the held-out sets of ScanNet [14], 3RScan [75], and MultiScan [55] following two principles: (1) the layout should be reasonable, neither overly cluttered nor too simple, with clear object mesh reconstructions; and (2) objects should be well-placed in the scene with balanced distribution over categories. This results in 30 high-quality scenes in diverse styles from the three datasets. Next, we identify potential target objects by excluding: (i) background objects like walls and floors, (ii) objects that are difficult to distinguish via text (*e.g.*, multiple chairs around a table), and (iii) objects with comparatively low-quality reconstructions, resulting in 837 unique target object instances. We then build an annotation tool following [50] (see details in the *Appendix*) for human annotators to annotate three G-Chains and GQA-Chains for each object instance, totaling 2511 G-Chains and 2511 GQA-Chains. To address prior data flaws, we establish detailed annotation guidelines, ensuring precise and natural language, the indispensability of visual modality in QA, and also balanced answer distributions. Each annotation is cross-validated by two human reviewers.

**Metrics** In addition to the conventional per-case average metrics, we adopt an object-centric evaluation scheme, requiring models to accurately predict over **all three** grounding or QA test cases per object. Our task-specific metrics are computed as follows:

- **Grounding:** For each grounding text, the model is considered correct if the predicted object is included within the candidate object set. For the object-centric metrics, we first derive per-object results according to whether **all three** predictions on the **primary grounding texts** are correct, and then average the results over all objects. We

also report per-case metrics by averaging the results over all **primary and simplified grounding texts**.

- **Question Answering:** We first evaluate each QA pair using GPT-Score [52], yielding a score $M$ between 1 to 5 from GPT-4 [60]. The corresponding per-case accuracy is then calculated as $\frac{M-1}{4}$ following [52]. We derive a binary per-object accuracy if $M \geq 4$ for **all three** QA pairs. We report object-centric metrics by averaging per-object accuracies, as well as per-case average accuracy over all **individual QA pairs**.

## 4.2. Data Quality Check and Statistics

To assess the quality of the data collected in BEACON3D, we have a separate group of human annotators evaluate it based on clarity, naturalness, and human accuracy, following metrics used in Sec. 3.1. For a fair comparison, we sample the same quantity of data from the same scenes. As shown in Fig. 6 and 7, BEACON3D significantly outperforms existing 3D grounding and QA benchmarks in terms of language clarity, naturalness, and especially human accuracy metric where nearly ∼95% of the data labeled as correct upon reexamination. We also visualize the statistics of BEACON3D in Fig. 8, including object counts by domains, knowledge types, data counts by knowledge types, and the proportion of QA pairs requiring *extra knowledge*.

## 5. Experiments

Our experiments aim to address the following questions:

- How does the object-centric evaluation scheme differ from conventional case-centric metrics in revealing model performance? (Sec. 5.1)
- How do models perform when handling language variations in the G-Chains? (Sec. 5.2)
- Do models show performance coherence between grounding and QA on GQA-Chains? (Sec. 5.2)
- Do LLMs affect the model performance? (Sec. 5.3)

Table 4. **Evaluation results of grounding on BEACON3D.** The "Obj." column reports object-centric metrics. The columns of knowledge types report per-case averages over each type.

| | Knowledge type | | | | Overall | |
|---|---|---|---|---|---|---|
| | Class | App. | Geo. | Spa. | Case | Obj. |
| *w/o LLM* | | | | | | |
| ViL3DRel [7] | 61.8 | 66.9 | 46.5 | 59.5 | 61.8 | 39.8 |
| 3D-VisTA [98] | 71.0 | 64.6 | 56.3 | 68.9 | 71.0 | 50.9 |
| PQ3D [99] | **76.1** | **71.2** | **66.0** | **74.5** | **76.1** | **57.2** |
| SceneVerse [35] | 73.4 | 64.9 | 64.6 | 71.9 | 73.5 | 52.1 |
| *LLM-based* | | | | | | |
| LEO-multi | 14.3 | 10.9 | 15.3 | 15.1 | 14.3 | 2.8 |
| LEO-curricular | 22.0 | 22.2 | 20.8 | 15.4 | 22.0 | 3.8 |
| PQ3D-LLM | 70.3 | 66.2 | 53.5 | 68.3 | 70.2 | 47.4 |
| Chat-Scene [27] | 62.7 | 57.3 | 56.3 | 57.8 | 62.7 | 44.3 |

Table 5. **Evaluation results of QA on BEACON3D.** Object-centric metrics ("Obj.") are drastically lower than case-centric metrics. [†] indicates text input (*i.e.*, object locations and attributes) instead of 3D point cloud.

| | Knowledge type | | | | | Overall | |
|---|---|---|---|---|---|---|---|
| | Class | App. | Geo. | Spa. | Exi. | Case | Obj. |
| *w/o LLM* | | | | | | | |
| 3D-VisTA [98] | 20.5 | 33.5 | 52.1 | 33.8 | 36.5 | 35.3 | 8.1 |
| PQ3D [99] | **36.4** | 28.0 | 27.8 | 11.9 | 45.5 | 27.8 | 3.5 |
| SceneVerse [35] | 35.6 | 41.7 | 48.9 | 41.9 | 35.7 | 40.3 | 6.6 |
| *LLM-based* | | | | | | | |
| GPT-4o[†] [60] | 33.3 | **49.9** | 54.9 | **52.1** | **73.8** | **57.1** | **20.2** |
| LEO-multi | 25.8 | 37.7 | 52.8 | 46.2 | 37.4 | 41.1 | 3.5 |
| LEO-curricular | 17.4 | 41.0 | 53.2 | 48.7 | 39.7 | 43.2 | 7.8 |
| PQ3D-LLM | 28.0 | 30.8 | 35.2 | 25.2 | 26.2 | 27.9 | 2.3 |
| Chat-Scene [27] | **36.4** | 39.8 | **56.7** | 47.6 | 48.8 | 45.8 | 7.8 |

To explore these questions, We select a variety of state-of-the-art 3D-VL models as baselines, categorizing them based on their use of LLM. We make the necessary adjustments to ensure that most baselines can handle both grounding and QA tasks with the same set of model weights (see implementation details in *Appendix*). Specifically, we consider the following baseline categories in our experiments:

- **Without LLM.** This category includes four baselines: ViL3DRel [7], 3D-VisTA [98], PQ3D [99], and SceneVerse [35]. ViL3DRel is selected as a grounding specialist and evaluated using its original checkpoint. For 3D-VisTA, we multi-task fine-tune the model to make it a generalist capable of handling both grounding and QA tasks. For PQ3D, we directly use its pre-trained checkpoint as it is already a generalist model. For SceneVerse, we freeze the backbone pre-trained for grounding and add an additional head for fine-tuning it on the QA task.
- **LLM-based.** This category includes five models: GPT-4o [60], LEO-multi, LEO-curricular, PQ3D-LLM, and Chat-Scene [27]. GPT-4o is prompted with object lists with locations and attributes for question answering. The object attributes are sourced from MSQA [42], which were generated using GPT-4V. LEO-multi and LEO-curricular are implemented by extending LEO [29] to grounding through contrastive learning between object tokens and language embeddings. LEO-multi is trained with both tasks jointly while LEO-curricular is trained first on grounding and then on QA with the backbone frozen. PQ3D-LLM is adapted from PQ3D by replacing T5-Small [67] with Vicuna-7B [12]. Chat-Scene is evaluated directly with its checkpoint.

## 5.1. Object-centric *vs.* Conventional Metrics

As shown in Tabs. 4 and 5, we observe a significant performance drop of all 3D-VL models by simply switching from per-case metrics to object-centric metrics in both grounding and QA. In 3D grounding, we observe an average performance drop by 20%, with LLM-based methods experiencing a more pronounced decline. For 3D-QA, model performance nearly drops to zero for all models after the metric switch, except for the 2D baseline GPT-4o. These findings highlight that existing 3D-VL models lack a comprehensive understanding of objects and are prone to variations in language descriptions and questions. The results underscore the importance of the object-centric evaluation scheme in pinpointing these limitations of 3D-VL models. We provide additional analyses in *Appendix*, such as discussions on outliers and the effect of LLMs.

## 5.2. Chain-of-analysis for Coherence Evaluation

**Grounding Chains.** We aggregate the evaluation results along G-Chains and categorize them into four types based on the grounding results on coarse (simplified grounding texts) and fine-grained texts (primary grounding texts). We leave out LEO variants in our chain analysis considering their weakness in grounding. We show the chained accuracy statistics in Fig. 10. We demonstrate that models struggle with the increased granularity in the G-Chain, where more failures in fine-grained primary grounding texts occur than in coarse simplified grounding texts. This indicates the difficulty of grounding primary grounding texts despite more detailed contexts, suggesting that understanding complex texts and maintaining model performance coherence across text granularities is still a challenge for 3D-VL models.

**Grounding-QA Chains.** We aggregate the results across GQA-Chains to study the gap between grounding and QA. As shown in Fig. 9, we categorize the results into four types based on the results of grounding and QA. We observe a large proportion of broken coherence between tasks, echoing Sec. 3.3. In particular, we design two metrics for evaluating the grounding-QA coherence: $R_1$ for the proportion of GQA-Chains where grounding is correct and QA is incorrect, indicating insufficient QA skills; $R_2$ for the proportion of GQA-Chains where grounding is incorrect but QA is correct, suggesting shortcuts. We find both $R_1$ and $R_2$ are close

Figure 9. **Chain-of-analysis for Grounding-QA-Chains.** The left figure visualizes the evaluation results across GQA-Chains, which exhibit a large proportion of broken grounding-QA coherence. The right figure shows two metrics for evaluating broken coherence: $R_1$ for the proportion of QA failures from insufficient QA skills, and $R_2$ for the proportion of QA successes from shortcuts.



Figure 10. **Chain-of-analysis for Grounding-Chains.**

to 50%, revealing a substantial gap between the skills of grounding and QA, as well as the prevalence of shortcuts in QA. This advocates deeper explorations in enhancing QA skills and mitigating shortcuts for 3D-VL models.

## 5.3. Effect of LLMs

**LLMs hinder grounding.** Tab. 4 and Fig. 10 show that LLM-based models perform worse than those without LLM. This includes (1) models that explicitly use LLM for grounding, such as Chat-Scene, which underperforms compared to non-LLM models like PQ3D and SceneVerse, despite excelling on existing benchmarks [5, 91]; and (2) models indirectly influenced by LLM, such as PQ3D-LLM, which performs worse than PQ3D, suggesting that integrating LLM parameters may bias the learning of grounding. These findings indicate that LLM-based models face a heightened risk of overfitting in grounding tasks.

**LLMs do not fundamentally enhance QA.** While LLM-based models achieve higher per-case accuracy, this is expected given their inherent language modeling capability. However, they have not shown a fundamentally better capability in 3D QA, as evidenced by their limited accuracy in object-centric metrics (Sec. 5.1) and poor grounding-QA coherence (Sec. 5.2). This suggests that the primary bottleneck in 3D QA lies in 3D perception and VL alignment rather than language modeling, where LLMs excel. Moreover, prior works [35, 99] show that simple QA heads (*e.g.*, T5-Small [67] and MCAN [88]) can already achieve competitive performance, indicating that 3D QA requires only

basic language modeling. Therefore, improving 3D QA may depend more on advancing 3D vision foundation models than on leveraging LLMs.

## 5.4. Additional Insights

**Task.** Results in Tab. 4 and Fig. 10 highlight the strong grounding capabilities of PQ3D and SceneVerse, suggesting that scaling up 3D-VL data is a promising strategy for grounded 3D scene understanding. This supports training 3D vision foundation models without integrating LLMs, which proves redundant and even detrimental. On the other hand, 3D QA remains highly challenging due to severe overfitting and shortcut learning in current 3D-VL models. A practical solution is to start with a pre-trained backbone with strong grounding capability and then perform lightweight finetuning. This is supported by (1) SceneVerse (finetuning QA head on top of grounding pretraining) shows best QA performances among non-LLM models, and (2) LEO-curricular (grounding-then-QA) outperforms LEO-multi (multi-task).

**Knowledge types.** We observe that geometry (Geo.) is the most challenging aspect in grounding task, probably because geometric features are rarely referenced in training data. In contrast, geometry-related questions in QA involve less diverse answers, potentially reducing the challenge. Conversely, the diverse answers in class and appearance (App.) increase the task difficulty and lead to lower accuracy.

## 6. Conclusion

We propose BEACON3D, a novel benchmark for 3D grounding and QA tasks, delivering an evaluation paradigm shift to object-centric evaluation and analysis across grounding-QA chains. BEACON3D is driven by a detailed investigation into the limitations of existing 3D-VL benchmarks, addressing flawed test data, vulnerable evaluation metrics, and the isolation of grounding and QA tasks. Our evaluation of state-of-the-art 3D-VL models highlights model pitfalls like insufficient object-level understanding, weak grounding-QA coherence, and limited effect of LLM on 3D-VL tasks.

# Acknowledgments

# References

[1] Ahmed Abdelreheem, Ujjwal Upadhyay, Ivan Skorokhodov, Rawan Al Yahya, Jun Chen, and Mohamed Elhoseiny. 3dref-transformer: Fine-grained object identification in real-world scenes using natural language. In *Proceedings of Winter Conference on Applications of Computer Vision (WACV)*, 2022. 3

[2] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 3, 13

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *International Conference on Computer Vision (ICCV)*, 2015. 3

[4] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3, 13

[5] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 3, 8, 13

[6] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 3

[7] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Language conditioned spatial relation reasoning for 3d object grounding. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2, 3, 7, 13, 15, 16

[8] 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. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3

[9] Yixin Chen, Siyuan Huang, Tao Yuan, Siyuan Qi, Yixin Zhu, and Song-Chun Zhu. Holistic++ scene understanding: Single-view 3d holistic scene parsing and human pose estimation with human-object interaction and physical commonsense. In *International Conference on Computer Vision (ICCV)*, 2019. 2

[10] Yixin Chen, Junfeng Ni, Nan Jiang, Yaowei Zhang, Yixin Zhu, and Siyuan Huang. Single-view 3d scene reconstruction with high-fidelity shape and texture. In *International Conference on 3D Vision (3DV)*, 2024. 3

[11] Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 13

[12] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90% chatgpt quality. https://lmsys.org/blog/2023-03-30-vicuna, 2023. 7, 13

[13] Tao Chu, Pan Zhang, Xiaoyi Dong, Yuhang Zang, Qiong Liu, and Jiaqi Wang. Unified scene representation and reconstruction for 3d large language models. *arXiv preprint arXiv:2404.13044*, 2024. 3

[14] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 6, 15

[15] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 3

[16] Weipeng Deng, Runyu Ding, Jihan Yang, Jiahui Liu, Yijiang Li, Xiaojuan Qi, and Edith Ngai. Can 3d vision-language models truly understand natural language? *arXiv preprint arXiv:2403.14760*, 2024. 3

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019. 13

[18] Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. Pla: Language-driven open-vocabulary 3d scene understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[19] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 3

[20] Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. Scene-llm: Extending language model for 3d visual understanding and reasoning. *arXiv preprint arXiv:2403.11401*, 2024. 3

[21] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[22] Ran Gong, Jiangyong Huang, Yizhou Zhao, Haoran Geng, Xiaofeng Gao, Qingyang Wu, Wensi Ai, Ziheng Zhou, Demetri Terzopoulos, Song-Chun Zhu, et al. Arnold: A benchmark for language-grounded task learning with continuous states in realistic 3d scenes. In *International Conference on Computer Vision (ICCV)*, 2023. 2

[23] Huy Ha and Shuran Song. Semantic abstraction: Open-world 3d scene understanding from 2d vision-language models. In *Conference on Robot Learning (CoRL)*, 2022. 3

[24] Yining Hong, Chunru Lin, Yilun Du, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. 3d concept learning and reasoning from multi-view images. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3

[25] 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. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2, 3

[26] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. 13

[27] 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. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 2, 3, 7, 13

[28] Jiangyong Huang, William Yicheng Zhu, Baoxiong Jia, Zan Wang, Xiaojian Ma, Qing Li, and Siyuan Huang. Perceive, ground, reason, and act: A benchmark for general-purpose visual representation. *arXiv preprint arXiv:2211.15402*, 2022. 3

[29] 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 *International Conference on Machine Learning (ICML)*, 2024. 2, 3, 7, 13

[30] Siyuan Huang, Siyuan Qi, Yixin Zhu, Yinxue Xiao, Yuanlu Xu, and Song-Chun Zhu. Holistic 3d scene parsing and reconstruction from a single rgb image. In *European Conference on Computer Vision (ECCV)*, 2018. 2

[31] Shijia Huang, Yilun Chen, Jiaya Jia, and Liwei Wang. Multiview transformer for 3d visual grounding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[32] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[33] Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Alaa Maalouf, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, et al. Conceptfusion: Open-set multimodal 3d mapping. In *Robotics: Science and Systems (RSS)*, 2023. 3

[34] Baoxiong Jia, Yu Liu, and Siyuan Huang. Improving object-centric learning with query optimization. In *International Conference on Learning Representations (ICLR)*, 2023. 3

[35] 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. In *European Conference on Computer Vision (ECCV)*, 2024. 2, 3, 7, 8, 13, 15, 16

[36] Weitai Kang, Haifeng Huang, Yuzhang Shang, Mubarak Shah, and Yan Yan. Robin3d: Improving 3d large language model via robust instruction tuning. *arXiv preprint arXiv:2410.00255*, 2024. 3

[37] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *International Conference on Computer Vision (ICCV)*, 2023. 3

[38] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *International Conference on Computer Vision (ICCV)*, 2023. 3

[39] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations (ICLR)*, 2022. 3

[40] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning (ICML)*, 2023. 3

[41] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Annual Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. 3

[42] Xiongkun Linghu, Jiangyong Huang, Xuesong Niu, Xiaojian Ma, Baoxiong Jia, and Siyuan Huang. Multi-modal situated reasoning in 3d scenes. *Advances in Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS Datasets and Benchmarks)*, 2024. 3, 7, 13

[43] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 3

[44] 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. 3

[45] Yu Liu, Baoxiong Jia, Yixin Chen, and Siyuan Huang. Slotlifter: Slot-guided feature lifting for learning object-centric radiance fields. In *European Conference on Computer Vision (ECCV)*, 2024. 3

[46] Yu Liu, Baoxiong Jia, Ruijie Lu, Junfeng Ni, Song-Chun Zhu, and Siyuan Huang. Building interactable replicas of complex articulated objects via gaussian splatting. In *International Conference on Learning Representations (ICLR)*, 2025.

[47] Ruijie Lu, Yixin Chen, Junfeng Ni, Baoxiong Jia, Yu Liu, Diwen Wan, Gang Zeng, and Siyuan Huang. Movis: Enhancing multi-object novel view synthesis for indoor scenes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[48] Tiange Luo, Chris Rockwell, Honglak Lee, and Justin Johnson. Scalable 3d captioning with pretrained models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 3

[49] Ruiyuan Lyu, Tai Wang, Jingli Lin, Shuai Yang, Xiaohan Mao, Yilun Chen, Runsen Xu, Haifeng Huang, Chenming Zhu, Dahua Lin, et al. Mmscan: A multi-modal 3d scene dataset with hierarchical grounded language annotations. *Advances in Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS Datasets and Benchmarks)*, 2024. 3

[50] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. In *International Conference on Learning Representations (ICLR)*, 2023. 2, 3, 4, 6, 13

[51] Xianzheng Ma, Yash Bhalgat, Brandon Smart, Shuai Chen, Xinghui Li, Jian Ding, Jindong Gu, Dave Zhenyu Chen, Songyou Peng, Jia-Wang Bian, et al. When llms step into the 3d world: A survey and meta-analysis of 3d tasks via multi-modal large language models. *arXiv preprint arXiv:2405.10255*, 2024. 2

[52] Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, and Others. Openeqa: Embodied question answering in the era of foundation models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3, 6

[53] Yunze Man, Shuhong Zheng, Zhipeng Bao, Martial Hebert, Liang-Yan Gui, and Yu-Xiong Wang. Lexicon3d: Probing visual foundation models for complex 3d scene understanding. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 3

[54] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[55] Yongsen Mao, Yiming Zhang, Hanxiao Jiang, Angel Chang, and Manolis Savva. Multiscan: Scalable rgbd scanning for 3d environments with articulated objects. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2, 6, 15

[56] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[57] David Marr. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press, 2010. 2

[58] Junfeng Ni, Yixin Chen, Bohan Jing, Nan Jiang, Bin Wang, Bo Dai, Puhao Li, Yixin Zhu, Song-Chun Zhu, and Siyuan Huang. Phyrecon: Physically plausible neural scene reconstruction. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 3

[59] Junfeng Ni, Yu Liu, Ruijie Lu, Zirui Zhou, Song-Chun Zhu, Yixin Chen, and Siyuan Huang. Decompositional neural scene reconstruction with generative diffusion prior. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[60] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3, 6, 7, 13, 15, 16

[61] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[62] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[63] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 3

[64] Hang Qi, Yuanlu Xu, Tao Yuan, Tianfu Wu, and Song-Chun Zhu. Scene-centric joint parsing of cross-view videos. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2018. 2

[65] Zekun Qi, Runpei Dong, Shaochen Zhang, Haoran Geng, Chunrui Han, Zheng Ge, Li Yi, and Kaisheng Ma. Shapellm: Universal 3d object understanding for embodied interaction. In *European Conference on Computer Vision (ECCV)*, 2024. 2

[66] 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 (ICML)*, 2021. 3, 13

[67] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research (JMLR)*, 2020. 7, 8, 13

[68] Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind. In *Asian Conference on Computer Vision (ACCV)*, 2024. 3

[69] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d: Mask transformer for 3d semantic instance segmentation. In *International Conference on Robotics and Automation (ICRA)*, 2023. 3

[70] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[71] Simranjit Singh, Georgios Pavlakos, and Dimitrios Stamoulis. Evaluating zero-shot gpt-4v performance on 3d visual question answering benchmarks. *arXiv preprint arXiv:2405.18831*, 2024. 3

[72] Julian Straub, Daniel DeTone, Tianwei Shen, Nan Yang, Chris Sweeney, and Richard Newcombe. Efm3d: A benchmark for measuring progress towards 3d egocentric foundation models. *arXiv preprint arXiv:2406.10224*, 2024.

[73] Emilia Szymanska, Mihai Dusmanu, Jan-Willem Buurlage, Mahdi Rad, and Marc Pollefeys. Space3d-bench: Spatial 3d question answering benchmark. *arXiv preprint arXiv:2408.16662*, 2024. 3

[74] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[75] Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Nießner. Rio: 3d object instance re-localization in changing indoor environments. In *International Conference on Computer Vision (ICCV)*, 2019. 2, 6, 15

[76] Diwen Wan, Ruijie Lu, and Gang Zeng. Superpoint gaussian splatting for real-time high-fidelity dynamic scene reconstruction. *arXiv preprint arXiv:2406.03697*, 2024. 3

[77] Qi Wang, Ruijie Lu, Xudong Xu, Jingbo Wang, Michael Yu Wang, Bo Dai, Gang Zeng, and Dan Xu. Roomtex: Texturing compositional indoor scenes via iterative inpainting. In *European Conference on Computer Vision (ECCV)*, 2024. 3

[78] Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, et al. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3

[79] Yan Wang, Baoxiong Jia, Ziyu Zhu, and Siyuan Huang. Masked point-entity contrast for open-vocabulary 3d scene understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[80] Le Xue, Ning Yu, Shu Zhang, Artemis Panagopoulou, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, et al. Ulip-2: Towards scalable multimodal pre-training for 3d understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[81] 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. *arXiv preprint arXiv:2406.05132*, 2024. 2, 3

[82] Jianing Yang, Xuweiyi Chen, Shengyi Qian, Nikhil Madaan, Madhavan Iyengar, David F Fouhey, and Joyce Chai. Llm-grounder: Open-vocabulary 3d visual grounding with large language model as an agent. In *International Conference on Robotics and Automation (ICRA)*, 2024. 3

[83] Zhengyuan Yang, Songyang Zhang, Liwei Wang, and Jiebo Luo. Sat: 2d semantics assisted training for 3d visual grounding. In *International Conference on Computer Vision (ICCV)*, 2021. 3

[84] Shuquan Ye, Dongdong Chen, Songfang Han, and Jing Liao. 3d question answering. *IEEE Transactions on Visualization and Computer Graph (TVCG)*, 2022. 3

[85] Huangyue Yu, Baoxiong Jia, Yixin Chen, Yandan Yang, Puhao Li, Rongpeng Su, Jiaxin Li, Qing Li, Wei Liang, Song-Chun Zhu, Tengyu Liu, and Siyuan Huang. Metascenes: Towards automated replica creation for real-world 3d scans. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[86] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision (ECCV)*, 2016. 3

[87] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 3

[88] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3, 8, 15, 16

[89] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[90] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations (ICLR)*, 2022. 3

[91] Yiming Zhang, ZeMing Gong, and Angel X Chang. Multi3drefer: Grounding text description to multiple 3d objects. In *International Conference on Computer Vision (ICCV)*, 2023. 2, 3, 8, 13

[92] Zhuofan Zhang, Ziyu Zhu, Pengxiang Li, Tengyu Liu, Xiaojian Ma, Yixin Chen, Baoxiong Jia, Siyuan Huang, and Qing Li. Task-oriented sequential grounding in 3d scenes. *arXiv preprint arXiv:2408.04034*, 2024. 3, 14

[93] Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3dvg-transformer: Relation modeling for visual grounding on point clouds. In *International Conference on Computer Vision (ICCV)*, 2021. 3

[94] Youjun Zhao, Jiaying Lin, Shuquan Ye, Qianshi Pang, and Rynson WH Lau. Openscan: A benchmark for generalized open-vocabulary 3d scene understanding. *arXiv preprint arXiv:2408.11030*, 2024. 3

[95] Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3d: Exploring unified 3d representation at scale. In *International Conference on Learning Representations (ICLR)*, 2024. 3

[96] Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness. *arXiv preprint arXiv:2409.18125*, 2024. 3, 4

[97] Yixin Zhu, Tao Gao, Lifeng Fan, Siyuan Huang, Mark Edmonds, Hangxin Liu, Feng Gao, Chi Zhang, Siyuan Qi, Ying Nian Wu, et al. Dark, beyond deep: A paradigm shift to cognitive ai with humanlike common sense. *Engineering*, 2020. 2

[98] Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3d-vista: Pre-trained transformer for 3d vision and text alignment. In *International Conference on Computer Vision (ICCV)*, 2023. 2, 3, 7, 13, 15, 16

[99] Ziyu Zhu, Zhuofan Zhang, Xiaojian Ma, Xuesong Niu, Yixin Chen, Baoxiong Jia, Zhidong Deng, Siyuan Huang, and Qing Li. Unifying 3d vision-language understanding via promptable queries. In *European Conference on Computer Vision (ECCV)*, 2024. 2, 3, 4, 7, 8, 13, 15, 16

## A. Annotation Tool

We set up an interactive annotation tool for data collection based on SQA3D [50]. We present a visualization of the user interface (UI) in Fig. A.1, including a 3D scene viewer (left), an annotation editor (middle), and object information (right). There are three G-Chains and three GQA-Chains to be annotated in the annotation editor for each target object.

Two panels on the right exhibit details of each annotation:
- For the grounding task, the human annotator is supposed to fill the referential text with precise and natural language, and then select the involved knowledge types and a list of objects that match the referential text.
- For the QA task, the human annotator first generates a QA pair based on the "grounding text", which lists three *primary grounding texts* from the G-Chains. Then, the annotator labels the knowledge type and the flag of *extra knowledge*, *e.g.*, "no" if the answer is covered by the "grounding text".

## B. Baselines

**ViL3DRel [7].** This is a 3D-VL specialist model for grounding, trained in a single-task scheme. We use the official checkpoint trained on ScanRefer [5].

**3D-VisTA [98].** While 3D-VisTA adopts task-specific fine-tuning for downstream tasks by default, we perform multi-task training by aggregating the datasets it uses. The datasets for grounding include ScanRefer, Nr3D [2], Sr3D [2], and Multi3DRefer [91]. The datasets for QA include ScanQA [4] and SQA3D [50].

**PQ3D [99].** PQ3D is a 3D-VL generalist model that supports both grounding and QA tasks. We directly use the checkpoint after pretraining and multi-task training. The training datasets include Scan2Cap [11] in addition to the datasets for 3D-VisTA.

**SceneVerse [35].** SceneVerse is a 3D-VL model pretrained on large-scale grounding datasets. To make it a generalist model for grounding and QA, we finetune a QA head while freezing the pretrained backbone weights to preserve its grounding ability. The datasets for fine-tuning include ScanQA and SQA3D.

**GPT-4o [60].** As a state-of-the-art LLM, GPT-4o is selected as a specialist model for QA to probe the upper bound of LLMs. We adopted the evaluation pipeline outlined in [42] to assess GPT-4o's performance. In our evaluation, we prompt GPT-4o to answer the questions based on a collection of objects, which comprises the category, location, size, and attributes of each object. The object attributes are extracted with GPT-4V [60].

**LEO-multi.** To address the lack of grounding capability in LEO [29], we design a grounding loss alongside the original autoregressive language modeling loss. The grounding loss resembles contrastive learning (CLIP [66]) on the alignment between object tokens (the input to LLM) and text embeddings. With the multi-task objective, we train LEO-multi by combining grounding (ScanRefer and Nr3D) with instruction-tuning tasks (ScanQA, SQA3D, 3RScan-QA [29], 3RScan-Plan [29], and 3RScan-Dialog [29]).

**LEO-curricular.** Similar to LEO-multi, LEO-curricular incorporates the contrastive grounding loss but learns grounding and QA in a curricular strategy. We first train the 3D encoder of LEO-curricular with grounding loss on ScanRefer and Nr3D. We then freeze the 3D encoder and finetune the LLM with LoRA [26] on instruction-tuning datasets.

**PQ3D-LLM.** This is a model variant based on PQ3D, substituting the original T5-Small [67] with Vicuna-7B [12], which is finetuned with LoRA. The training setting is identical to PQ3D.

**Chat-Scene [27].** Chat-Scene is designed to be a 3D-VL generalist model, using object identifiers and LLM to perform grounding. The training datasets include ScanRefer, Multi3DRefer, Scan2Cap, ScanQA, and SQA3D. We directly use its released checkpoint for evaluation.

## C. Additional Analyses

### C.1. Outliers and Prospective Questions

We observe several outliers in our evaluation results. Below, we address these outliers and answer prospective questions:

***Poor grounding for LEO-multi and LEO-curricular.*** The grounding performance of these two models falls significantly below that of others. We attribute this to our implementation of the grounding task learning, which employs contrastive learning between object tokens and text embeddings of pretrained LLM (*e.g.*, Vicuna). We receive two lessons from this: (1) contrastive learning demands large-scale data while the scarce 3D-VL data proves insufficient; and (2) unlike CLIP, the text embeddings of pretrained LLM may not be suitable for contrastive learning.

***Poor QA for PQ3D and PQ3D-LLM.*** Despite the strong performance in grounding for these two models, their performance in QA is notably weak. We attribute this to the choice of language encoder. Compared to 3D-VisTA, PQ3D adopts a similar overall architecture but differs in language encoder: 3D-VisTA uses BERT [17], whereas PQ3D uses CLIP. The reasonable QA performance of 3D-VisTA indicates that the

Figure A.1. **Overview of our annotation tool.** The interface includes a 3D viewer (left), an annotation editor (middle), and object information (right). Two panels on the right exhibit details of each annotation for the grounding and QA task, respectively.

CLIP language encoder is suboptimal for QA task, despite being adequate for grounding. This further underscores the linguistic gap between grounding and QA tasks: grounding texts encompass descriptive language while questions involve diverse querying patterns. It reveals the limitations of the CLIP language encoder in addressing this disparity.

***Why is PQ3D-LLM worse than PQ3D in grounding?*** While the LLM incorporated by PQ3D-LLM is only used for QA, it introduces a significant number of extra parameters for optimization, which may hinder the learning of grounding during multi-task learning and consequently weaken the grounding performance.

***Why is PQ3D-LLM not better than PQ3D in QA?*** In PQ3D, the input to the QA head (*e.g.*, LLM) only comprises object tokens, which can be regarded as foreign language for LLM. The challenge of utilizing these tokens for QA cannot be alleviated by incorporating LLM, despite its strength in language processing. Additionally, incorporating LLM for QA is prone to overfitting given the scarcity of 3D QA data.

***Strong performance of GPT-4o in QA.*** We observe that GPT-4o significantly outperforms 3D-VL models in QA, especially in questions related to appearance (App.) and existence (Exi.). This showcases the upper bound of using explicit textual information (*e.g.*, object lists with attributes), which bypasses 3D perception. The considerable gap between GPT-4o and 3D-VL models further suggests that 3D perception remains a key bottleneck in 3D-VL models.

## C.2. Discussion on the Effect of LLM

**LLM hinders grounding.** This conclusion is drawn from the consideration of two categories of models:

- *LLM directly used for grounding.* Models that perform grounding based on LLM (*e.g.*, Chat-Scene) exhibit less robust performance compared to models without LLM. Specifically, despite the close performances on ScanRefer, Chat-Scene lags behind PQ3D and SceneVerse on BEA-CON3D, which implies the potential risk of overfitting for LLM-based grounding. However, LLM may be beneficial in more complex grounding tasks that require high-level reasoning or planning, *e.g.*, sequential grounding [92]. This suggests that the effect of LLM-based grounding varies according to task complexity.

- *LLM not directly used for grounding.* In models that do not rely on LLM for grounding (*e.g.*, PQ3D-LLM), we observe a weaker performance in grounding after incorporating LLM. This shows the negative effect of LLM's parameters on the learning of grounding during multi-task learning. A practical solution is to decompose multi-task learning into curricular learning, which disregards LLM's parameters during the learning of grounding.

**LLM does not truly improve QA.** We elaborate on this conclusion from three aspects: clarification on how we draw the conclusion, explanation on why per-case metrics do not matter, and analysis on why LLM may not help 3D QA.

- *How we draw the conclusion.* The evidence mainly comes from two observations: (1) the results of LLM-based models are comparable to those without LLM under object-centric metrics; and (2) fragile grounding-QA coherence.

Table A.1. **Evaluation results of grounding on BEACON3D (3RScan).** The settings and metrics follow the main paper. ** denotes models that have never been trained in 3RScan. * denotes models that have been trained in 3RScan but not on grounding. ‡ denotes only point feature is available.

| | Knowledge type | | | | Overall | |
|---|---|---|---|---|---|---|
| | Class | App. | Geo. | Spa. | Case | Obj. |
| *w/o LLM* | | | | | | |
| ViL3DRel** [7] | 41.5 | 44.9 | 37.4 | 37.3 | 41.5 | 18.4 |
| 3D-VisTA** [98] | 45.6 | 38.3 | 37.4 | 40.9 | 45.6 | 21.7 |
| PQ3D**‡ [99] | 38.3 | 28.0 | 36.4 | 35.3 | 38.3 | 13.6 |
| SceneVerse [35] | **61.8** | **51.4** | **53.3** | **57.3** | **61.8** | **37.5** |
| *LLM-based* | | | | | | |
| LEO-multi* | 10.1 | 9.9 | 9.7 | 8.8 | 10.1 | 0.4 |
| LEO-curricular* | 15.3 | 17.7 | 11.8 | 9.3 | 15.3 | 1.1 |
| PQ3D-LLM**‡ | 30.3 | 27.6 | 24.6 | 25.5 | 30.3 | 8.5 |

Table A.2. **Evaluation results of QA on BEACON3D (3RScan).** † indicates text input (*i.e.*, object locations and attributes) instead of 3D point cloud. ** denotes models that have never trained in 3RScan. * denotes models that have been trained in 3RScan but not on QA. ‡ denotes only point feature is available.

| | Knowledge type | | | | | Overall | |
|---|---|---|---|---|---|---|---|
| | Class | App. | Geo. | Spa. | Exi. | Case | Obj. |
| *w/o LLM* | | | | | | | |
| 3D-VisTA** [98] | 15.2 | 24.1 | 28.2 | 25.3 | 28.9 | 25.7 | 3.3 |
| PQ3D**‡ [99] | 6.5 | 19.6 | 13.6 | 16.6 | 52.6 | 25.7 | 0.7 |
| SceneVerse* [35] | 28.3 | 32.3 | 34.6 | 38.9 | 44.6 | 37.4 | 0.4 |
| *LLM-based* | | | | | | | |
| GPT-4o† [60] | 34.8 | 38.2 | 40.0 | 45.4 | **60.7** | **46.1** | **11.0** |
| LEO-multi | **37.0** | 35.0 | 51.8 | 48.5 | 46.5 | 44.1 | 1.8 |
| LEO-curricular | 19.6 | **41.8** | 48.2 | 48.5 | 50.7 | 45.6 | 7.4 |
| PQ3D-LLM**‡ | 13.0 | 21.4 | 17.3 | 21.4 | 33.2 | 23.4 | 1.8 |

- *Why per-case metrics do not matter.* While LLM-based models show slightly better results in per-case metrics, these metrics do not reliably indicate true 3D QA capability. As demonstrated in the main paper, per-case metrics are not robust enough due to their vulnerability to shortcuts. Moreover, the advantage of LLM-based models in per-case metrics is marginal, which is intuitive given LLM's strength in general QA. We believe the marginal gap in per-case metrics cannot evidence a gap in the true capability of 3D QA.
- *Why LLM may not help 3D QA.* We conjecture the bottleneck in 3D QA lies in the alignment between 3D features and QA modules, rather than language generation, where the primary strength of LLM resides. Prior works [35, 98, 99] have shown that simple QA heads (*e.g.*, T5-Small or MCAN [88]) perform well in 3D QA, as the task demands only a basic level of language generation. This explains the minimal contribution of LLM to 3D QA.

**Harnessing LLM for 3D-VL tasks.** We first identify a critical problem in current 3D LVLMs and then propose an effective solution to harness LLM for 3D-VL tasks.

- *Problem.* Our investigation in the main paper reveals that overfitting to text is a critical problem in current 3D LVLMs. This implies a significant imbalance between 3D encoder and LLM, that is, LLM often overshadows 3D encoder during training. This issue is less pronounced in 2D LVLMs owing to the robust 2D features learned through extensive pretraining, which is infeasible for 3D encoders.
- *Solution.* We propose curricular learning, progressing from grounding to QA, as an effective solution to mitigate this issue by shielding 3D features from LLM interference. The effectiveness is evidenced by the advantages of SceneVerse and LEO-curricular.

## C.3. Limitations and Future Work

First, our benchmark prioritizes focused and systematic analysis, which involves trade-offs in task scope and complexity. Our object-centric evaluation excludes more advanced tasks, such as multi-object grounding and complex reasoning. Extending this evaluation framework to include more complex tasks will be a key direction for future work. Second, our baselines may not cover the wide range of existing 3D-VL models. We will evaluate and analyze more models in the future. Third, we consider the performance of the grounding task as a proxy for the grounding implicitly performed in the QA task. This may be unfair to models whose grounding performance is locked due to issues like improper implementation (*e.g.*, LEO-multi and LEO-curricular). Nonetheless, we believe our approach remains practical for assessing grounding-QA coherence in most 3D-VL generalist models.

## D. Domain Transfer

We follow the setting outlined in the main paper to evaluate the baselines in two novel domains: 3RScan [75] and MultiScan [55]. This evaluation is referred to as *domain transfer* since most baselines are only trained on ScanNet [14]. Notably, as Chat-Scene only provides model features for ScanNet, its evaluation on 3RScan and MultiScan is not feasible. We distinguish between two types of domain transfer:
- **: the model has never been trained in the target domain.
- *: the model has been trained in the target domain but on tasks other than the specific one.

**Results.** We present the domain transfer results for 3RScan in Tabs. A.1 and A.2, and MultiScan in Tabs. A.3 and A.4. The overall trends are consistent with those reported in the main paper for ScanNet. For example, while models without

Table A.3. **Evaluation results of grounding on BEACON3D (MultiScan).** The settings and metrics follow the main paper. ** denotes models that have never been trained in MultiScan. Only SceneVerse has been trained in MultiScan.

| | Knowledge type | | | | Overall | |
|---|---|---|---|---|---|---|
| | Class | App. | Geo. | Spa. | Case | Obj. |
| *w/o LLM* | | | | | | |
| ViL3DRel** [7] | 33.2 | 34.4 | 25.0 | 32.0 | 33.2 | 13.2 |
| 3D-VisTA** [98] | 40.8 | 30.5 | 28.1 | 38.0 | 40.8 | 18.9 |
| PQ3D** [99] | 56.3 | 53.9 | 37.5 | 52.8 | 56.3 | 34.0 |
| SceneVerse [35] | **59.5** | **54.6** | **53.1** | **56.6** | **59.5** | **35.9** |
| *LLM-based* | | | | | | |
| LEO-multi** | 9.0 | 9.1 | 9.4 | 9.0 | 9.0 | 1.3 |
| LEO-curricular** | 11.7 | 11.0 | 6.3 | 9.0 | 11.7 | 0 |
| PQ3D-LLM** | 51.0 | 46.8 | 37.5 | 49.0 | 51.0 | 25.8 |

Table A.4. **Evaluation results of QA on BEACON3D (MultiScan).** [†] indicates text input (*i.e.*, object locations and attributes) instead of 3D point cloud. ** denotes models that have never been trained in MultiScan. * denotes models that have been trained in MultiScan but not on QA.

| | Knowledge type | | | | | Overall | |
|---|---|---|---|---|---|---|---|
| | Class | App. | Geo. | Spa. | Exi. | Case | Obj. |
| *w/o LLM* | | | | | | | |
| 3D-VisTA** [98] | 6.5 | 22.6 | 16.7 | 13.2 | 28.8 | 19.1 | 0 |
| PQ3D** [99] | 21.0 | 16.8 | 16.7 | 9.6 | 39.0 | 20.8 | 0.6 |
| SceneVerse* [35] | 16.2 | 32.1 | 12.5 | 26.5 | 38.1 | 28.9 | 3.1 |
| *LLM-based* | | | | | | | |
| GPT-4o[†] [60] | **29.0** | **41.6** | 33.3 | 25.7 | **59.3** | 39.4 | **7.6** |
| LEO-multi** | 12.9 | 24.1 | 41.7 | 24.3 | 32.2 | 25.6 | 2.5 |
| LEO-curricular** | 8.1 | 27.0 | **50.0** | **28.7** | 41.5 | 29.8 | 3.8 |
| PQ3D-LLM** | 6.5 | 21.9 | 8.3 | 11.0 | 25.4 | 17.0 | 0.6 |

LLM (*e.g.*, SceneVerse) excel in grounding, LLM-based models (*e.g.*, LEO-curricular) perform better under per-case metrics but struggle with object-centric metrics in QA. In particular, we report several specific findings regarding the domain transfer results:

- *Challenge of domain transfer.* All models exhibit notable performance declines, emphasizing the challenge of domain transfer (ScanNet → 3RScan; MultiScan). SceneVerse surpasses PQ3D owing to its comprehensive pretraining across diverse scene domains. Moreover, training on 3RScan-QA improves QA performance on 3RScan (LEO-multi and LEO-curricular). These findings highlight the inevitable domain gap and the benefits of cross-domain pretraining.
- *Limitations of feature-dependent models.* PQ3D and PQ3D-LLM experience considerable performance drops on 3RScan due to a lack of image and voxel features. While this issue results in only a marginal drop on ScanNet, as reported in the original paper [99], the considerable drop on 3RScan indicates the heightened challenges of transferring to novel domains for feature-dependent models such as PQ3D and Chat-Scene.
- *More challenging 3D perception in MultiScan.* Performance on MultiScan is consistently lower than on 3RScan, reflecting the increased difficulty of 3D perception in the domain of MultiScan. SceneVerse, despite using a simple QA head [88], outperforms LEO-multi and matches LEO-curricular. This suggests that the bottleneck in QA lies in 3D perception, suppressing the contribution of LLM. It further underscores the need for more powerful 3D encoders to address this bottleneck.
- *Performance degradation of GPT-4o.* GPT-4o exhibits noticeably lower performance on 3RScan and MultiScan compared to ScanNet, with the results on 3RScan approached by LEO-curricular. We attribute this degradation to incomplete object attributes stemming from insufficient multi-view images, which limits the object attribute extrac-

tion by GPT-4V. This reveals that, despite their strengths in 3D QA, LLMs and 2D LVLM are constrained by the availability of high-quality multi-view images.

# E. Illustration of Data and Evaluation

We present a video demo to illustrate the process of data collection and evaluation (see attachment). Here we show the static overview in Fig. A.2 and A.3.

**Data Collection**

Figure A.2. **Static overview of data collection.** Check the dynamic process in our video demo in the attachment.



**Evaluation**

Figure A.3. **Static overview of evaluation.** Check the dynamic process in our video demo in the attachment.