SURPRISE3D: A Dataset for Spatial Understanding and Reasoning in Complex 3D Scenes

Jiaxin Huang[1†], **Ziwen Li**[1†], **Hanlve Zhang**[1†], **Runnan Chen**[2], **Zhengqing Gao**[1], **Xiao He**[3], **Yandong Guo**[3], **Wenping Wang**[4], **Tongliang Liu**[1,2‡], **Mingming Gong**[1,5‡]

[1]MBZUAI, [2]The University of Sydney, [3]AI2Robotic
[4]Texas A&M University, [5]The University of Melbourne

Abstract

The integration of language and 3D perception is critical for embodied AI and robotic systems to perceive, understand, and interact with the physical world. Spatial reasoning, a key capability for understanding spatial relationships between objects, remains underexplored in current 3D vision-language research. Existing datasets often mix semantic cues (e.g., object name) with spatial context, leading models to rely on superficial shortcuts rather than genuinely interpreting spatial relationships. To address this gap, we introduce SURPRISE3D, a novel dataset designed to evaluate language-guided spatial reasoning segmentation in complex 3D scenes. SURPRISE3D consists of more than 200k vision language pairs across 900+ detailed indoor scenes from ScanNet++ v2, including more than 2.8k unique object classes. The dataset contains 89k+ human-annotated spatial queries deliberately crafted without object name, thereby mitigating shortcut biases in spatial understanding. These queries comprehensively cover various spatial reasoning skills, such as relative position, narrative perspective, parametric perspective, and absolute distance reasoning. Initial benchmarks demonstrate significant challenges for current state-of-the-art expert 3D visual grounding methods and 3D-LLMs, underscoring the necessity of our dataset and the accompanying 3D Spatial Reasoning Segmentation (3D-SRS) benchmark suite. SURPRISE3D and 3D-SRS aim to facilitate advancements in spatially aware AI, paving the way for effective embodied interaction and robotic planning.

1 Introduction

Spatial reasoning lies at the heart of embodied AI and robotic systems [5, 31, 33, 8]. For agents to navigate real-world environments, manipulate objects, or interact with humans, they must interpret instructions that are rooted in 3D spatial layouts and context. Consider a robot waiter tasked with serving drinks to *the table to the left of the couch*, or a household assistant that infers from *I am thirsty* the intention of regaining the nearest cup. In both scenarios, the agent must go beyond identifying object categories, reasoning about spatial relationships, viewpoint-dependent references, and pragmatic intent. This type of language-guided spatial understanding is critical for tasks such as navigation, manipulation, and human-robot interaction [39].

However, despite its importance, existing 3D vision-language grounding benchmarks do not faithfully capture or evaluate spatial reasoning. Most prior datasets rely heavily on explicit object references,

[†]Equal contribution.
[‡]Corresponding author
[*]Code available at: https://github.com/tmllab/2025_NeurIPS_SURPRISE3D

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

allowing models to localize targets by matching named categories or rigid templates [42, 41, 25], without spatial inference. As a result, current models achieve strong performance not by understanding spatial context but by exploiting superficial textual patterns or semantic shortcuts. This **shortcut bias** has been discussed in [49], showing that models tend to rely on object names rather than appearance or spatial relations, leading to imbalanced learning. When object names are replaced by neutral placeholders (e.g., "object"), model performance drops sharply, confirming the dominance of name-based shortcuts in existing 3D vision-language grounding benchmarks.

We identify three major limitations of current 3D vision-language datasets:

(1) **Overreliance on explicit queries**. Datasets such as ScanRefer [6] and ReferIt3D [3] provide queries that including a object name (e.g., chair). Although challenging before the emergence of large language models (LLMs), such datasets have become increasingly solvable through category detection, often requiring little or no reasoning.

(2) **Limited and shallow reasoning coverage**. Recent efforts such as Intent3D [27], ScanReason [60], Reason3D [21] and Instruct3D [15] have taken steps towards incorporating implicit queries and common sense. However, these benchmarks remain narrow in scope. For example, ScanReason contains only 10K Q&A pairs across five loosely defined reasoning types, and does not provide a rigorous or fine-grained definition of spatial reasoning in 3D contexts [58].

(3) **Template-driven or trivial spatial queries**. Many datasets [17] formulate spatial queries using basic patterns (e.g., 'the object to the left') that do not require understanding geometric variability or occlusion and can often be resolved using positional heuristics.

These limitations lead to a recurring problem: models perform well by leveraging semantic priors and dataset biases rather than learning spatial reasoning. There is a critical need for a benchmark that disentangles spatial understanding from semantic recognition and poses queries that necessitate interpreting the scene geometry in context. To address this gap, we introduce SURPRISE3D: a large-scale dataset and benchmark designed from the first principles to evaluate language-guided spatial reasoning in complex 3D scenes. Built on top of 900+ richly annotated indoor environments from ScanNet++ v2 [52], SURPRISE3D includes more than 200,000 query-object mask pairs, covering 2,800+ object classes. It is the first benchmark to support spatial reasoning segmentation at this scale, breadth, and level of annotation precision. Key features of SURPRISE3D include:

Complex spatial queries: In particular, we find that LLMs and MLLMs are incapable of generating spatial reasoning annotations with sufficient fidelity, necessitating a human-in-the-loop annotation process. 89K+ human-generated questions that require varied spatial reasoning. Queries investigate relative position recall (*'the vase next to the left door'*), narrative perspective reasoning (describing objects from a moving observer's point of view), parametric perspective reasoning (specifying angles or offsets) and absolute distance reasoning (*'the plamp 2 meters above the floor'*).

Pragmatic language quality with Human Check. All expressions follow Gricean conversational maxims (clarity, relevance, and brevity) and are vetted by multiple annotators to avoid ambiguity or bias. To resolve ambiguous references in cluttered environments, we adopt Gricean maxims to ensure clarity and informativeness, and introduce category-specific disambiguation rules that favor spatial, functional, or visual attributes depending on context.

3D spatial reasoning segmentation benchmark. We introduce a formal evaluation framework tailored for 3D spatial reasoning segmentation(3D-SRS). It includes task definitions, metrics such as mask IoU and grounding precision, and diagnostic breakdowns between reasoning types. Our results demonstrate that state-of-the-art 3D vision-language models, when deprived of explicit naming, perform significantly worse, revealing their limitations in spatial understanding.

Knowleage-based reasoning. Beyond spatial queries, we incorporate 110K LLM-generated questions focusing on knowleage-based reasoning: for example, 'the object used to sit' (common sense) or 'the item someone might be reaching for' (human intention). These complement spatial cues with functional and behavioral semantics, which are essential for embodied interaction.

By combining linguistic subtlety and geometric complexity, SURPRISE3D sets a new standard for 3D spatial reasoning benchmarks. Unlike existing datasets, our queries are implicit, ambiguous, and semantically lightweight: any correct object that satisfies the spatial constraint is valid, and object names are deliberately avoided. This forces models to rely on reasoning rather than recognition, and

Table 1: Comparison of major 3D vision-language datasets. 'Multi-target' indicates if a query refers to multiple objects. 'Lang Source' denotes whether language queries are human-annotated or generated via template or LLM. '1st observ.' refers to descriptions written from a first-person observer perspective (i.e., egocentric viewpoint). 'Shortcut Free' indicates if object names are avoided in queries.

Dataset	Output	Multi-target	Lang. Source	Spatial Reason	1st observ.	Shortcut Free
CLEVR3D [50]	Lang	-	Template	✗	-	-
Scan2Cap [7]	Lang	-	Human	✗	-	-
ScanQA [4]	Lang	✗	Human	✗	-	-
3DVQA [12]	Lang	✗	Human	✗	-	-
SQA3D [40]	Lang	-	Human	✓	✓	✗
ScanScribe [61]	Lang	✗	Template,LLM	-	-	-
3DMV-VQA [16]	Lang	-	Template	✗	✗	✗
M3DBench [30]	Lang	✓	LLM	✗	✗	✗
SceneVerse [24]	Lang	✗	Template,LLM	✗	✗	-
MSQA [34]	Lang	-	Human,LLM	✓	✓	✗
VLA-3D [56]	Lang	✓	Template	✓	✓	✗
ExCap3D [53]	Lang	✓	LLM	✗	✓	✗
ReferIt3D [3]	BBox	✓	Human,Template	✗	✓	✗
ScanRefer [6]	BBox	✗	Human	✗	✓	✗
3D-DenseOG [23]	BBox	✓	Human	✗	✓	✗
Grounded 3D-LLM [11]	Mask	✗	Template,LLM	✓	✗	✗
ScanEnts3D [1]	BBox	✓	Human	✓	✗	✗
PhraseRefer [54]	BBox	✓	Human	✗	✗	✗
EmbodiedScan [46]	BBox	✓	Template,LLM	✓	✓	✗
3D-LLM [17]	Lang + BBox	✓	LLM	✓	✗	✗
LL3DA [9]	Lang + BBox	✓	Template,LLM	✗	✓	✗
3DMIT [32]	Lang + BBox	✗	LLM	✗	✓	✗
3D-GRAND [51]	Lang + BBox	✓	Template,LLM	✗	✗	✗
Segpoint [15]	Mask	✗	LLM	✓	✗	✓
ScanReason [60]	BBox	✗	LLM	✓	✗	✓
Reason3D [21]	Mask	✗	-	✓	✗	✓
Intent3D [27]	BBox	✓	LLM	✗	✗	✓
SURPRISE3D (Ours)	Mask	✓	Human,LLM	✓	✓	✓

aligns with recent calls to evaluate deeper spatial understanding [60]. Table 1 summarizes the key differences from previous benchmarks.

Error Analysis and Future Directions. Despite its overall challenge, SURPRISE3D reveals consistent weaknesses in perspective-taking queries, where all evaluated models perform poorly under both zero-shot and fine-tuned settings. We hypothesize that this stems from the lack of explicit egocentric modeling and viewpoint transformation capabilities in current architectures. Future work may incorporate learned egocentric priors, dynamic pose embeddings, or multi-view reasoning modules to improve spatial grounding from first-person descriptions.

Empirical evaluations with the state-of-the-art expert 3D visual grounding (VG) model and 3D-LLMs further confirm that performance degrades dramatically when explicit semantic cues are removed, underscoring the need for spatially grounded reasoning. Our main contributions are as follows.

- We introduce **SURPRISE3D**, a novel dataset of more than 200K language-guided 3D segmentation queries that cover spatial, common sense, and human intention reasoning.
- We define the **3D-SRS benchmark**, a standardized protocol for evaluating spatial reasoning segmentation in 3D point clouds.
- We empirically demonstrate shortcut bias in existing benchmarks and show that current 3D vision-language models significantly underperform on SURPRISE3D, highlighting the need for models capable of implicit and relational reasoning.

We hope SURPRISE3D will serve as a foundation for future research in spatially grounded 3D understanding and drive progress in embodied AI, robotics, and world model.

2 Related Works

2.1 Spatial Reasoning in 3D Vision-Language Model

Understanding natural language in 3D scenes has focused on referring to objects using explicit names or attributes. Recent million-level extensions such as 3D-GRAND [51] and SceneVerse [24] introduce richer annotations, including multi-object grounding. However, these data sets still mainly rely on straightforward object references rather than complex spatial relationships. Consequently, they allow models to exploit shortcut biases, limiting their ability to evaluate true spatial understanding. Recently, ExCap3D [53] has explored expressive captioning in 3D scenes at multiple levels of detail that covers more than 3k objects. This work is inspiring because of the rich language descriptions for abundant object classes, but does not directly focus on segmentation and spatial reasoning. Recent advances like Intent3D [27] introduce grounding based on human intention, but still rely on semantic cues for detecting object categories, ignoring the spatial relationship inherent in 3D. ScanReason expands this by proposing spatial and safety reasoning. However, its spatial component remains relatively coarse and lacks detailed supervision, such as segmentation masks.

2.2 3D Large Language Model

The advent of large language models (LLMs) has led to more sophisticated 3D-VL models emphasizing spatial and reasoning capabilities [19, 24, 29, 30, 36, 37, 38, 14**?**]. For example, 3D-LLM injects 3D spatial knowledge into pre-trained LLMs, facilitating nuanced scene understanding. Models like LEO [19] integrate language models with 3D understanding, enabling open-ended scene reasoning and interaction capabilities. Chat-3D [47], Chat-Scene [18], and Grounded 3D-LLM [11], incorporate explicit identifiers or referent tokens to bridge linguistic queries and specific scene elements. Recent efforts explicitly focus on reasoning-guided segmentation tasks. For example, Reason3D [21] and SegPoint [15] employ LLM-driven frameworks to guide point cloud segmentation based on natural language instructions. MORE3D [26] and MLLM-For3D [20] adopt multimodal LLMs (MLLMs) to simultaneously reason about complex spatial relationships and output detailed segmentation masks.

3 3D Spatial Reasoning Segmentation

3.1 Task Definition

We define 3D Spatial Reasoning Segmentation (3D-SRS) as follows. Given a 3D scene $\mathcal{S}$ (e.g., a reconstructed indoor room) and a language query q describing a spatial relation, the goal is to produce a segmentation mask M that highlights all object(s) in $\mathcal{S}$ satisfying q. Formally, we learn a function $f(\mathcal{S}, q) \rightarrow M$, where $M \subseteq \mathcal{S}$ consists of the 3D points (or object regions) referred to by the query. For example, if $q =$"The chair closest to the door", then M contains the points of the chair nearest the door. If multiple objects satisfy q, the mask covers all of them.

Narrative perspective: interpreting egocentric references from a described point of view. For example, a query might implicitly place an agent in the scene (e.g., sitting on the black sofa facing the blackboard, the object to your left used for teaching). The model should simulate the narrator's point of view and understand terms like 'to your left' or 'in front of you' in this context.

Parametric perspective: understanding the current orientation and position parameters given in the description. For instance, *facing away from the door towards the cabinet* specifies a camera orientation. The model must parse such instructions about where the observer is looking or located and use them to ground other spatial terms.

Relative position: reasoning about spatial relations between objects. Queries frequently use relational phrases such as "on the table", "behind the sofa", "to the left of the cabinet", etc. The model must identify reference objects (table, sofa, cabinet) and understand directional or topological relations (on, behind, left of) to find the target. This includes handling occlusion (e.g., an object "behind the sofa") that may be partially or fully hidden from certain viewpoints, requiring true 3D understanding beyond a single-camera view.

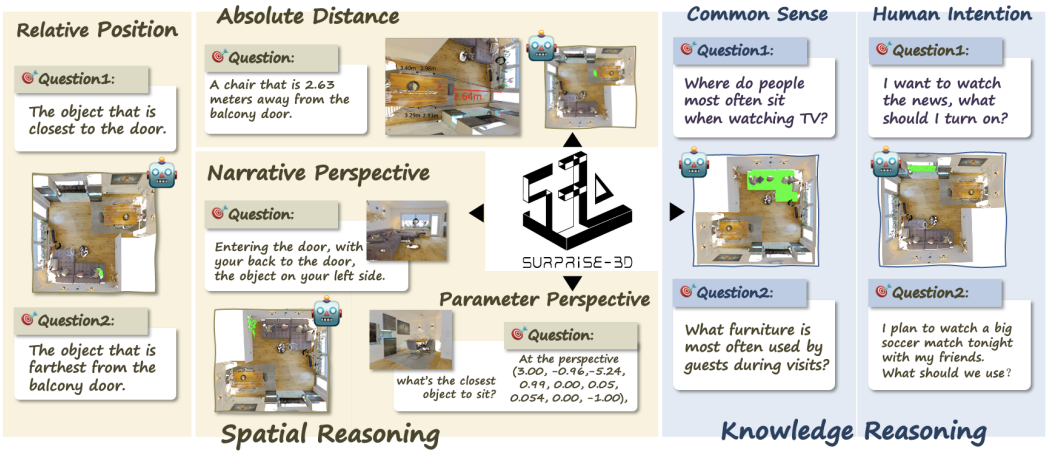


Figure 1: Examples of query categories in 3D spatial reasoning segmentation (3D-SRS) task.

Absolute Distance: interpreting absolute or comparative distance cues. Many queries involve terms such as 'closest', 'furthest', or 'near'. The model should be able to compare distances between multiple objects with the same semantic category. For example, *'the table closest to the bed'* requires finding all tables in the scene and selecting the one with minimal distance to the bed. Importantly, such distance terms are defined in an absolute spatial sense (the physical distances between objects), rather than relative to the viewer's perspective. The model must therefore compute and compare 3D distances or understand spatial superlatives in context.

3.2 Benchmark Protocol and Evaluation

We divide the data into disjoint training and validation splits at the scene level to prevent overlap. For evaluation, we adopt standard segmentation metrics: the primary score is *mean Intersection-over-Union* (mIoU) between the predicted and ground-truth masks across all queries. We also report precision and recall at fixed IoU thresholds (e.g., 0.5 and 0.75) to analyze performance sensitivity. These metrics follow common practice in referring segmentation benchmarks [22, 48, 43].

To support a leaderboard, we release train/val annotations. Participants train models on the public data and submit predicted masks on the test scenes. The benchmark server computes all metrics on the withheld ground truth to rank submissions. This protocol ensures fair comparison with identical data splits.

4 SURPRISE3D: Dataset Construction and Annotation

We constructed SURPRISE3D with two parallel annotation pipelines to capture spatial reasoning, common sense, and human intention reasoning. One pipeline focuses on spatial reasoning, where annotators formulate four type of questions mentioned the section 3 and mask the corresponding target objects in 3D scenes. The other pipeline focuses on common sense and human intention queries, where the questions probe typical human knowledge or intent in the scene and require identifying the object that satisfies the query. By design, these pipelines operate independently but on the same set of scenes, ensuring a rich and complementary set of annotations. In total, SURPRISE3D provides a balanced mix of query types (spatial vs. knowledge-based) and a ground-truth target object for each query. In the following, we describe each annotation process and then analyze the coverage and balance of the dataset.

4.1 Annotation Pipeline

As shown in Figure 2 and Figure 3, we use separate annotation pipelines for spatial and knowledge queries.

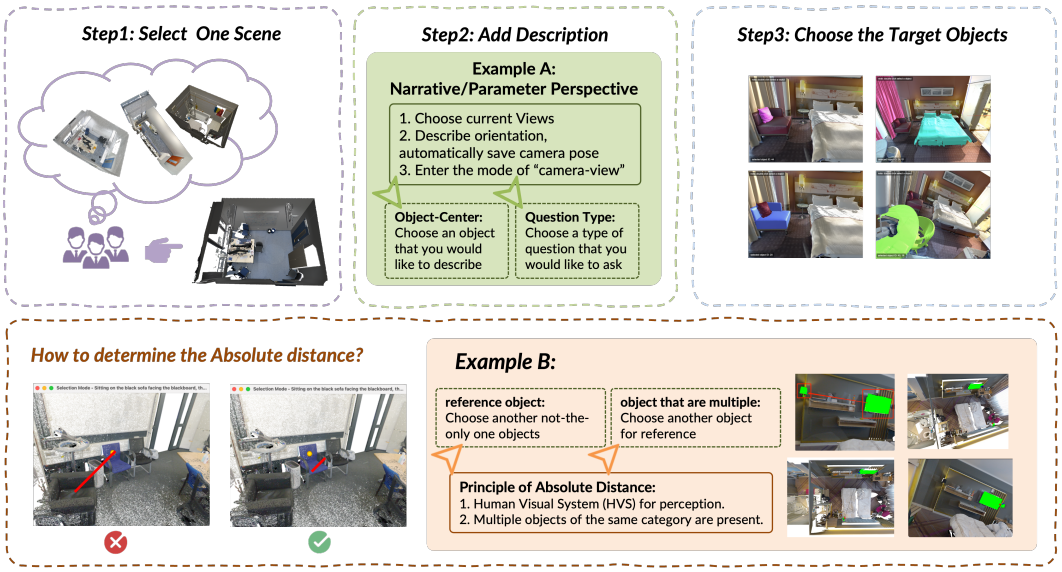


Figure 2: **Spatial reasoning annotation pipeline.** Human Annotators select a scene and target object, then write a question that identifies the object via spatial context, and finally mask the object.

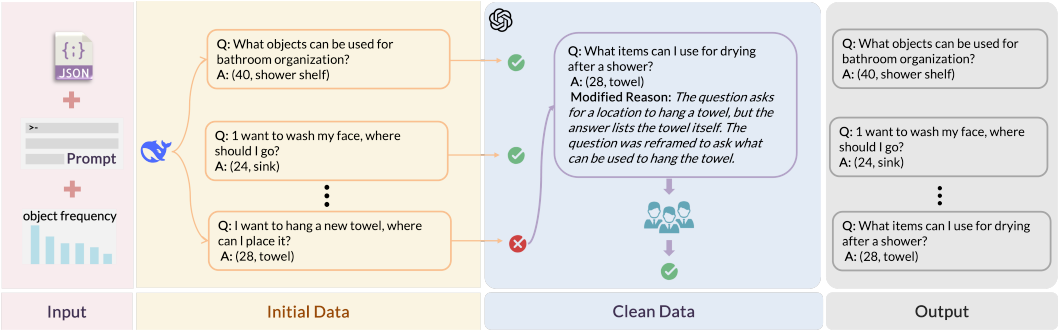


Figure 3: Overview of the common sense and human intention reasoning queries generation pipeline.

Spatial reasoning. Annotators view each 3D scene from a fixed camera viewpoint (e.g., top-down or entry view) and manually identify the object(s) satisfying the spatial query. The interface allows for clicking or highlighting points in the rendered scene. For example, to answer 'closest to the door', the annotator selects the object closest to the entrance point of the door. This produces a ground-truth mask for the target object. Using a fixed viewpoint ensures that spatial relations can be assessed consistently from one perspective.

First, annotators could freely navigate or choose a fixed camera viewpoint in the 3D scene. Note that the target objects can be visible or invisible from this camera view. Locking this point of view is critical, as it establishes a clear 'frame' of reference for egocentric directions such as left or right. The interface then allowed the annotator to enter a description of a target object based on that perspective, and finally to mark the object's mask directly in the point cloud of 3D scenes. We recorded the camera parameters (extrinsics and intrinsics) alongside each query, so that any model processing the data can interpret the spatial language from the correct viewpoint context. The annotation UI thus lets the annotator 'be the agent' in the scene, selecting an orientation and then writing a query as if they were there (for example, an annotator might place the camera facing a wall and then describe 'the chair to my right' relative to that orientation). Once the description was written, the annotator highlighted the referenced object by drawing its segmentation mask over the 3D point cloud (the tool projected the mask onto the 3D object surface). This yields a ground-truth mask for the query.

Knowledge reasoning. We use an LLM-augmented pipeline inspired by previous work [51]. We first generate candidate question-answer pairs using a large language model, given scene metadata

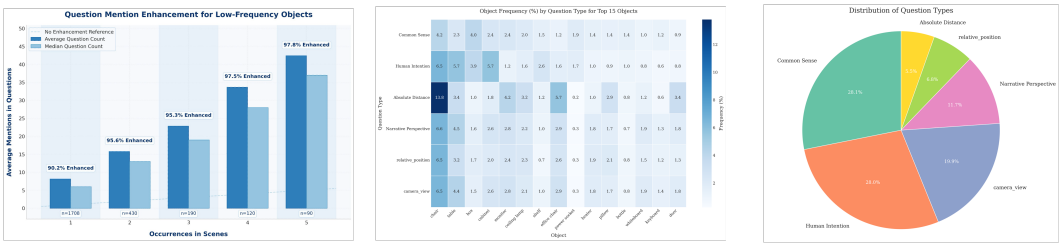


Figure 4: SURPRISE3D Dataset Statistics and Enhancements.

and object labels as context. The LLM proposes questions (e.g., 'What objects can be used for bathroom organization? A: (shower shelf)'). These are automatically filtered by basic rules (removing duplicates or nonsensical queries). The remaining questions are then passed on to human annotators for validation. Figure 3 illustrates this two-stage process: LLM outputs are checked by humans, who either accept them or refine the question/answer. Invalid or ambiguous queries (marked with ✗) are corrected or discarded, while valid ones (✓) become part of the dataset. This workflow scales up annotation efficiently while retaining human oversight.

4.2 Annotation Strategy and Quality Control

To ensure clarity and coverage, we apply: (1) **Disambiguation Guidelines:** We require that the questions be unambiguous. Annotators are given rules (e.g., explicitly naming reference objects) to avoid confusion. Queries with multiple plausible answers are rephrased or omitted. (2) **Rare-Object Sampling:** To improve the representation of uncommon classes, we identify objects with low overall frequency and sample scenes containing them. We then generate additional queries targeting these rare objects. This rare object boost increases their mentions (on average by $\sim$ 90–100%), improving the learning of these categories, as shown in Figure 4. (3) **Human Verification:** All annotations are double-checked. Spatial queries receive multiple annotator checks for the target mask, and discrepancies are resolved by consensus. LLM-generated queries are reviewed by editors to ensure that the question matches the intended object. This multi-stage review yields high-quality, consistent annotations.

5 Experiments

In this section, we first provide a brief overview of the advanced methods evaluated on our proposed dataset, models setting including zero-shot and fine tuning are also introduced. We then describe the evaluation metrics and criteria used to assess the performance of these methods on knowledge and spatial reasoning tasks. Finally, we present the quantitative and qualitative results obtained by the aforementioned methods to comparatively analyze their performance on the newly introduced reasoning challenges.

5.1 Baselines

Spatial reasoning tasks require a deeper understanding of semantic relationships and spatial configurations within 3D scenes. To evaluate the effectiveness of existing approaches on these tasks, we conduct comprehensive experiments on several advanced methods, including MLLMfor3D [20], 3D-Vista [61], Reason3D [10], ChatScene [18] and Intent3D [27]. All these methods take natural language questions or descriptions as input, and output masks or bounding boxes of objects that are the answer of given prompts in 3D scenes. MLLMfor3D [20] is a label-free paradigm of 3D understanding, it projects 2D pseudo-masks of objects to 3D scene. 3D-Vista [61] directly encourages the alignment of masked text with masked 3D scenes. Reason3D [10] utilizes a pre-trained LLM to process input point and text features, and predict segmentation masks of targeted objects. Intent3D [27] aligns the point features of scenes and candidate box features integrated with the encoded prompts. For **Reason3D** [21] and **UniDet3D** [28], superpoint extraction was necessary. We employed the Segmentor tool from previous works [13] to extract superpoints for each scene. However, due to the higher granularity of ScanNet++ [52] scenes, the default parameters for superpoint extrac-

Table 2: The results of various methods on different reasoning tasks under **zero-shot** setting.

Task Type	Metric	Method					
		MLLMfor3D	3D-Vista	Reason3D	Intent3D	ChatScene	Avarage
Knowledge Reasoning							
Common-sense	A25	**20.42**	6.14	6.97	10.01	7.86	10.28
	A50	**13.42**	6.14	3.11	3.24	4.01	5.98
	mIoU	**12.75**	–	4.79	–	–	8.77
Human Intention	A25	**22.38**	8.26	11.33	15.84	1.64	11.89
	A50	**13.62**	8.26	6.03	5.82	1.00	6.95
	mIoU	**11.91**	–	7.51	–	–	9.71
Spatial Reasoning							
Narrative Perspective	A25	**15.07**	6.50	8.39	2.65	0.00	6.52
	A50	**13.62**	6.50	4.88	0.77	0.00	5.15
	mIoU	**11.56**	–	5.63	–	–	8.60
Parametric Perspective	A25	4.25	3.65	**7.91**	2.62	0.00	3.69
	A50	3.20	3.65	**4.82**	0.79	0.00	3.12
	mIoU	2.93	–	**5.68**	–	–	4.31
Relative Position	A25	7.78	6.52	**9.57**	4.30	1.38	5.91
	A50	5.81	**6.52**	6.38	0.97	0.00	3.94
	mIoU	4.92	–	**6.78**	–	–	5.85
Absolute Distance	A25	**11.90**	7.61	9.10	2.41	1.39	6.48
	A50	**10.62**	7.61	2.60	0.74	0.00	4.32
	mIoU	**9.24**	–	5.25	–	–	7.25
Overall Reasoning							
Overall	A25	**13.63**	6.40	9.09	8.63	3.59	8.27
	A50	**10.05**	6.40	4.57	2.94	1.77	5.15
	mIoU	**8.89**	–	6.08	–	–	7.49

tion generated an excessively large number of superpoints, leading to significant GPU memory consumption.

5.2 Model settings

Zero-shot. To assess the capabilities of existing methods in performing knowledge-based and spatial reasoning tasks, we directly evaluate pre-trained models, including MLLMfor3D [20], 3D-Vista [61], Reason3D [21], ChatScene [18] and Intent3D [27] on our proposed dataset, without training or finetuning on our datasets. The inputs consist of 3D scenes represented as point clouds, paired with images and textual questions that focus on knowledge and spatial reasoning. As reported in previous work [20, 21, 27, 61, 18], Intent3D, 3D-Vista, ChatScene and Reason3D demonstrate partial abilities to understand human intentions and common sense knowledge. Moreover, MLLMfor3D and 3D-Vista show promising results in reasoning about relative object positions. However, none of them has been evaluated or shown to have the ability to comprehend narrative perspectives, parametric perspectives, or absolute spatial distances. Our dataset serves as the first benchmark to systematically evaluate these unexplored dimensions of spatial and knowledge understanding in existing 3D vision language models. Since their official implementations did not provide specific instructions for handling the ScanNet++ [52] dataset, we adapted the data preprocessing for each baseline based on their unique characteristics.

Fine tuning.To explore the ability of current advanced methods to learn knowledge and spatial reasoning, we finetune MLLMfor3D, 3D-Vista, Reason3D, and Intent3D on our dataset. The implementation details, as well as the data preprocessing steps, are provided in the appendix E.

5.3 Evaluation metrics

For the segmentation tasks (MLLMfor3D and Reason3D), we adopt both Mean Intersection over Union (MIoU) [45] and Accuracy (Acc) as evaluation metrics. MIoU measures the average overlap between the predicted and true 3D volumes, while Accuracy evaluates precision across varying

Table 3: The results of various methods on different reasoning tasks under **fine-tuned** setting.

Task Type	Metric	Method					
		MLLMfor3D	3D-Vista	Reason3D	Intent3D	ChatScene	Avarage
Knowledge Reasoning							
Common-sense	A25	25.41	19.36	18.08	**30.09**	13.56	21.30
	A50	**23.92**	19.36	8.97	15.22	4.37	14.37
	mIoU	**19.40**	–	11.92	–	–	15.66
Human Intention	A25	28.42	22.36	17.98	**31.16**	13.80	22.74
	A50	**24.51**	22.36	10.81	18.08	4.70	16.09
	mIoU	**19.47**	–	11.98	–	–	15.73
Spatial Reasoning							
Narrative Perspective	A25	22.38	**25.77**	11.30	24.91	13.98	19.67
	A50	20.40	**25.77**	7.71	15.58	4.28	14.75
	mIoU	18.44	–	9.03	–	–	13.74
Parametric Perspective	A25	10.04	3.87	11.52	**12.57**	4.58	8.52
	A50	**9.33**	3.87	7.59	6.13	2.36	5.86
	mIoU	7.50	–	**8.85**	–	–	8.18
Relative Position	A25	22.61	**23.86**	11.51	12.90	12.48	16.67
	A50	18.76	**23.86**	8.18	7.49	1.39	11.94
	mIoU	14.70	–	8.88	–	–	11.79
Absolute Distance	A25	**25.30**	18.92	12.80	5.75	7.42	14.04
	A50	**20.37**	18.92	4.45	1.86	3.71	9.86
	mIoU	**19.23**	–	8.16	–	–	13.70
Overall Reasoning							
Overall	A25	22.36	18.34	16.14	**23.98**	11.58	18.48
	A50	19.55	18.34	9.06	**13.10**	3.99	12.81
	mIoU	16.46	–	11.00	–	–	13.73

confidence thresholds, which we obtain from different intersection proportions ($e.g.$, 0.25 and 0.50) of the predicted and ground-truth volumes. For the detection tasks (3D-Vista, Intent3D, and ChatScene), we use Accuracy as the sole evaluation metric.

5.4 Results analysis

As shown in Table 2, the zero-shot results indicate that all models demonstrate relatively weak overall spatial reasoning capabilities compared with knowledge reasoning capabilities. Although these methods have been trained on spatial description datasets, such as ScanRefer [6] for most models and SQA3D [40] for ChatScene. After fine-tuning on our dataset, as presented in Tables 3 and Figure 5, all models demonstrate substantial improvements in reasoning abilities. This improvement is particularly significant in spatial reasoning, with an average performance increase of approximately three times. We hypothesize that this improvement is due to a key difference between the existing datasets and our proposed dataset. Although prior datasets contain questions involving spatial information, they often retain excessive semantic cues that can serve as shortcuts for models to arrive at answers without fully exercising spatial reasoning. In contrast, our dataset deliberately removes such shortcuts, ensuring that the training process emphasizes and encourages the spatial reasoning capability of the models. These findings highlight that current methods still have substantial room for improvement in terms of spatial reasoning, revealing an important avenue for future research.

6 Conclusion and Limitations

We introduced **SURPRISE3D**, a large-scale dataset and benchmark for evaluating spatial and knowledge reasoning in 3D scenes. Our benchmark defines the 3D Spatial Reasoning Segmentation (3D-SRS) task, which includes various types of spatial queries, including relative position, absolute distance, narrative perspective, and parametric viewpoint, as well as commonsense and human intention grounding. Through a dual annotation pipeline and rare-object enhancement, SURPRISE3D provides high-quality, diverse, and spatially language-query and segmentation pairs. Extensive

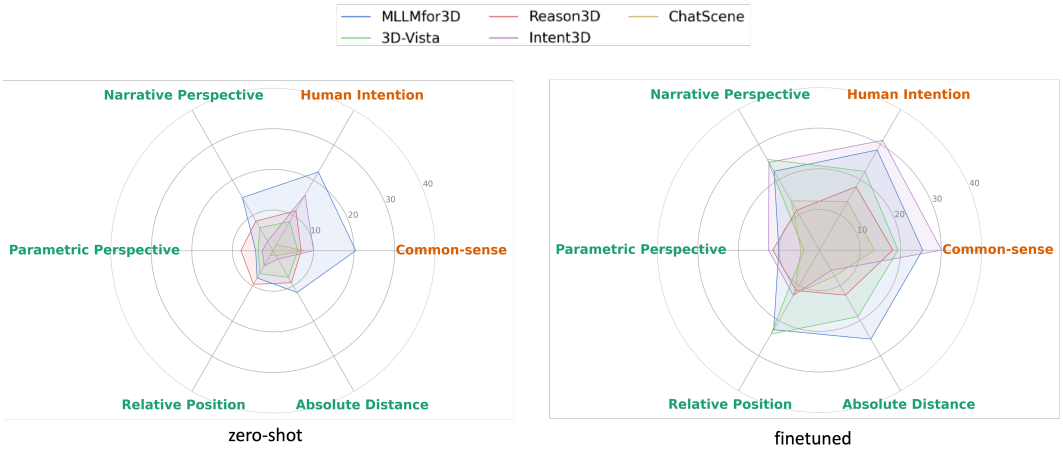


Figure 5: The comparison between zero-shot and fine-tuned models on all reasoning tasks.

analysis confirms its balanced coverage and strong potential for advancing spatial intelligence in 3D vision-language models.

Limitations. While our human annotation ensures quality, it limits scalability. Some query types (e.g., parametric view) may be less natural for real-world deployment. In addition, annotations are restricted to indoor scenes from ScanNet++, which may not generalize to outdoor or dynamic environments. We leave domain transfer, temporal reasoning, and multi-turn interaction as future directions.

References

[1] Ahmed Abdelreheem, Kyle Olszewski, Hsin-Ying Lee, Peter Wonka, and Panos Achlioptas. Scanents3d: Exploiting phrase-to-3d-object correspondences for improved visio-linguistic models in 3d scenes. *WACV*, 2022.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[3] 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 *ECCV*, pages 422–440. Springer, 2020.

[4] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *CVPR*, 2022.

[5] Alisson Azzolini, Hannah Brandon, Prithvijit Chattopadhyay, Huayu Chen, Jinju Chu, Yin Cui, Jenna Diamond, Yifan Ding, Francesco Ferroni, Rama Govindaraju, et al. Cosmos-reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558*, 2025.

[6] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *ECCV*, pages 202–221. Springer, 2020.

[7] Dave Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X. Chang. Scan2cap: Context-aware dense captioning in rgb-d scans, 2020.

[8] Haodong Chen, Runnan Chen, Qiang Qu, Zhaoqing Wang, Tongliang Liu, Xiaoming Chen, and Yuk Ying Chung. Beyond gaussians: Fast and high-fidelity 3d splatting with linear kernels. *arXiv preprint arXiv:2411.12440*, 2024.

[9] 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 *CVPR*, pages 26428–26438, 2024.

[10] Tianrun Chen, Chunan Yu, Jing Li, Jianqi Zhang, Lanyun Zhu, Deyi Ji, Yong Zhang, Ying Zang, Zejian Li, and Lingyun Sun. Reasoning3d–grounding and reasoning in 3d: Fine-grained zero-shot open-vocabulary 3d reasoning part segmentation via large vision-language models. *arXiv preprint arXiv:2405.19326*, 2024.

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

[12] Yasaman Etesam, Leon Kochiev, and Angel X. Chang. 3dvqa: Visual question answering for 3d environments. In *Conference on Robots and Vision (CRV)*, 2022.

[13] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 59:167–181, 2004.

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

[15] Shuting He, Henghui Ding, Xudong Jiang, and Bihan Wen. Segpoint: Segment any point cloud via large language model. In *ECCV*, pages 349–367. Springer, 2024.

[16] Yining Hong, Chunru Lin, Yilun Du, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. 3d concept learning and reasoning from multi-view images. In *CVPR*, pages 9202–9212, 2023.

[17] 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. In *NeurIPS*, pages 20482–20494, 2023.

[18] Haifeng Huang, Yilun Chen, Zehan Wang, Rongjie Huang, Runsen Xu, Tai Wang, Luping Liu, Xize Cheng, Yang Zhao, Jiangmiao Pang, et al. Chat-scene: Bridging 3d scene and large language models with object identifiers. In *NeurIPS*, 2024.

[19] 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 *ICML*, pages 20413–20451, 2024.

[20] Jiaxin Huang, Runnan Chen, Ziwen Li, Zhengqing Gao, Xiao He, Yandong Guo, Mingming Gong, and Tongliang Liu. Mllm-for3d: Adapting multimodal large language model for 3d reasoning segmentation, 2025.

[21] Kuan-Chih Huang, Xiangtai Li, Lu Qi, Shuicheng Yan, and Ming-Hsuan Yang. Reason3d: Searching and reasoning 3d segmentation via large language model. *arXiv preprint arXiv:2405.17427*, 2024.

[22] Pin-Hao Huang, Han-Hung Lee, Hwann-Tzong Chen, and Tyng-Luh Liu. Text-guided graph neural networks for referring 3d instance segmentation. *AAAI*, 35(2):1610–1618, May 2021.

[23] Wencan Huang, Daizong Liu, and Wei Hu. Dense object grounding in 3d scenes. *ACM Multimedia*, 2023.

[24] 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 *ECCV*, pages 289–310. Springer, 2024.

[25] Li Jiang, Shaoshuai Shi, and Bernt Schiele. Open-vocabulary 3d semantic segmentation with foundation models. In *CVPR*, pages 21284–21294, 2024.

[26] Xueying Jiang, Lewei Lu, Ling Shao, and Shijian Lu. Multimodal 3d reasoning segmentation with complex scenes. *arXiv preprint arXiv: 2411.13927*, 2024.

[27] Weitai Kang, Mengxue Qu, Jyoti Kini, Yunchao Wei, Mubarak Shah, and Yan Yan. Intent3d: 3d object detection in rgb-d scans based on human intention. In *ICLR*, 2025.

[28] Maksim Kolodiazhnyi, Anna Vorontsova, Matvey Skripkin, Danila Rukhovich, and Anton Konushin. Unidet3d: Multi-dataset indoor 3d object detection. In *AAAI*, volume 39, pages 4365–4373, 2025.

[29] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. 2023.

[30] Mingsheng Li, Xin Chen, Chi Zhang, Sijin Chen, Hongyuan Zhu, Fukun Yin, Zhuoyuan Li, Gang Yu, and Tao Chen. M3dbench: Towards omni 3d assistant with interleaved multi-modal instructions. In *ECCV*, pages 41–59. Springer, 2024.

[31] Yanbang Li, Ziyang Gong, Haoyang Li, Xiaoqi Huang, Haolan Kang, Guangping Bai, and Xianzheng Ma. Robotic visual instruction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12155–12165, 2025.

[32] Zeju Li, Chao Zhang, Xiaoyan Wang, Ruilong Ren, Yifan Xu, Ruifei Ma, and Xiangde Liu. 3dmit: 3d multi-modal instruction tuning for scene understanding. *arXiv preprint arXiv:2401.03201*, 2024.

[33] Ziwen Li, Jiaxin Huang, Runnan Chen, Yunlong Che, Yandong Guo, Tongliang Liu, Fakhri Karray, and Mingming Gong. Urbangs: Semantic-guided gaussian splatting for urban scene reconstruction. *arXiv preprint arXiv:2412.03473*, 2024.

[34] Xiongkun Linghu, Jiangyong Huang, Xuesong Niu, Xiaojian Ma, Baoxiong Jia, and Siyuan Huang. Multi-modal situated reasoning in 3d scenes. *NeurIPS*, 2024.

[35] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

[36] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023.

[37] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36:34892–34916, 2023.

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

[39] Wufei Ma, Haoyu Chen, Guofeng Zhang, Celso M de Melo, Alan Yuille, and Jieneng Chen. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. *arXiv preprint arXiv:2412.07825*, 2024.

[40] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. In *ICLR*, 2023.

[41] Phuc Nguyen, Tuan Duc Ngo, Evangelos Kalogerakis, Chuang Gan, Anh Tran, Cuong Pham, and Khoi Nguyen. Open3dis: Open-vocabulary 3d instance segmentation with 2d mask guidance. In *CVPR*, pages 4018–4028, 2024.

[42] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *CVPR*, pages 815–824, 2023.

[43] Zhipeng Qian, Yiwei Ma, Jiayi Ji, and Xiaoshuai Sun. X-refseg3d: Enhancing referring 3d instance segmentation via structured cross-modal graph neural networks. In *AAAI*, volume 38, pages 4551–4559, 2024.

[44] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d: Mask transformer for 3d semantic instance segmentation. In *ICRA*, pages 8216–8223. IEEE, 2023.

[45] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE TPAMI*, 39(4):640–651, 2017.

[46] Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, Xihui Liu, Cewu Lu, Dahua Lin, and Jiangmiao Pang. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In *CVPR*, 2024.

[47] Zehan Wang, Haifeng Huang, Yang Zhao, Ziang Zhang, and Zhou Zhao. Chat-3d: Data-efficiently tuning large language model for universal dialogue of 3d scenes. *arXiv preprint arXiv:2308.08769*, 2023.

[48] Changli Wu, Yiwei Ma, Qi Chen, Haowei Wang, Gen Luo, Jiayi Ji, and Xiaoshuai Sun. 3d-stmn: Dependency-driven superpoint-text matching network for end-to-end 3d referring expression segmentation. In *AAAI*, volume 38, pages 5940–5948, 2024.

[49] Yanmin Wu, Xinhua Cheng, Renrui Zhang, Zesen Cheng, and Jian Zhang. Eda: Explicit text-decoupling and dense alignment for 3d visual grounding. In *CVPR*, pages 19231–19242, 2023.

[50] Xu Yan, Zhihao Yuan, Yuhao Du, Yinghong Liao, Yao Guo, Zhen Li, and Shuguang Cui. Comprehensive visual question answering on point clouds through compositional scene manipulation, 2023.

[51] 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, 2024.

[52] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *ICCV*, pages 12–22, 2023.

[53] Chandan Yeshwanth, David Rozenberszki, and Angela Dai. Excap3d: Expressive 3d scene understanding via object captioning with varying detail. *arXiv preprint arXiv: 2503.17044*, 2025.

[54] Zhihao Yuan, Xu Yan, Zhuo Li, Xuhao Li, Yao Guo, Shuguang Cui, and Zhen Li. Toward explainable and fine-grained 3d grounding through referring textual phrases. *arXiv preprint arXiv:2207.01821*, 2022.

[55] Ed Zalta. Stanford encyclopedia of philosophy. 2012.

[56] Haochen Zhang, Nader Zantout, Pujith Kachana, Zongyuan Wu, Ji Zhang, and Wenshan Wang. Vla-3d: A dataset for 3d semantic scene understanding and navigation. *arXiv preprint arXiv: 2411.03540*, 2024.

[57] Haochen Zhang, Nader Zantout, Pujith Kachana, Zongyuan Wu, Ji Zhang, and Wenshan Wang. Vla-3d: A dataset for 3d semantic scene understanding and navigation. *arXiv preprint arXiv:2411.03540*, 2024.

[58] Jiahui Zhang, Yurui Chen, Yanpeng Zhou, Yueming Xu, Ze Huang, Jilin Mei, Junhui Chen, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, et al. From flatland to space: Teaching vision-language models to perceive and reason in 3d. *arXiv preprint arXiv:2503.22976*, 2025.

[59] Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3d: Exploring unified 3d representation at scale. *arXiv preprint arXiv:2310.06773*, 2023.

[60] Chenming Zhu, Tai Wang, Wenwei Zhang, Kai Chen, and Xihui Liu. Scanreason: Empowering 3d visual grounding with reasoning capabilities. *arXiv preprint arXiv: 2407.01525*, 2024.

[61] 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 *ICCV*, pages 2911–2921, 2023.

A Data Anaylsis

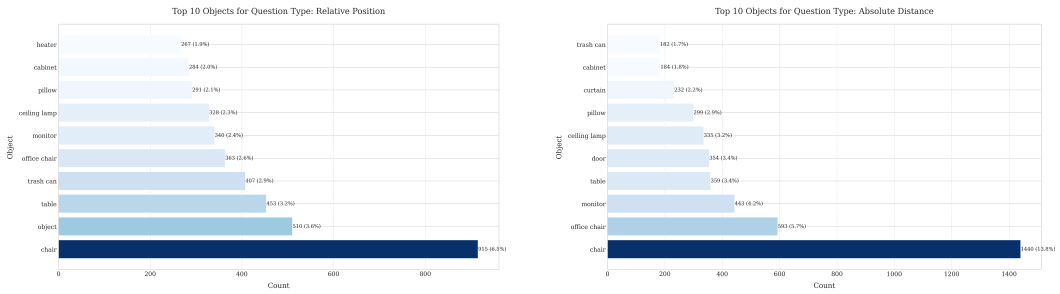
We conduct a detailed analysis of object distributions across six distinct reasoning question types in the **SURPRISE-3D** dataset. The results are visualized in Figures 6, 7, and 8, each highlighting the Top-10 most queried objects for two representative reasoning tasks.

Figure 6 focuses on *absolute distance* and *relative position* queries. In both subtasks, the object `chair` dominates, appearing in 1,440 (13. 8%) and 915 (6. 5%) queries, respectively. This prominence probably reflects its frequent presence and stable placement within indoor scenes. In the absolute distance setting, other frequently queried objects include `office chair` (5.7%), `monitor` (4.2%), and `table` (3.4%), all of which are static and spatially anchored elements. The relative position task features more generic terms such as `object` (3.6%) and `trash can` (2.9%), indicating that spatial relations are sometimes queried in less specific terms. These findings suggest that spatial reasoning is closely tied to furniture and desktop objects that are familiar and contextually grounded.

Figure 7 extends the spatial reasoning perspective to viewpoint-based descriptions. In the *narrative perspective* sub-task (i.e., first-person view), `chair` again leads with 1,535 queries (6.6%), followed by `table`, `office chair`, and `monitor`. These results underscore the natural alignment between egocentric vision and human interaction with seating and work surfaces. Similarly, the *parametric perspective* task (i.e., camera-relative view) highlights `chair` (6.5%) and `table` (4.4%) as dominant targets. Other notable objects such as `monitor`, `object`, and `cabinet` reflect a mixture of coarse and fine-grained references. Collectively, these patterns indicate that perspective-based spatial reasoning queries tend to center on objects within immediate visual or functional reach, validating our emphasis on camera-pose alignment in dataset construction.

Figure 8 presents two complementary reasoning categories: *common sense* and *human intention*. In both cases, `chair` remains the most frequently queried object, with 9,474 (4.2%) common sense and 9,882 (6.5%) human intention queries. Notably, in the human intention setting, objects such as `cabinet` (5.7%), `table` (5.7%), and `box` (3.9%) also emerge as key targets. This distribution suggests a strong correlation between human actions and functional furniture or storage items. For instance, queries might include "Where can I find something inside the cabinet?" or "What is the user sitting on?". In the common sense sub-task, a more diverse set of objects appears, including `monitor`, `power socket`, `heater`, and `bottle`, indicating broader expectations around affordances, typical object usage, and environment context.

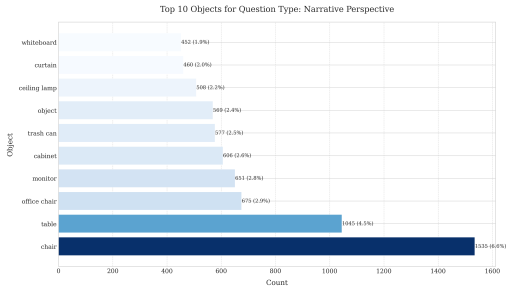
Summary Insights. Across all six reasoning categories, `chair` consistently ranks as the most queried object, underscoring its central role in indoor human-scene interactions. Other recurring objects include `table`, `monitor`, `cabinet`, and `office chair`, suggesting a core subset of spatially grounded and semantically meaningful targets. These observations guide the selection of high-coverage object classes for training, support curriculum learning strategies focused on frequently referenced targets, and motivate the integration of egocentric and functional priors in model development.



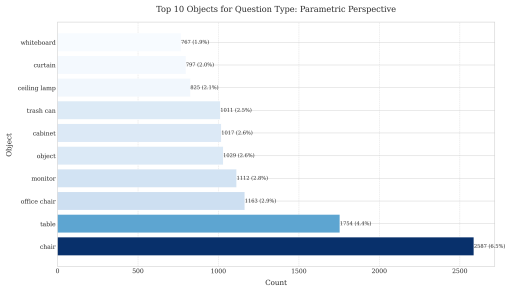
(a) Top 10 objects for relative position.

(b) Top 10 objects for Absolute Distance.

Figure 6: Top objects associated with spatial reasoning queries, highlighting absolute distance and relative positional understanding.

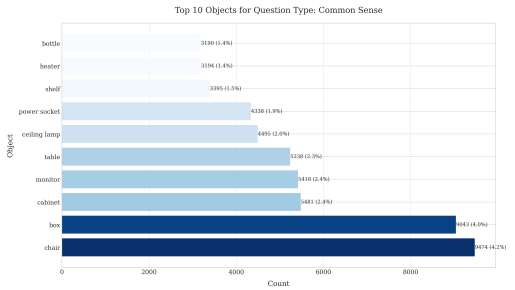


(a) Top 10 objects for narrative perspective (first-person view).

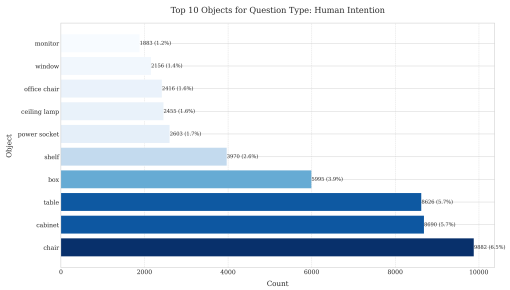


(b) Top 10 objects for parametric perspective (camera pose).

Figure 7: Top objects in viewpoint-based spatial reasoning queries, emphasizing egocentric and camera-relative descriptions.



(a) Top 10 objects for common sense reasoning.



(b) Top 10 objects for human intention reasoning.

Figure 8: Top objects in knowledge-based reasoning queries. Common sense tasks emphasize general knowledge, while human intention tasks focus on purposeful interaction.

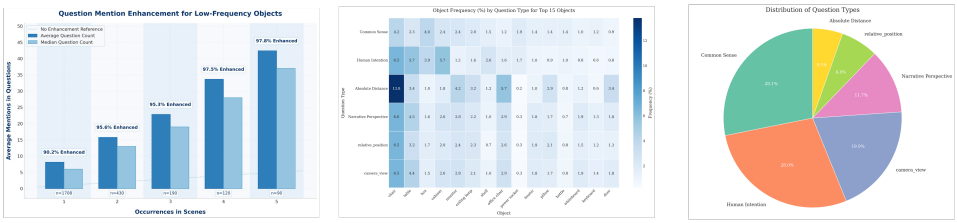


Figure 9: SURPRISE-3D Dataset Statistics and Enhancements.

B Annotation pipeline

B.1 Philosophy of Human Annotaion

(1) Object Name Bias: In most 3D referring expression and QA datasets, the language query explicitly names the target object category. Models can exploit this by simply detecting the named class (e.g. 'chair') without truly understanding the spatial context or attributes. This overreliance on object name cues means that the task sometimes reduces to object classification rather than relational reasoning.

(2) Coarse Localization Annotations: Many benchmarks use coarse labels like 3D bounding boxes or object IDs as the target output (e.g. ScanRefer, ReferIt3D, Intent3D). Such annotations do not evaluate the ability of a model to precisely delineate the shape of the object or handle overlapping objects. Fine-grained segmentation masks are largely missing, which lowers the bar for spatial understanding (since predicting a loose box is easier than predicting an exact mask).

(3) Limited Spatial Language: While some datasets include relative spatial phrases (e.g., "next to the bed"), they often lack a rich variety of spatial reasoning challenges. The absolute distance (reference cues such as 'on the north wall') is usually absent. Thus, models are not fully tested on

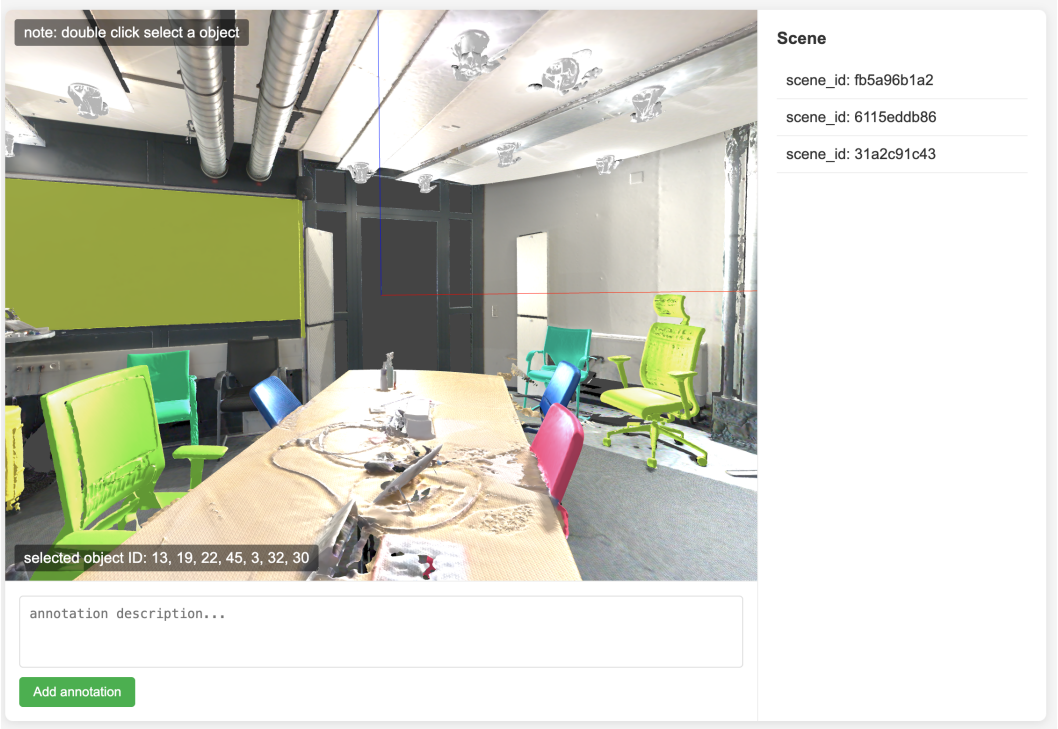


Figure 10: UI annotation Interface.

understanding diverse spatial relations or room-centric directions, and they struggle when queries require reasoning about geometric layout beyond simple pairwise relations.

(4) Insufficient Ambiguity and Intent Understanding: Disambiguation between multiple similar objects is a core challenge often underrepresented. ReferIt3D explicitly ensured contrasting instances in each scene, but many other datasets have numerous utterances referring to objects that are unique in their scene (trivializing the task). Moreover, genuine human intentions or functional descriptions (e.g. 'something to cook with' for a stove) are not prevalent in most benchmarks, aside from specialized cases like Intent3D which still treat it as a detection task. As a result, current models are not fully tested in understanding the purpose or intent behind a reference, especially in conjunction with spatial cues.

B.2 Language Annotation

To ensure consistency, we followed a standardized three-step protocol for the spatial query annotations (illustrated in Figure 2). (1) Select a Scene and Viewpoint: The annotator begins by selecting one scene and positioning themselves (virtually) at a vantage point that provides a clear context. They also consider their orientation in the scene – for example, an annotator might imagine "standing in front of the sofa and facing the blackboard" as their narrative context. This step establishes the reference frame (what is "left" vs "right", etc.) and highlights salient scene anchors (notable objects or landmarks) visible from that view. (2) Compose the Description: Next, the annotator writes a spatial query referring to a target object without naming it outright. The description leverages the established context and anchors. For instance, given the orientation above, they might describe "an object on the left used for teaching," which implicitly refers to the blackboard without saying "blackboard." Alternatively, they might use purely relative terms like "the table closest to the bed", if multiple tables are present. The key is to include enough information (spatial relations, attributes, or object use) so that the query uniquely identifies one object in the scene. We emphasize relational and perspective words here: terms like "closest/farthest," "left/right of X," "behind Y," etc., as well as descriptive qualifiers (e.g., color, affordance) to avoid ambiguity. (3) Segment the Object: Finally, the annotator highlights the referred object by creating a 3D segmentation mask. They ensure the correct object (and only that object) is masked, and then save the mask along with the query and the camera

Figure 11: An example of the annotation result for the **Narrative Perspective (NP)** question type. The **left side** illustrates the global position and orientation of the camera within the 3D scene, where the red sphere represents the camera's position, and the cone-shaped arrow indicates its orientation. The **right side** shows the scene as observed from the camera's perspective. In both images, the green object represents the annotated object, which corresponds to the answer for the perspective description and question.

parameters for that description. Throughout this process, we enforced strict annotation criteria to maintain the clarity and uniqueness of the referring expressions. In particular: Avoid ambiguous viewer-relative terms: Queries were not allowed to depend on uncertain notions of distance from the observer. Annotators should not say "the far object" or "the big chair near me" without a clear reference, since "far" or "near" might change with viewpoint. Any egocentric directional term had to be grounded in the fixed orientation (e.g., "to your left" is fine if the orientation is stated). Ensure uniqueness within context: Each query had to single out one object. If a description was initially ambiguous (e.g., "the plant in the corner" when two plants are in the corner), annotators would refine it by adding a distinguishing relation: "the closest plant on the left", for example, if there are multiple plants. We instructed that referring terms (like "the one on the left," "the taller lamp") must be unique given the scene, leveraging superlatives or additional attributes as needed to achieve this. Keep descriptions concise yet specific: The goal was a minimal description that still disambiguates the target. Annotators were told not to add extraneous detail beyond what's necessary. For instance, "the small red stool behind the sofa" is acceptable (provides size, color, and relation to sofa to pick out the stool), but adding unrelated details or full narratives is discouraged. Define distance relations by objects, not the camera: For phrases like "closest" or "furthest," we clarified that distance is measured between objects in the scene, not relative to the annotator's viewpoint. Figure 2 (orange inset) shows how annotators determined the "closest X" – by comparing actual distances among candidates. For example, "the table closest to the bed" means the table with the smallest 3D distance to the bed (out of all tables in the scene), irrespective of which table appears closest in the camera view. This rule prevents confusion where an object might look near in the 2D view but isn't the nearest in 3D terms. Following this protocol, we collected a large set of (query, mask) pairs covering diverse spatial reasoning cases. The resulting dataset includes queries that span a range of difficulty – from simple ones like "the blue object on the table" (which require identifying an attribute and a support relation) to complex ones like the earlier example of an egocentric perspective ("Facing away from the door toward the cabinet, the item on your left used for teaching"). We include several illustrative examples of queries and their corresponding 3D segmentations in the Appendix for clarity.

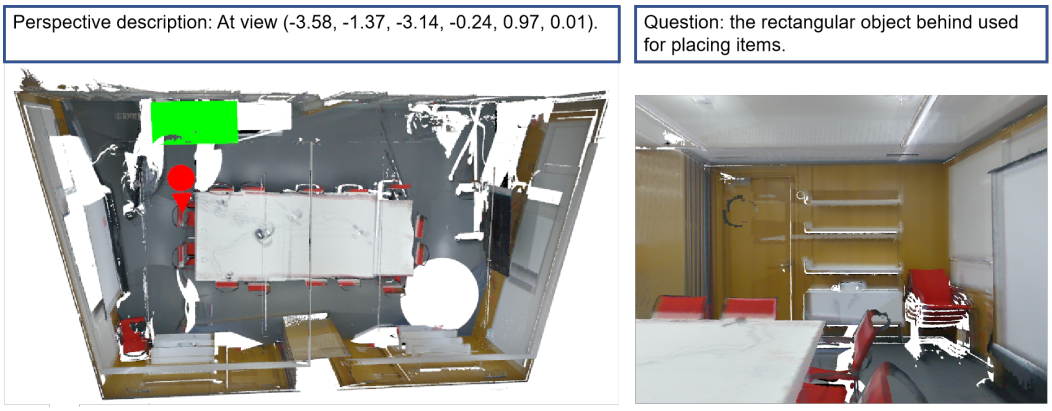


Figure 12: An example of the annotation process for the **Parametric Perspective** question type. The **left side** illustrates the camera's position and orientation within the 3D scene, where the red sphere represents the camera's position, and the cone-shaped arrow indicates its orientation. This position is challenging for annotators to describe accurately using natural language. The **right side** displays the view observed from the current camera perspective. The green object in the image is the annotated object, corresponding to the perspective description and question provided by the annotator.

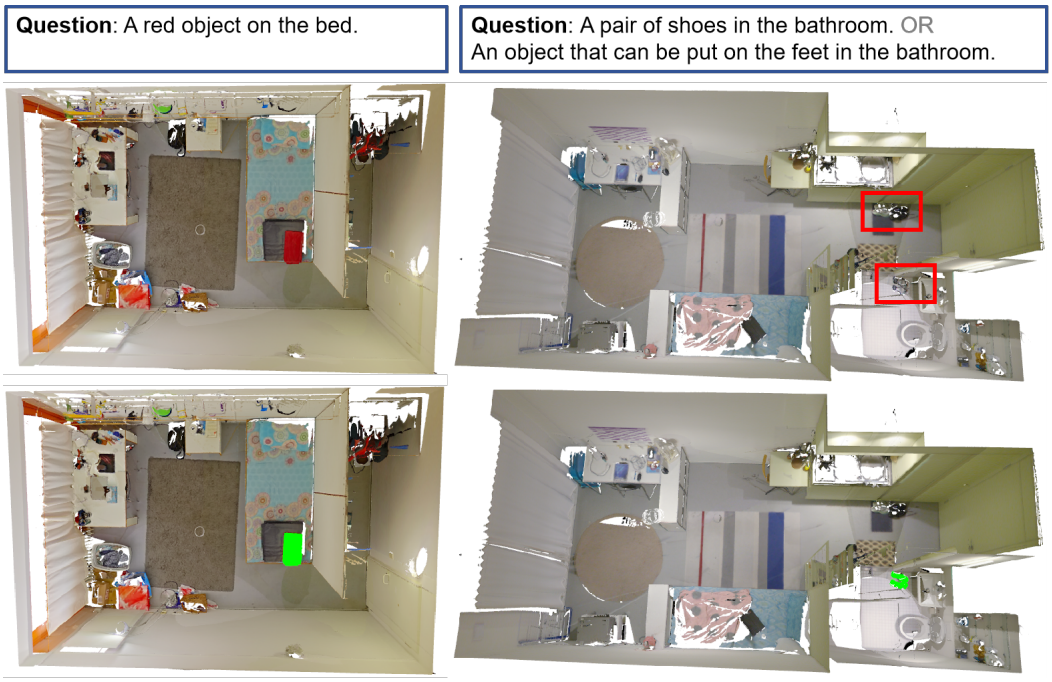


Figure 13: An example of the annotation result for the **Relative Position** question type. The **left side** of the figure demonstrates a case where the answer is described solely based on its physical attributes in addition to relative position relationships. In this situation, it is unnecessary for the answer to have multiple instances of its kind in the scene, as there are other objects with the same physical attributes present (e.g., the red object in the bottom-left corner and the red object in the top-right corner). The **right side** of the figure shows a case where multiple similar objects (e.g., several pairs of shoes, marked within the red boxes) exist in the scene. In this case, a semantic description is used to differentiate the target object. By incorporating spatial position relationships, the annotation avoids shortcuts and ensures the specified pair of shoes can be uniquely identified.

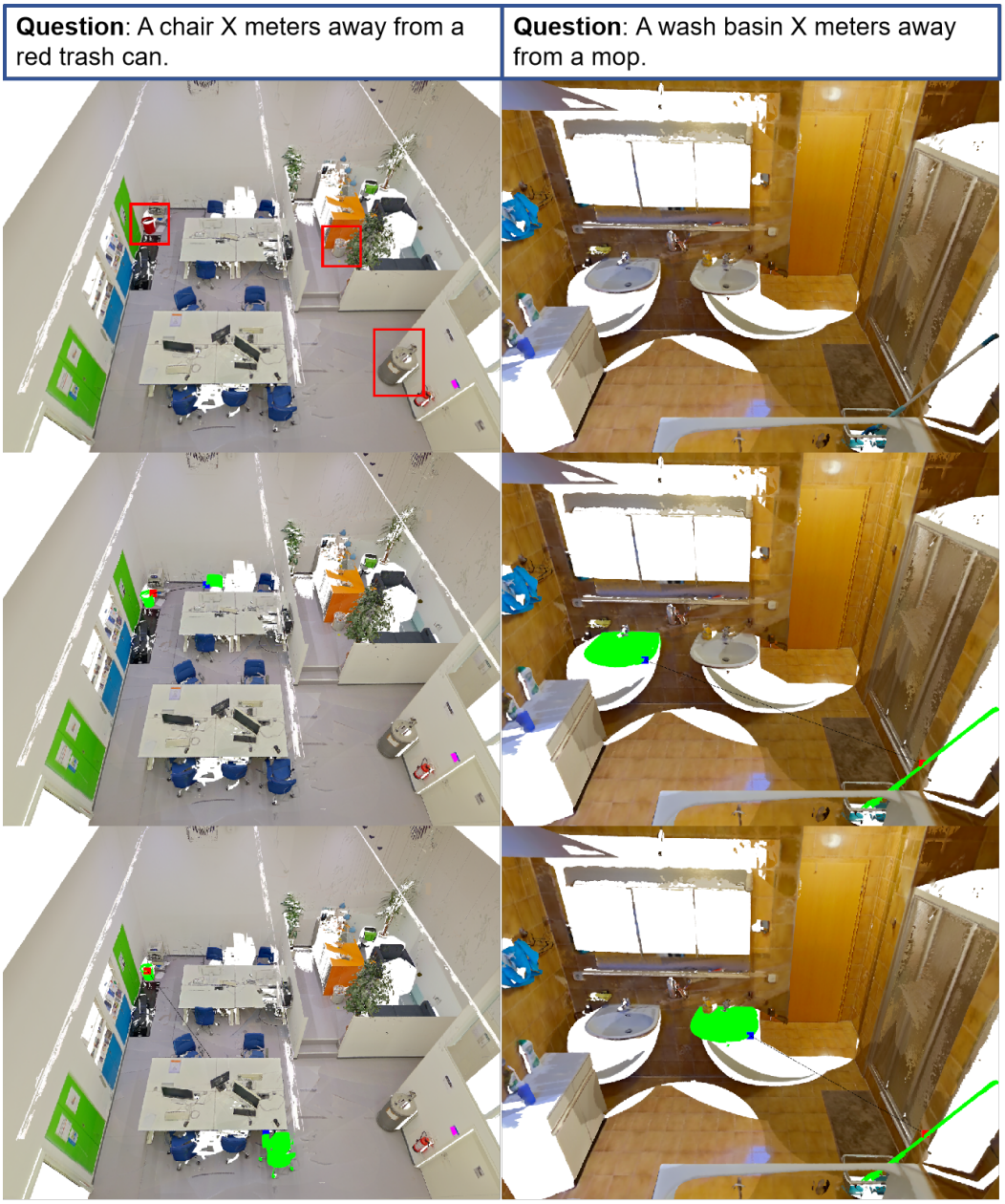


Figure 14: An example of the annotation process for **Absolute Distance** questions. The **first row** shows the original scene, the **second row** illustrates the first set of absolute distance annotations, and the **third row** depicts the second set of annotations. In each case, the red point represents the reference object chosen for calculating the absolute distance, while the blue point represents the target object. The left column shows a scenario where multiple trash bins exist in the scene; the annotator distinguishes a specific trash bin as the reference through a detailed description. The target object in this case is a chair, which is one of several instances in the scene. On the right, the scene contains only one mop, allowing it to be directly selected as the reference object without requiring a description. The target object in this case is a sink, which, like the chair, has multiple instances in the scene.

C Detail of Spatial Reasoning Annotations

C.1 Narrative Perspective.

For the **NP** question type, annotators were instructed to identify an unambiguous viewpoint within the scene that could be naturally described using human language. The annotation process involved the following steps:

- Annotators navigated through the 3D scene to locate a viewpoint that they deemed as the best representation for generating a clear and unambiguous natural language description.
- They adjusted the camera view by dragging and moving within the scene until they reached the desired perspective.
- Once the perspective was finalized, annotators provided a detailed description of the scene from that viewpoint, ensuring that the description was precise, complete, and free of ambiguities.

An example of this process is illustrated in Figure 11. The **left side** of the figure shows the global position and orientation of the camera within the 3D scene. The **right side** of the figure visualizes the scene as observed from the selected camera perspective.

The goal of this annotation process was to ensure the creation of high-quality data that links natural language descriptions to specific, unambiguous perspectives within 3D environments.

C.2 Parametric Perspective.

Every **NP** question can naturally be converted into a **PP** question. When annotators navigate the scene and adjust the camera to the desired viewpoint for a narrative perspective, the camera extrinsic parameters (i.e., **position**, **rotation**, and **up vector**) for that viewpoint are saved. By replacing the corresponding narrative perspective description with the format ‘‘`At perspective(position, rotation, upvector)`’’, we can generate the **PP description**. Combining this description with the same question from NP yields the corresponding **PP** question.

In addition to this conversion process, annotators were tasked with identifying other viewpoints within the scene that might not be easily describable using natural language. For such perspectives, annotators navigated to the desired viewpoint by dragging and moving the camera, and directly formulated a question based on the visual information.

An example of this process is illustrated in Figure 12. The **left side** of the figure shows the camera's position and orientation within the 3D scene. Notably, this position is challenging for annotators to describe accurately using natural language. The **right side** of the figure visualizes the scene as observed from the current camera perspective.

This dual approach ensures that the dataset captures a wide range of perspectives, including both linguistically describable viewpoints and purely parametric ones, enhancing the diversity and completeness of the dataset.

C.3 Relative Position.

Following the methodology proposed in **VLA-3D** [57], relative position relationships include **near**, **closest**, **furthest**, **above**, **on**, **in**, and **below**. And three key requirements for relative position questions are:

- **View-independent:** The relation predicate for the target object must not depend on the perspective from which the scene is viewed.
- **Unambiguous:** There must be only one possible target object in the specified region, ensuring the reference is clear and precise.
- **Minimal:** In accordance with Grice's maxim of manner [55], the statements should use the minimal number of descriptors necessary to disambiguate the target object.

On top of these requirements, we impose an additional constraint: annotators are prohibited from explicitly stating the answer within the question itself. Instead, they must either:

1. Provide a physical attributes description (e.g., color, size, shape, or texture) of the target object. An example of this approach is shown in Figure 13 (left), where the object is described based on its physical characteristics.

2. For scenes that contain multiple objects of the same type, annotators can directly mark the target object within the scene and append a # symbol at the end of the question. In such cases, we employ a

large language model (LLM) during post-processing to either generate a semantic description of the object or directly label it with its name. An example of this process is illustrated in Figure 13 (right).

This approach ensures that the relative position questions are both precise and contextually rich, while avoiding redundancy or ambiguity. By combining physical descriptors with LLM-based post-processing for complex cases, we achieve a robust and scalable annotation pipeline for relative position tasks.

C.4 Absolute Distance.

For absolute distance questions, we require annotators to select a single, unique object within the scene as the **reference object**. If the object is not inherently unique, annotators must provide a description that disambiguates it into a unique target. This reference object serves as the anchor for the distance calculation. Annotators then select another set of objects of the same type within the scene as the **target objects**. To aid understanding, the visualization of this process is shown in Figure 14. The **left side** of Figure 14 illustrates a scenario where the reference object is inherently unique within the scene. The **right side** of Figure 4 demonstrates a case where the reference object is made unique through a detailed description provided by the annotator. These visualizations highlight the steps annotators take to ensure unambiguous reference selection and accurate distance measurement.

To calculate the distance, annotators identify specific points on both the reference object and one of the target objects that they believe best represent the distance between the two. These points are then marked, and the distance between them is recorded as the absolute distance between the objects.

The question format is subsequently processed into the form: `‘‘Target object that is D units away from the reference object?’’`. This design ensures that the model cannot rely on shortcuts to directly identify the target object; instead, it must leverage spatial reasoning and calculate distances to determine which object satisfies the question.

We hypothesize that models trained on such questions will develop a stronger spatial understanding of the scene, as they are required to integrate information about object positions, relative distances, and scene geometry to arrive at the correct answer.

C.5 Post Processing.

For preprocessing, we address the overlap between the **narrative perspective** and the **parametric perspective** by dividing the issues into two separate formats:

1. **Parametric Perspective Format:** Parametric perspective problems are standardized into the format:

   ```
   "at perspective (X, X, X, X, X, X, X, X, X), + original question"
   ```

   This ensures consistency in representing parametric data across the dataset.

2. **Object Name Replacement:** For descriptions without the # symbol, the original object name is inserted directly into the sentence. For example:

   ```
   "Facing away from the door upon entry, the nearest [32]"
   ```

   is transformed into:

   ```
   "Facing away from the door upon entry, the nearest [chair]"
   ```

   Following C.3, the object names are processed in the prompt with a 50% probability to:

   - Retain the original object name (e.g., "chair").
   - Replace it with a functional description (e.g., "object for sitting").

3. **Absolute Distance Adjustment:** Descriptions involving absolute distances are reformatted for clarity. For example:

   ```
   "The [table] 1.92 meters away from the door"
   ```

This ensures consistent representation of spatial relationships.

After initial preprocessing, we employ a Large Language Model (LLM) for further refinement. The specific prompt used for this step is detailed in Section G. The LLM is tasked with ensuring natural language fluency, applying the appropriate object descriptions, and verifying consistency across the dataset.

D Detail of Knowledge Reasoning Annotations

Knowledge Reasoning. For Knowledge Reasoning questions, we automatically generate the questions using a Large Language Model (LLM). The process begins by simplifying the data for each scene based on the categories and IDs of all instances in that scene. Each scene is summarized as follows:

- The scene contains N objects, distributed across C categories:
  - x_1 bed
  - x_2 blanket
  - x_3 set of books
- Instance details:
  - **BED** (x_1):
    - ID: id_1
  - **BLANKET** (x_2):
    - ID: id_2
  - **BOOKS** (x_3):
    - ID: id_3

Next, we calculate the global frequency of each instance across all scenes. Objects with a total occurrence count of less than 20 are defined as **low-frequency objects**. Using this preprocessed data, we employ **DeepSeek V3** [35] to generate questions for each scene. The question generation is divided into two categories: Common Sense Questions G and Human Intention Questions G.

The DeepSeek V3 API iterates through all scene data to produce the initial set of questions. However, this initial dataset often contains issues that need to be addressed:

1. The category of the answer object is explicitly mentioned in the question.
2. Logical errors, where the answer does not align with human intuition or reasoning standards.

To address these problems, we utilize **GPT-4o** [2] to clean the initial dataset. The cleaning process uses a specific prompt G. If the modified questions are still judged to contain issues, we perform a second round of corrections using targeted prompts based on the identified errors:

- For **Direct Mention Errors** G.
- For **Logical Errors** G.

Finally, for the rare cases where questions still contain answers, we apply an enhanced prompt G. This step ensures that all remaining problematic questions are rigorously refined to meet the desired quality and reasoning standards.

Through this multi-step process, we ensure that the Knowledge Reasoning dataset is free from logical inconsistencies and direct answer mentions, enabling the generation of high-quality questions that require sophisticated reasoning to solve.

E Implementation of Baselines

E.1 Data Preprocess

We evaluated five baseline models: MLLMfor3D [20], 3D-VISTA [61], Reason3D [21], Intent3D [27], and ChatScene [18]. Since their official implementations did not provide specific instructions for

handling the ScanNet++ [52] dataset, we adapted the data preprocessing for each baseline based on their unique characteristics. The details of our data preprocessing pipeline are described below.

E.1.1 Bounding Box Prediction

For **3D-VISTA** [61] and **Intent3D** [27], pretrained models were required to predict bounding boxes (predicted bbox). To achieve this, we utilized UniDet3D [28] to predict the bounding boxes for each scene in the ScanNet++ [52] dataset.

E.1.2 Instance Segmentation and Feature Extraction

For **ChatScene** [18], pretrained models were required to predict both per-scene instance segmentation and 3D features. Following the methodology outlined in the original paper, we used Mask3D [44] to predict instance segmentation and Uni3D [59] to extract 3D features.

E.1.3 Superpoint Extraction

For **Reason3D** [21] and **UniDet3D** [28], superpoint extraction was necessary. We employed the Segmentor tool from previous works [13] to extract superpoints for each scene. However, due to the higher granularity of ScanNet++ [52] scenes, the default parameters for superpoint extraction generated an excessively large number of superpoints, leading to significant GPU memory consumption. To mitigate this issue, we adjusted the parameters from `kThresh=0.01` and `segMinVerts=20` to `kThresh=10` and `segMinVerts=200`. This adjustment reduced memory consumption while maintaining sufficient granularity. Additionally, we modified the dataloaders in the source code of different methods to ensure compatibility with the ScanNet++ [52] dataset.

E.1.4 Annotation Processing for Intent3D

For **Intent3D** [27], we made several modifications to the annotation processing script to handle the specific characteristics of our dataset. These include:

- **Handling Missing "I" in Sentences:** The original implementation assumed that every sentence would contain the subject "I" and relied on it for verb-object parsing. To handle cases where "I" was absent, we introduced a fallback mechanism that assigns a default value of `all_verb_obj = [(1, (0, 0, 0, 0))]` when no valid verb-object pairs are found.
- **Handling `camera_view` Question Type:** For annotations with the `camera_view` question type, we skipped the verb-object parsing altogether and directly assigned the default value of `all_verb_obj = [(1, (0, 0, 0, 0))]`.
- **Improving Robustness:** We updated the code to initialize the subject variable (`i_subject`) as `None` to avoid undefined variable errors. This ensures robust handling of sentences where "I" is not present.

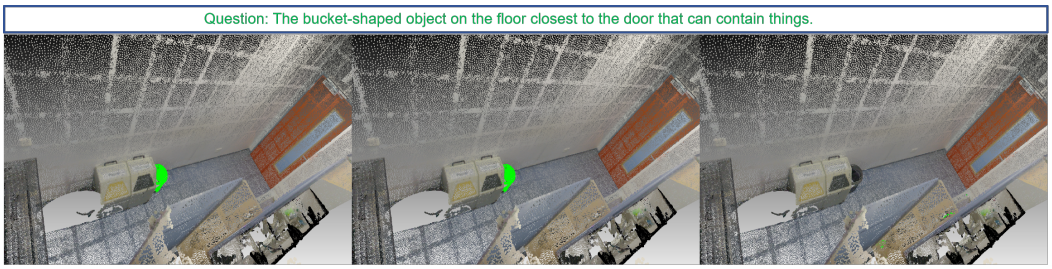
These modifications allowed us to adapt **Intent3D** [27] to the more complex and diverse annotations in the ScanNet++ [52] dataset while ensuring compatibility and robustness.

E.2 Implementation Details

For all methods, we followed the official hyperparameter settings unless otherwise specified. The only exception was that we reduced the batch size for **3D-VISTA** and **Reason3D**, as our hardware (A100 GPU with 40GB memory) could not accommodate the default settings due to memory constraints. Additionally, for **Reason3D**, we reduced the number of epochs from 100 to 15 due to its excessively long training time.

E.2.1 Handling Task-Specific Constraints

- **3D-VISTA:** Since **3D-VISTA** can only infer a single object at a time, tasks involving multiple objects required a specific adjustment. For such cases, we selected the label that appeared most frequently in the ground truth answer as the input to the model.



(a). Ground Truth (b). Reason3D + SURPRISE (c). Reason3D w/o SURPRISE

Figure 15: Visualization of a reasoning result. The figure illustrates three outputs: **Left:** Ground truth (GT) result, **Middle:** Reason3D's output after being trained on our dataset, and **Right:** Reason3D's output using the official checkpoint. The result demonstrates that after training on our dataset, Reason3D is better able to understand both **knowledge reasoning** (e.g., identifying functional attributes like "can contain things") and **spatial reasoning** (e.g., determining the object closest to the door). In contrast, the official checkpoint struggles with these aspects, highlighting the effectiveness of our dataset in enhancing Reason3D's reasoning capabilities.

- **ChatScene:** For **ChatScene**, only predicted instance masks with a mean Intersection over Union (mIoU) above a certain threshold compared to the ground truth were retained for training. This filtering process reduced the size of the final training set to approximately 10k samples.

E.2.2 Zero-Shot Experiments

For zero-shot experiments, we directly used the pretrained checkpoints provided on the official repositories of each method for evaluation without further fine-tuning or modification.

E.3 Visualization of Results.

The visualized results are shown in Figure 15, which demonstrate the effectiveness of training Reason3D using our dataset. After being trained on our dataset, Reason3D exhibits significantly improved capabilities in both **knowledge reasoning** and **spatial reasoning** tasks.

F Future Work

Our dataset provides a valuable resource for exploring spatial reasoning and knowledge understanding in 3D environments. However, there are still several challenges and potential directions for future work:

F.1 Addressing Parametric Perspective Challenges

One of the key issues lies in solving questions related to the **parametric perspective**, which currently proves to be highly challenging. This difficulty likely arises from the inherent complexity of using numerical representations to describe positions and orientations, which are hard for models to interpret effectively. Future work could consider embedding these parametric representations directly into the input features, allowing the model to better process and understand them. Alternatively, aligning these parametric representations with their corresponding narrative perspective descriptions may help bridge the gap and improve comprehension.

F.2 Advancing Spatial Understanding in 3D Scenes

Another critical direction is encouraging more models to explore how to better understand spatial information in 3D scenes. This includes developing methods that can handle complex relationships between objects, such as relative positions, orientations, and distances, as well as integrating multi-modal data (e.g., visual, textual, and parametric information) to achieve a more holistic understanding of the 3D environment.

F.3 Towards Unified Representations

Future research could also focus on creating unified representations that seamlessly combine parametric and narrative perspectives. Such representations would allow models to leverage both numerical and descriptive spatial information, potentially improving their reasoning capabilities in tasks involving complex 3D spatial relationships.

F.4 Expanding Benchmarks and Tasks

Finally, we encourage the development of additional benchmarks and tasks based on our dataset to further evaluate and enhance models' abilities to reason about 3D spaces. These could include tasks specifically designed to test models' understanding of fine-grained spatial relationships or their ability to generalize across diverse 3D environments.

By addressing these challenges and exploring these directions, we believe future research can unlock new possibilities for leveraging our dataset to advance spatial reasoning and knowledge understanding in 3D contexts.

G Prompt

Spatial Reasoning Prompt (Non-ABS Data)

Please process the following JSON data by:

1. **For descriptions that contain the # symbol:**
   - Maintain the complete semantic meaning while translating.
   - Do not alter the key object references.
   - Remove the # symbol in the final translation.
2. **For descriptions that do not contain the # symbol and have terms in square brackets []:**
   - When multiple instances of the same object type appear in brackets (like [monitor], [monitor]):
     - Treat them as a group rather than individually.
     - Use collective descriptions like "two monitors" or "multiple monitors" when appropriate.
     - Apply consistent styling to all instances of the same object type in a single description.
   - For each unique object type in brackets:
     - With 50% probability: Keep the original term or use a perfect synonym (e.g., "trash bin" might become "trash can" or "garbage bin").
     - With 50% probability: Replace the bracketed term with a functional description of the object (e.g., [trash bin] might become "an object used for disposing waste", [monitor] might become "a device used for displaying images").
   - Maintain the overall sentence structure and meaning.
   - **IMPORTANT:** Remove all square brackets [] in the final output while preserving the words inside them.
   - MUST ensure that half of bracketed terms are functional descriptions and the other half are left unchanged or replaced with synonyms.

Goal: Create natural-sounding English descriptions that preserve the original intent while introducing controlled variation to terms in brackets.
Examples:

- `"on the desks, [monitor], [monitor]"` might become `"Two monitors on the desk"` or `"Two display devices on the desk"`.
- `"on the table, there is [cup] [laptop]"` might become `"There is a cup and a portable computer on the table"`.

Here is the JSON data to process:
`{json_data}`
Instructions: Return the processed JSON with the same structure but with translated and modified descriptions according to the rules. Return **ONLY** the JSON data without any additional explanations.

Spatial Reasoning Prompt (ABS Data)

Please process the following JSON data by:

1. **For the word in square brackets [] in descriptions:**
   - With 50% probability: Keep the original term or use a perfect synonym (e.g., "trash bin" might become "trash can" or "garbage bin").
   - With 50% probability: Replace the bracketed term with a functional description of the object (e.g., [trash bin] might become "an object used for disposing waste").
   - Maintain the overall sentence structure and meaning.
2. **IMPORTANT:** Remove all square brackets [] in the final output.

Goal: Create natural-sounding English descriptions that preserve the original intent while introducing controlled variation to terms in brackets.
Here is the JSON data to process: `{json_data}`
Example:

```
{
    "object_id": [
        14,
        1
    ],
    "object_name": [
        "box",
        "monitor"
    ],
    "description": "A monitor that is 4.34 meter away from the box"
}
```

Instructions: Return the processed JSON with the same structure but with translated and modified descriptions according to these rules.

Common Sense Prompt

Generate diverse knowledge reasoning questions and answers based on a scene. You will receive a list of objects with their instance IDs and labels. Create questions covering multiple aspects of commonsense knowledge:

1. **Functionality** (e.g., *"What can be used to sit?"* or *"What can hold a drink?"*)
2. **Typical locations** (e.g., *"Which objects are typically found in a kitchen?"*)
3. **Safety concerns** (e.g., *"Which objects should children avoid for safety reasons?"*)
4. **Object relationships** (e.g., *"Which objects typically work together?"*)
5. **Temporal aspects** (e.g., *"Which objects are used daily?"* or *"Which objects are used seasonally?"*)

Ensure the questions are clear, diverse, and relevant to the scene. Avoid creating questions about *"wall"*, *"ceiling"*, *"floor"*, *"object"*, or *"remove"*.
Scene Data: `{Scene Data}`
Here are example question-answer pairs for reference. Please follow this style and difficulty level:
Example 1:

- **Question:** Objects that mostly used in the kitchen?
- **Answer:** `(61,mug),(1,coffee machine)`

Example 2:

- **Question:** Which objects are dangerous and should be kept away from children?
- **Answer:** `(62,heater),(85,heater)`

Instructions: Please generate 60 new questions and their answers. Use the following JSON format for your response:

```
[
    {
        "question": "Question 1",
        "answer": "(instance_id1,label_name1),..."
    },
    ...
]
```

Important: Prioritize questions about these rare objects: `{low frequency objects in this scene}`. Create at least 3 questions about them. Questions should have specific, informative answers.

Human Intention Prompt

Generate diverse knowledge reasoning questions and answers based on a scene. You will receive a list of objects with their instance IDs and labels. Create questions that cover multiple aspects of human intention. For example:

- *"I'm thirsty, what should I use?"*
- *"I'm sleepy, where should I go?"*

Ensure the questions are clear, diverse, and relevant to the scene. Avoid creating questions about *"wall"*, *"ceiling"*, *"floor"*, *"object"*, or *"remove"*.
Scene Data: `Scene Data`
Here are example question-answer pairs for reference. Please follow this style and difficulty level:
Example 1:

- **Question:** I need to put my books somewhere, where can I place them?
- **Answer:** `(18,storage cabinet),(26,storage cabinet)`

Example 2:

- **Question:** I want to cool myself while working, what can I use?
- **Answer:** `(84,pedestal fan)`

Instructions: Please generate 60 new questions and their answers. Use the following JSON format for your response:

```
[
    {
        "question": "Question 1",
        "answer": "(instance_id1,label_name1),..."
    },
    ...
]
```

Important: Prioritize questions about these rare objects: `{low frequency objects in this scene}`. Create at least 3 questions about them. Questions should have specific, informative answers.

Data Cleaning Prompt

I need you to identify and fix problems with questions about a 3D scene.
SCENE ID: {scene_id}
QUESTION TYPE: {question_type}

Here are all the questions with their answers:

```
 QUESTION 1:  {question}
ANSWER 1:  {answer}
OBJECTS IN ANSWER 1:  {objects}
...
```

For each question, identify if it has one of these problems:
1. **DIRECT MENTION:** The question explicitly mentions any specific object that appears in the answer. - Example: Where is the window located? when the answer includes "window" - Fix by completely rewording without using any terms from the answer (e.g., Where can I see outside from inside the house? instead of Where are the windows?)
2. **LOGICAL ERROR:** The question and answer have a fundamental mismatch in their logical relationship. - *For location questions:* If the question asks "where can I put/place/hang X?" but the answer only lists objects that are not locations, this is a logical error. - *For object questions:* If the question asks about using a specific function but the answer only includes objects that cannot fulfill that function, this is a logical error. - Fix by completely reframing the question to align with what the answer actually provides.
CRITICAL FORMATTING INSTRUCTIONS: - Provide ONLY the plain question text without any formatting markers. - NO quotation marks around the question. - NO prefixes like "Improved question:" or similar phrases. - NO explanations or meta-text in the question itself. - NO newlines or multiple sentences.
For "human_intention" questions especially: - If asking about "where to place X" and the answer only lists objects of type X (not locations), change to ask what can be used for the intended purpose. - If asking about finding a feature (like a window) and the answer lists those features, reword to ask about the function without naming the object.
Respond in this exact JSON format:

```
{
  "improvements": [
    {
      "question_number": 1,
      "original": "Original question text",
      "issue_type": "DIRECT MENTION or LOGICAL ERROR",
      "improved": "Clean improved question",
      "reason": "Brief explanation of the specific issue and fix"
    }
  ]
}
```

Only include questions with genuine issues that need improvement.

Second Correction Prompt (Direct Mention)

Your improved question still contains words from the answer objects.
Original question: `{original}`
Current improved version: `{improved}`
Objects to completely avoid mentioning: `{objects_in_answer}`

Please provide a completely different question that:

1. Achieves the same goal as the original question
2. Does **NOT** use **ANY** of the words from the answer objects list
3. Uses alternative descriptions or functions (e.g., *"Where can I see outside?"* instead of *"Where are the windows?"*)
4. Is concise and natural-sounding
5. Has **NO quotation marks, NO prefixes, NO explanations** - just the pure question

Just provide the clean question text with no additional formatting.

Second Correction Prompt (Logical Error)

Your improved question still doesn't fully address the logical mismatch with the answer.
Original question: `{original}`
Current improved version: `{improved}`
Answer objects: `{objects_in_answer}`

Especially for the human intention question, you need to completely reframe it to align with what the answer actually provides.
Please provide a completely different question that:

1. Makes logical sense given the objects in the answer
2. Doesn't ask about locations if the answer only provides objects
3. Asks about using, finding, or interacting with the type of objects in the answer
4. Is concise and natural-sounding
5. Has **NO quotation marks, NO prefixes, NO explanations** - just the pure question

Just provide the clean question text with no additional formatting.

Enhanced Prompt

URGENT: Your question still contains words from the answer objects.
Objects in answer: `{objects_in_answer}`
Current question: `{second_attempt}`

Provide a single question that **DOES NOT** contain **ANY** of these words or close synonyms.
For example, if the object is *"window"*, don't use *"window"*, *"windowsill"*, etc.
Write a question about the FUNCTION without naming the object:

- **For windows:** ask about *"seeing outside"* or *"natural light"*
- **For furniture:** ask about the function (*sitting, sleeping, storing*)
- **For appliances:** ask about the function (*cooking, cleaning, cooling*)

CRITICAL REQUIREMENTS:

1. ONE SENTENCE ONLY
2. NO QUOTATION MARKS
3. NO PREFIXES
4. NO EXPLANATIONS

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state our key contributions: the release of the SURPRISE-3D dataset, the definition of the 3D-SRS benchmark, and the evaluation of spatial reasoning capabilities in existing models. These claims are fully supported by the experiments and scope of our paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a brief discussion of limitations, noting that current large language and multimodal models lack spatial reasoning abilities, which necessitates human annotation. It also acknowledges the constraint of operating within static 3D scenes and the potential difficulty of generalizing to dynamic or outdoor environments.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper is focused on dataset construction, benchmark definition, and empirical evaluation. It does not present formal theoretical results or proofs.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides detailed descriptions of the dataset construction pipeline, benchmark task definition, evaluation protocol, and experimental settings for both zero-shot and fine-tuned models. We also specify implementation details for each baseline and report results using standardized metrics such as mIoU and accuracy at fixed thresholds. All dataset splits, annotation strategies, and benchmark interfaces are documented to ensure reproducibility.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We release the full SURPRISE-3D dataset, including all spatial and knowledge reasoning annotations, along with a public benchmark API and evaluation server. Code for reproducing the training and evaluation of all baselines, as well as detailed setup instructions and pretrained model checkpoints, are included in the supplemental material and will be hosted on a GitHub repository. This ensures that the community can faithfully reproduce all results.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper clearly specifies the training, validation, and test splits at the scene level to prevent overlap. We provide detailed descriptions of model settings, including hyperparameters, optimizer choices, and learning schedules, for both zero-shot and fine-tuning scenarios. Full implementation details and training procedures for each baseline are included in Appendix E, ensuring that readers can interpret and reproduce the reported results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: While the paper reports comprehensive evaluation metrics across multiple tasks and baselines, it does not include error bars, confidence intervals, or statistical significance tests.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper specifies that all experiments were conducted on 8 NVIDIA A100 GPUs (80GB), using 64-core AMD EPYC 7742 CPUs and 1TB RAM. Fine-tuning runs for each model took approximately 24–36 hours depending on model size and scene complexity. Inference for the test set averaged 5–8 minutes per scene. The appendix details compute usage across zero-shot and fine-tuned settings and includes notes on preliminary runs. This provides sufficient information for reproducibility.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

Justification: The research adheres to the NeurIPS Code of Ethics. All 3D scenes were sourced from publicly available datasets with proper licenses. Human annotations were collected with informed consent, and no personal or sensitive data is involved. The dataset and benchmarks are released under a permissive license for research purposes, ensuring transparency, fairness, and reproducibility.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: This work introduces a dataset and benchmark for 3D spatial reasoning in embodied AI, which is primarily foundational research. It is not tied to any specific application or deployment scenario, and as such, the societal impact is currently indirect.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The dataset introduced in this work does not pose significant risks of misuse. All scenes are sourced from publicly available indoor 3D datasets (e.g., ScanNet++), and annotations are focused on spatial reasoning with no personally identifiable or sensitive content. The work does not release generative models or scraped web data, and thus no specific safeguards are necessary.

Guidelines:

- The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All third-party assets used in this work are properly cited and comply with their respective licenses. The dataset is constructed based on ScanNet++ v2, which is released under a non-commercial Creative Commons license (CC BY-NC-SA 4.0). All baseline models (e.g., MLLMfor3D, 3D-Vista, Reason3D, Intent3D, ChatScene) are cited with appropriate references, and we adhere to their open-source licenses as specified in their official repositories.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We introduce SURPRISE-3D, a new large-scale dataset for spatial reasoning in 3D scenes. All assets are accompanied by comprehensive documentation, including annotation guidelines, data format specifications, licensing details, and usage instructions. We also include metadata for each scene, camera parameters for spatial grounding, and examples to support reproducibility.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.

- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No]

    Justification: While our dataset includes annotations collected from trained human annotators, the full instruction text, screenshots of the annotation interface, and compensation details are not included in the current submission due to space limitations. However, we ensured that all annotators were compensated fairly in accordance with local wage standards and followed clear guidelines. We will provide full annotation instructions and screenshots in the camera-ready version and upon dataset release.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [No]

    Justification: Our study only involved human annotators who contributed to data annotation and verification tasks, such as providing spatial descriptions and segmentation masks for 3D scenes. All procedures strictly adhered to ethical guidelines and were approved by an institutional ethics review process equivalent to IRB standards. Participants were fully informed about the nature and purpose of the tasks, provided their consent, and were fairly compensated for their contributions. No personal or identifiable information was collected during the study, ensuring complete privacy and confidentiality.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

Justification: Large Language Models (LLMs) were used as part of the annotation pipeline to generate candidate questions for knowledge reasoning (e.g., commonsense and intention-based queries). These LLM-generated samples were further filtered and validated by human annotators to ensure quality and correctness. This usage was central to scaling the dataset and is clearly documented in the annotation section of the paper.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.