Empowering Large Language Models with 3D Situation Awareness

Zhihao Yuan[1,2], Yibo Peng[1,2], Jinke Ren[1,2], Yinghong Liao[1,2], Yatong Han[1], Chun-Mei Feng[3], Hengshuang Zhao[4], Guanbin Li[5], Shuguang Cui[2,1], Zhen Li[2,1*]

[1] FNii-Shenzhen, CUHKSZ [2] SSE, CUHKSZ [3] IHPC, A*STAR, Singapore [4] HKU [5] SYSU

Abstract

Driven by the great success of Large Language Models (LLMs) in the 2D image domain, their application in 3D scene understanding has emerged as a new trend. A key difference between 3D and 2D is that the situation of an egocentric observer in 3D scenes can change, resulting in different descriptions (e.g., "left" or "right"). However, current LLM-based methods overlook the egocentric perspective and use datasets from a global viewpoint. To address this issue, we propose a novel approach to automatically generate a situation-aware dataset by leveraging the scanning trajectory during data collection and utilizing Vision-Language Models (VLMs) to produce high-quality captions and question-answer pairs. Furthermore, we introduce a situation grounding module to explicitly predict the position and orientation of the observer's viewpoint, thereby enabling LLMs to ground situation descriptions in 3D scenes. We evaluate our approach on several benchmarks, demonstrating that our method effectively enhances the 3D situational awareness of LLMs while significantly expanding existing datasets and reducing manual effort.

1. Introduction

Recently, Large Language Models (LLMs) [28, 34, 35] have revolutionized natural language processing, showcasing remarkable capabilities in image understanding tasks [22, 25, 29], such as image captioning, visual question answering (VQA), and engaging in dialogs about visual content. Building on this success, researchers have begun to explore the use of LLMs in the three-dimensional (3D) domain [8, 15, 39], aiming to bridge the gap between language and 3D visual data. In 3D vision and language tasks—including 3D visual grounding, 3D captioning, and 3D VQA—the objective is to unify these tasks under the framework of next-token prediction. To achieve this, researchers focus on learning a 3D representation that aligns with pre-trained text embedding spaces, enabling seamless integration with language models.

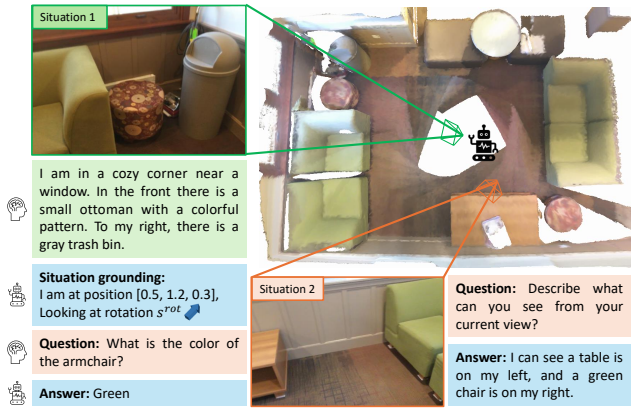


Figure 1. Illustration of 3D Situation Awareness. The LLMs can accurately ground situation descriptions to the observer's position and orientation, enabling context-aware question answering based on the observer's viewpoint.

However, a significant challenge in applying LLMs to 3D tasks is the scarcity of annotated 3D-text data, which is crucial for training such models effectively. Unlike 2D images, which have abundant paired data in the form of captions and annotations, 3D data lacks extensive textual descriptions. To overcome this limitation, researchers have explored methods to generate additional 3D-text data to augment the training process. For example, 3D-LLM [15] utilizes multi-view images of 3D scans to generate captions, leveraging rich visual information from different perspectives. Similarly, LEO [18] and SceneVerse [20] employ scene graphs and harness the capabilities of ChatGPT to generate captions and question-answer (QA) pairs, enriching the textual annotations associated with 3D content.

Despite these efforts, inherent limitations exist in the current data generation processes. Firstly, most approaches to 3D scene understanding emphasize global perspectives, providing comprehensive overviews of environments while overlooking the importance of situational con-

*Corresponding author.

texts—specific viewpoints or scenarios within a scene crucial for accurate interpretation. Unlike static images with fixed viewpoints, 3D scenes are dynamic, and the perceived situation can change when an embodied agent moves through the environment. As illustrated in Fig. 1, an object like a sofa may appear on the left side from one viewpoint (Situation 1) but on the right side from another viewpoint (Situation 2). Without specifying the situational context, such variations can lead to ambiguity during model training and degrade performance in tasks that require spatial understanding.

Moreover, current scene graph-based methods [15, 18, 20, 40] for data generation rely heavily on ground-truth 3D instance labels to construct accurate scene graphs. Acquiring these 3D labels is labor-intensive and costly and limits the scalability of the data generation process. Existing labeled datasets are often insufficient, lacking coverage of all objects within a scene, especially small objects and those belonging to rare categories. Additionally, the relationships between objects are typically predefined using fixed templates, which restricts the ability of models to handle open vocabulary scenarios and capture the richness of the real-world.

To overcome these challenges, we propose a novel approach to automatically generate a situational dataset, termed **View2Cap**. Our key insight is that 3D scans are commonly reconstructed from RGB-D videos, where the camera trajectory inherently represents an egocentric exploration of the environment by a human observer. By leveraging this naturalistic data, we can capture the situational context not included in existing datasets. Specifically, we utilize 2D Vision-Language Models (VLMs) to generate captions and QA pairs from individual frames of the RGB-D videos. This approach effectively distills knowledge from well-established 2D models into the 3D domain, capitalizing on the strengths of 2D VLMs. Concurrently, we record the camera pose of each frame, which, in combination with the depth information, allows us to extract the corresponding point cloud for that particular region. This methodology enables us to create a point cloud-text dataset with situational context, capturing the dynamic perspectives encountered by an embodied agent moving through a 3D environment. Importantly, this strategy reduces data generation costs and supports the scalable creation of datasets capable of handling open vocabulary scenarios without the need for extensive manual annotations or 3D labels.

In addition to the dataset creation, enhancing the situational understanding of LLMs requires models that can explicitly ground descriptions in the 3D space. To this end, we propose a **Situation Grounding (SG)** module that builds upon existing 3D LLM architectures. This module allows the model to predict situational positions and view rotations based on textual descriptions and the scene-level point cloud. By treating each object within the scene as an anchor point, the model can predict the distance and angle class offsets relative to the observer's viewpoint. This formulation transforms the complex problem of pose estimation into a more tractable classification task, simplifying the learning process and improving the model's ability to comprehend and reason about spatial relationships in 3D environments.

Our contributions can be summarized as follows:

- We introduce **View2Cap**, a scalable 3D dataset that provides paired data of situational positions, rotations, region point clouds, textual descriptions, and QA pairs. This dataset is generated automatically without the need for 3D labels or extensive manual annotations, enabling the study of situational context in 3D scene understanding.
- We propose the Situation Grounding **SG** module, that can be integrated into existing LLMs, allowing for explicit grounding of situational descriptions in 3D scenes. This module transforms pose estimation into a classification problem, facilitating easier training and improved spatial reasoning.
- Through extensive experiments, we demonstrate that combining our situational dataset and grounding module significantly enhances the situational awareness and performance of existing 3D LLMs in various tasks.

2. Related Work

Indoor 3D Scene Understanding. 3D scene understanding involves perceiving and interacting with 3D environments, encompassing tasks such as 3D segmentation [33], 3D visual grounding [5, 41, 42], 3D captioning [10, 43], and 3D visual question answering (VQA) [26, 40]. The development of large-scale RGB-D scan datasets has significantly advanced this field. Notably, ScanNet [11] provides extensive annotations for indoor scenes, facilitating research in 3D scene understanding. Matterport3D [4] offers a large-scale collection of richly annotated house-level environments, providing high-resolution RGB-D scans and detailed semantic labels, which have been instrumental in advancing tasks such as 3D reconstruction, navigation, and semantic understanding. EmbodiedScan [38] enriches these annotations by providing more fine-grained object bounding boxes, particularly focusing on small objects and diverse class labels, with assistance from models like SAM [21].

Large Language Models in 3D. Inspired by the success of LLMs in image understanding, researchers have begun exploring the integration of 3D inputs with LLMs to leverage their impressive reasoning and generalization capabilities for 3D understanding [44]. Models such as PointLLM [39] and GPT4Point [31] attempt to map point clouds into the token space of LLMs to generate captions for objects. However, they struggle to handle scene-level point clouds due to the complexity of indoor environments. 3D-LLM [15] pioneers the use of LLMs for scene understanding but still

relies on 2D features. Recent models like LL3DA [8], Chat-3D [17], and LEO [18] have investigated the use of 3D encoders for scene-level tasks. To enhance grounding abilities, Grounded 3D-LLM [9] introduces referent tokens and employs contrastive learning to unify grounding with textual responses. Similarly, Chat-3D [17] proposes the use of object identifiers (object IDs) to facilitate referring expressions and grounding mechanisms. However, these models still lack a comprehensive understanding of situation awareness in 3D space.

Situation Awareness in 3D Space. A key difference between 2D and 3D scene understanding lies in situation awareness. In 2D images, the viewpoint is fixed, making spatial relationships like left and right straightforward to determine. In contrast, in 3D spaces, these relationships can change with the observer's position. For example, left/right relationships can reverse when moving from one side of a room to another. Some works have addressed this problem using data augmentation techniques, such as MVT [19] and ViewRefer [14]. SQA3D [26] first proposes to describe the situation in text and then conduct tasks like VQA. However, it relies on human annotators to write these descriptions, making it costly and challenging to scale for the large-scale training required by 3D LLMs. Our method addresses this limitation by using an automatic situation dataset generation pipeline that leverages the capabilities of 2D VLMs and the trajectories inherent in RGB-D dataset collection processes.

3. Method

To enable LLMs to comprehend situational contexts within 3D spaces, we propose a method that involves an automatic data generation pipeline for building a situation dataset and a novel module for situation grounding.

3.1. Situation Dataset

Automatic Data Generation. Our data generation process leverages the natural exploratory behavior inherent in RGB-D videos, which serves as first-person navigations through 3D environments. For each video frame, we extract the situational context directly from the camera extrinsic, capturing precise positional and rotational information. Moreover, we employ VLMs to generate captions from the corresponding 2D images extracted from the video frames. It provides elaborate information about the environment beyond 3D labels. Additionally, we utilize VLMs to generate QA pairs related to each situation, providing more specific and direct supervision compared to captions alone. Specifically, we use Llava-onevision [23] as VLM for its comprehensive caption ability and open-sourced.

For situation descriptions, we generate two types of captions: simple and detailed. The simple captions focus on the primary objects and their relationships within the scene, ensuring that the model grasps the essential components of the environment. The detailed captions provide an elaborate account of all visual information presented in the image, including background elements like the floor and overall environment, as shown in Fig. 4.

Statistic	SQA3D	View2Cap
Total s^{txt}	20,369	231,184
Total q	33,403	553,779
Unique q	26,091	92,877
Total scenes	650	2,841
Total objects	14,925	71,376
Average s^{txt} length	17.49	54.73
Average q length	10.49	8.55
Average a length	1.10	5.74

Table 1. Dataset Statistics for SQA3D and View2Cap.

For the situation QA pairs, we define four categories of questions: (1) object identification, prompting recognition of objects presented in the image; (2) spatial relationships, describing the positions of objects relative to each other from the viewer's perspective (e.g., left, right, front, behind); (3) visual features, showing distinctive attributes such as colors, shapes, sizes, and textures; (4) insights into the overall layout of the room, enhancing the model's understanding of the scene as a whole.

Dataset Verification and Refinement. To ensure the quality and reliability of the generated dataset, we implement a verification and refinement process leveraging GPT-4. Specifically, we utilize the 3D labels of point clouds and employ GPT-4 to evaluate whether the generated captions include all objects present in the scene. We use annotations from EmbodiedScan [38], as it annotate more fine-grained categories and small objects compared with the labels from the original scan datasets. The scoring criteria are adopted from PointLLM [39], including correctness, hallucination, and general considerations, range from 0-5. The average score for View2Cap is 3.09, indicating that captions derived from 2D images can effectively serve as high-quality supervision for training 3D models. We also ask GPT-4 to refine the validation set based on the labels, which increases the average score to 3.31. This refinement step enhances the reliability of our evaluations when assessing 3D models on validation sets. Detailed information about our scoring process is provided in the Supplementary Material.

For the ViewQA dataset, we aim to enrich the diversity and informativeness of the QA pairs. For each image, the VLM generates 10 questions for each predefined type. We then utilize GPT4 to rank these QA pairs based on criteria such as relevance, clarity, and informativeness. Items that fall below a certain threshold are excluded from the dataset. This ranking and filtering process ensures that only

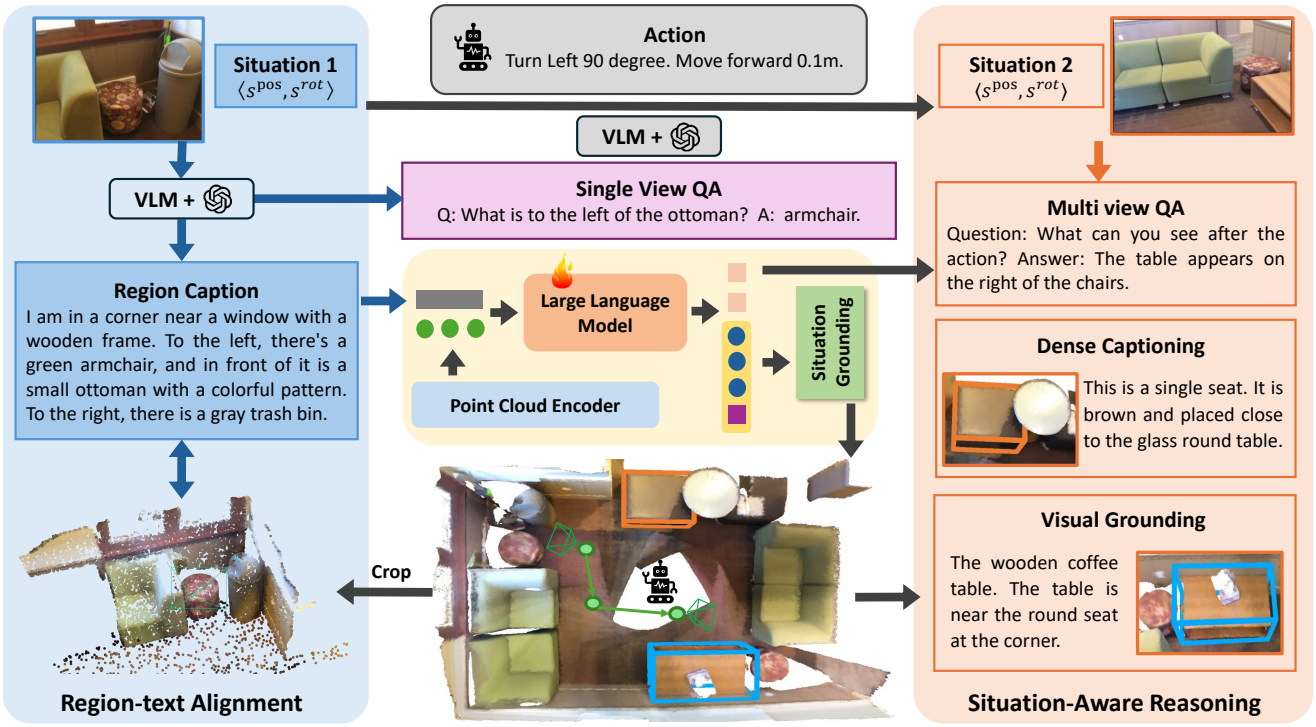


Figure 2. **Overview of our method.** The left part illustrates the process of region-text alignment. Paired point cloud and caption data are generated using VLM and RGB-D videos. The LLM is fine-tuned to align features from the point cloud encoder and generated region caption. For situation grounding, the region caption is fed into the LLM to predict the viewpoint of the observer $\mathbf{s}^{pos}$ and $\mathbf{s}^{rot}$. The right part shows the situation-aware instruction tuning process, where QA data is generated using multi-view images and corresponding actions.

the most pertinent and well-constructed QA pairs are retained, thereby enhancing the overall quality of the dataset.

In total, we collect more than 200K situation descriptions and 550K QA pairs using 2,841 scans from ScanNet [10], 3RScan [37], and Matterport3D [4] datasets, which is 10x large than the SQA [26]. Our average situation text length (54.73) is much longer than SQA3D (17.49). Our average answer length of 5.74 is also longer. More detailed statistics are provided in the Supplementary Material.

3.2. Model Architecture

Our model takes as input a point cloud $\mathbf{P} \in \mathbb{R}^{N\times 6}$, representing a 3D scan where each point includes spatial coordinates and RGB color information. Additionally, it processes a situation description s and a task instruction t. The goal is to generate the task answer a. In addition, we ask the model to predict the situation position $\mathbf{s}^{\text{pos}} = (x, y, z) \in \mathbb{R}^3$ and the rotation of the front view direction represented as a quaternion $\mathbf{s}^{\text{rot}} = (q_x, q_y, q_z, w)$, where q_x, q_y, and q_z are the vector components representing the axis of rotation scaled by w.

Point Cloud Encoder. The point cloud $\mathbf{P}$ is segmented into K instances $\{\mathbf{P}_k\}$, each representing an object in the scene. For each instance, we group the points into local patches $\{\mathbf{p}_{k,l}\}_{l=1}^{L_k}$, where L_k is the number of patches in instance k. Each patch contains neighboring points grouped based on spatial proximity. The patch sequence along with a special token [CLS] is processed by a Vision Transformer (ViT) [13] to obtain instance feature $\mathbf{v}_k$. The point cloud encoder [45] is fine-tuned on a large-scale 3D object dataset [12] in the classification task.

Connector. To enhance the spatial relationships between objects [17, 18], we employ spatial attention layers [6] to fuse the coordinates of objects with their semantic features. This fusion effectively integrates spatial and semantic information, which is crucial for comprehensive scene understanding. Subsequently, to map the instance-level features into the embedding space of the LLM, we utilize a simple multilayer perceptron (MLP) f_{proj} to obtain the processed visual tokens corresponding to the object instances in the scene:

$$\tilde{\mathbf{v}}_k = f_{\text{proj}}(\mathbf{v}_k). \tag{1}$$

This projection ensures that the visual features are compatible with the LLM's embedding space.

Large Language Model. Following multimodal LLM approaches, we tokenize the visual features and interleave them with text tokens to serve as input to the LLM. Specifically, the input sequence begins with system messages and

the situation description t_s, followed by the visual tokens, and concludes with the task instruction t:

$$\text{Input} = [t_s, \tilde{\mathbf{v}}_1, \tilde{\mathbf{v}}_2, \ldots, \tilde{\mathbf{v}}_K, t]. \tag{2}$$

This structure allows the LLM to process the visual context within the flow of textual information. During the forward pass, the LLM processes the entire input sequence and generates hidden states for each token. We introduce a special grounding token [GRD] in the output sequence. Let $\{\mathbf{h}_k\}$ represent the hidden state corresponding to the visual tokens $\{\tilde{\mathbf{v}}_k\}$ and $\mathbf{h}_{\text{GRD}}$ as the hidden state corresponding to this grounding token.

Situation Grounding Module. Directly predicting the situation's absolute position $\mathbf{s}^{pos}$ and rotation $\mathbf{s}^{pos}$ in 3D space is challenging due to the complexity of estimating precise spatial coordinates and orientations. To simplify this task, we propose using anchor points derived from objects within the scene. Specifically, we treat each object as an anchor, utilizing its center coordinates $\mathbf{a}_k^{\text{pos}}$ and rotation $\mathbf{a}_k^{\text{rot}}$ as reference points. Consequently, we only need to predict the offset $\Delta\mathbf{p}_k \in \mathbb{R}^3$ from the anchor position to the situation position and the angular difference θ_k between the anchor rotation and the situation rotation, as illustrated in Fig. 3. The blue circles and arrows shows the anchor $\mathbf{a}_k^{\text{pos}}$ and $\mathbf{a}_k^{\text{rot}}$. The greens are ground truth. The dotted line shows the offset from $\mathbf{a}_k^{\text{pos}}$ to $\mathbf{s}_k^{\text{pos}}$. The solid arrow shows the predicted rotation is rotated by θ from $\mathbf{a}_k^{\text{rot}}$. We set each $\mathbf{a}_k^{\text{rot}}$ point to the center of the room. Beacuse estimating the front face of each object is a difficult problem.

For rotation prediction, we convert the regression problem into a classification problem. Considering rotations around the vertical axis (*i.e.*, yaw rotation) as an example (see Fig. 3), we discretize the rotation angle into B bins ranging from $-\pi$ to π. The target angular difference θ_k can then be represented by the bin index $\hat{b}_k$ for anchor k:

$$\hat{\theta}_k = -\pi + \frac{2\pi}{B}\left(\hat{b}_k + \frac{1}{2}\right). \tag{3}$$

We employ another MLP f_{grd} to predict the confidence score $c_k \in [0, 1]$, the position offset $\Delta\mathbf{p}_k$, and the rotation bin $\hat{b}_k$. Here, $\mathbf{h}_{\text{GRD}}$ is the hidden state of the grounding token from the LLM's output, and $\mathbf{h}_k$ is the hidden state corresponding to the visual token $\tilde{\mathbf{v}}_k$ of anchor k:

$$(c_k,\ \Delta\mathbf{p}_k,\ \hat{b}_k) = f_{\text{grd}}\left([\mathbf{h}_{\text{GRD}};\ \mathbf{h}_k]\right), \tag{4}$$

where $[\mathbf{h}_{\text{GRD}};\ \mathbf{h}_k]$ denotes the concatenation of the two hidden states. The predicted situation position for anchor k is then computed as:

$$\hat{\mathbf{s}}_k^{\text{pos}} = \mathbf{a}_k^{\text{pos}} + \Delta\mathbf{p}_k. \tag{5}$$

During inference, we select the most confident anchor $k^* = \arg\max_k c_k$, and use its predictions for the situation

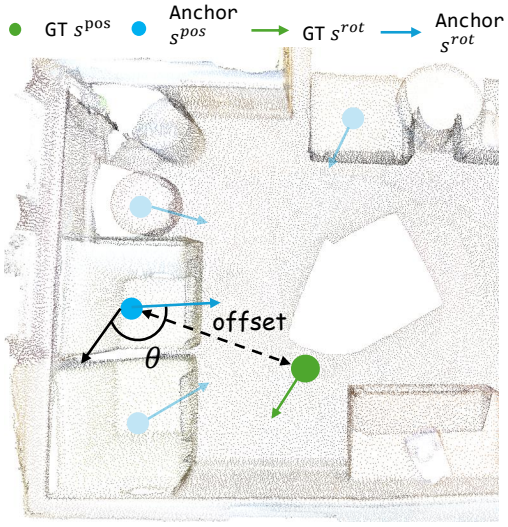


Figure 3. **Situation prediciton.** We treat each object as anchor (blue), where the center is $\mathbf{a}_k^{\text{pos}}$. We set each $\mathbf{a}_k^{\text{rot}}$ point to the center of the room. Ground truth position and rotation are shown in green. The dotted line shows the offset from $\mathbf{a}_k^{\text{pos}}$ to $\mathbf{s}_k^{\text{pos}}$. The solid arrow shown the predicted rotation is rotated by θ from $\mathbf{a}_k^{\text{rot}}$.

position and rotation, *i.e.*, $\hat{\mathbf{s}}^{\text{pos}} = \hat{\mathbf{s}}_{k^*}^{\text{pos}}$, $\hat{\theta} = \hat{\theta}_{k^*}$. Then, the predicted angular difference $\hat{\theta}$ is mapped back to a quaternion representation for the situation rotation. Assuming the anchor's rotation $\mathbf{a}_{k^*}^{\text{rot}}$ is known, the situation rotation $\hat{\mathbf{s}}^{\text{rot}}$ is calculated by applying the rotation difference to the anchor's rotation:

$$\delta\hat{\mathbf{s}}^{\text{rot}} = \left(0,\ 0,\ \sin\left(\frac{\hat{\theta}}{2}\right),\ \cos\left(\frac{\hat{\theta}}{2}\right)\right), \tag{6}$$

$$\hat{\mathbf{s}}^{\text{rot}} = \mathbf{a}_{k^*}^{\text{rot}} \otimes \delta\hat{\mathbf{s}}^{\text{rot}}, \tag{7}$$

where $\otimes$ denotes quaternion multiplication. By transforming the prediction task into estimating relative offsets and rotations with respect to anchor points, we reduce the complexity associated with directly predicting absolute positions and orientations in 3D space.

3.3. Training

Our training process comprises three stages. In the first stage, we train the connector to align the point cloud features with the text embedding space. In the second stage, we train the situation grounding module using the situation grounding task, enhancing the LLM's situational awareness within 3D environments. In the final stage, we fine-tune the entire model using downstream instruction data to further improve its performance. Throughout all stages, we fine-tune the LLM using LoRA [16], and we utilize LLaMa 3.1 [35] as the base LLM.

Region Text Alignment. To achieve point cloud-text alignment, we map the point cloud to images, ensuring that we retain only the objects that are visible within the images. We also incorporate depth information to filter out object point clouds that are completely occluded in the corresponding view. This results in region-text pairs, linking specific regions of the point cloud to their corresponding textual descriptions. Compared to prior methods [8, 18] that rely on scene captions, our approach simplifies the training process by reducing the number of objects, eliminating ambiguity in spatial relationships due to explicit viewpoint information, and increasing the number of training samples. A single scene can be divided into multiple regions, providing diverse contexts for training.

Situation Grounding. After establishing the point cloud-text alignment, we proceed with the situation grounding. During training, we supervise only the anchors whose positions are within a distance threshold D of the ground truth situation position $\mathbf{s}^{\text{pos}}$. Formally, the set of supervised anchors is given by:

$$\mathcal{K} = \left\{ k \;\middle|\; \left\| \mathbf{a}_k^{\text{pos}} - \mathbf{s}^{\text{pos}} \right\|_2 \leq D \right\}.$$

For these anchors, the position loss is defined as:

$$\mathcal{L}_{\text{pos}} = \sum_{k \in \mathcal{K}} \left\| \hat{\mathbf{s}}_k^{\text{pos}} - \mathbf{s}^{\text{pos}} \right\|_2^2, \tag{8}$$

where $\hat{\mathbf{s}}_k^{\text{pos}}$ is the predicted situation position for anchor k. The rotation loss is given by:

$$\mathcal{L}_{\text{rot}} = -\sum_{k \in \mathcal{K}} \sum_{b=1}^{B} y_b \log p_{k,b}^{\text{rot}}, \tag{9}$$

where $p_{k,b}^{\text{rot}}$ is the predicted probability for rotation bin b for anchor k, and y_b is the one-hot encoding of the ground truth rotation bin corresponding to the angular difference between $\mathbf{a}_k^{\text{rot}}$ and $\mathbf{s}^{\text{rot}}$. The confidence loss encourages the confidence scores c_k to reflect the anchors' proximity to the ground truth position:

$$\mathcal{L}_{\text{conf}} = \sum_{k} \left| c_k - \exp\left(-\alpha \left\| \mathbf{a}_k^{\text{pos}} - \mathbf{s}^{\text{pos}} \right\|_2 \right) \right|, \tag{10}$$

with α being a scaling factor controlling the decay rate. During inference, we select the most confident anchor $k^* = \arg\max_k c_k$ and use its predictions for the situation position and rotation.

Instruction Tuning. After training with situation awareness, we finetune the LLM on downstream 3D reasoning tasks. The task answer loss is defined as the standard cross-entropy loss for language modeling:

$$\mathcal{L}_{\text{ans}} = -\sum_{i=1}^{T} \log P(a_i \mid a_{<i}, \text{ Input}), \tag{11}$$

where T is the length of the task answer a, and $P(a_i \mid a_{<i}, \text{ Input})$ is the probability of generating token a_i.

4. Experiments

4.1. 3D Scene Understanding

Overview. We evaluate our method on three well-established 3D scene understanding tasks. Specifically, Scan2Cap [10] requires the model to generate captions of each object in the scene regarding their category, attributes, and neighbor content. ScanQA [1] requires the model to answer questions related to objects in 3D. SQA3D [26] requires the model to answer questions under particular situations described by the text. We investigate how well our method can perform 3D VL understanding and reasoning tasks, especially when compared against task-specific models [3, 27] and existing generalist models [15, 18]. We evaluate our model using conventional text generation metrics, including CIDEr [36], BLEU [30], METEOR [2], and ROUGE-L [24], and open-ended generation metric Sentence-Sim [32] and refined exact-match accuracy [18]. Following 3D-VisTA [46], we utilize object proposals from Mask3D [33] instead of ground-truth object segments for evaluation.

Results and Analysis. Existing methods for 3D scene understanding can be categorized into two streams: specialist models and generalist models. Specialist models are designed specifically for individual tasks only. Generalist models *i.e.* LEO [18] allow for joint training and inference across different datasets without changing the network structure. Compared with LEO, our method surpasses 2.8 CIDEr scores on Scan2Cap and 4% on EM@1 on SQA3D. This improvement underscores the effectiveness of integrating 3D situational awareness into LLMs for enhancing 3D scene understanding and reasoning capabilities.

4.2. Situation Grounding

Overview. In this part, we evaluate our method's ability to predict the position and orientation of agents based on textual descriptions using the SQA3D dataset. This dataset provides 26,000 situational descriptions, making it a comprehensive benchmark for 3D scene understanding from an egocentric perspective. To assess our model's performance on situation grounding, we use the four metrics: Acc@0.5m and Acc@1.0m, which are the percentages of position predictions within 0.5 meters and 1.0 meters of the ground truth on the x-y plane; Acc@15° and Acc@30°, which are the percentages of rotation predictions within 15 degrees and 30 degrees of the ground truth around the z-axis (yaw rotation). The experiments in Table 3 demonstrate that our model effectively grounds textual descriptions into accurate spatial positions and orientations within 3D scenes. This highlights

Model	Scan2Cap (val)					ScanQA (val)					SQA3D (test)
	C	B-4	M	R	Sim	C	B-4	M	R	EM@1	EM@1
Task-specific models											
Scan2Cap [10]	35.2	22.4	21.4	43.5	-	-	-	-	-	-	41.0
3DJCG [3]	47.7	31.5	24.3	51.8	-	-	-	-	-	-	-
Vote2Cap-DETR [7]	61.8	34.5	26.2	54.4	-	-	-	-	-	-	-
ScanRefer+MCAN	-	-	-	-	-	55.4	7.9	11.5	30.0	18.6	-
ClipBERT	-	-	-	-	-	-	-	-	-	-	43.3
ScanQA [1]	-	-	-	-	-	64.9	10.1	13.1	33.3	21.1	47.2
SIG3D[27]	-	-	-	-	-	68.8	12.4	13.4	35.9	-	52.6
Generalist models											
3D-VisTA [46]	66.9	34.0	27.1	54.3	53.8	69.6	10.4	13.9	35.7	22.4	48.5
3D-LLM (FlanT5)	69.4	12.0	14.5	35.7	-	-	-	-	-	-	-
LL3DA [8]	65.2	36.8	26.0	55.1	-	76.8	13.5	15.9	37.3	-	-
LEO [18]	72.4	38.2	27.9	58.1	55.3	101.4	13.2	20.4	49.2	24.5 (47.6)	50.0 (52.4)
Ours	**75.2**	**38.9**	**29.0**	**58.7**	**56.3**	**89.8**	**14.6**	**17.5**	**42.9**	**22.9 (40.2)**	**54.0 (56.0)**

Table 2. Quantitative comparison with state-of-the-art models on 3D VL understanding tasks. "C" stands for "CIDEr", "B-4" for "BLEU-4", "M" for "METEOR", "R" for "ROUGE", "Sim" for sentence similarity, and "EM@1" for top-1 exact match. The n-gram metrics for Scan2Cap are governed by IoU@0.5. Entries in gray indicate using ground truth question-relative object annotations.

Model	Localization		Orientation	
	Acc@0.5m	Acc@1.0m	Acc@15°	Acc@30°
Random	7.2	25.8	8.4	16.9
SQA3D [26]	9.5	29.6	8.7	16.5
SQA3D *(separate)*	10.3	31.4	17.1	22.8
3D-VisTA [46]	11.7	34.5	16.9	24.2
SIG3D* [27]	16.8	35.2	23.4	26.3
Ours	**17.4**	**36.9**	**24.1**	**28.5**

Table 3. Performance comparison of different models on localization and orientation metrics. * indicates our reproduction using their open-sourced code.

Type	Data	C	B-4	M	R	Sim
Simple	SceneCap	5.7	2.5	12.9	18.5	49.7
	SceneVerse	5.7	2.5	12.9	18.5	49.7
	View2Cap (S)	31.3	14.1	17.8	38.1	64.0
	View2Cap	**36.0**	**15.0**	**18.5**	**38.5**	**65.4**
Detail	SceneCap	1.5	0.5	8.6	15.8	50.9
	SceneVerse	4.3	1.5	8.6	15.8	50.9
	View2Cap (D)	11.2	11.8	19.1	29.8	69.0
	View2Cap	**12.5**	**12.2**	**19.9**	**29.9**	**70.0**

Table 4. Result on situation captioning training with different data.

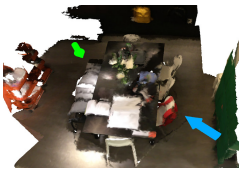
its potential for real-world applications where precise localization and orientation based on language inputs are crucial.

4.3. Situation Captioning

Describing the 3D environment from a first-person perspective is a critical task in embodied applications, such as navigation. In this experiment, we evaluate 3D LLMs on the task of generating region-level captions conditioned on a given situation. We start by using the agent's positional and rotational data, along with the camera position, to filter the point cloud to only the region visible to the agent. This filtered point cloud is then fed into the model to generate captions. The evaluation is conducted on 7,074 samples from the ScanNet validation set. We consider two types of captions: simple captions, which focus only on the main objects, and detailed captions, which capture the full environment in greater detail. The results, presented in Table 4, show that LEO trained solely on object and scene-level captioning data perform poorly on situation captioning tasks. In contrast, our View2Cap data significantly improves performance. Additionally, training with both simple and detailed captioning data yields better results compared to using only one type of captioning data, suggesting that a diverse range of captioning styles can enhance model training and performance.

4.4. Situation Question Answering

We evaluate our model's performance on the situational VQA task using our generated ViewQA dataset. For the single-view QA, we input the region point cloud into the model, similar to the procedure in the situational captioning task. For the multi-view QA, we provide the model with a scene-level point cloud along with a situational description that includes the starting point and actions as conditions. The model is then tasked with answering questions about the differences or new views resulting from this action. The validation set comprises 8,287 questions derived from the ScanNet dataset. As shown in Table 6, incorporating the Situation Grounding (SG) Module for situational prediction

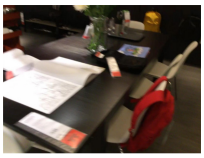


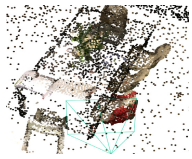
SceneVerse

In this Dining Room scene, there are 5 chairs, 2 carts, 1 cabinet, and 3 backpacks. The cabinet is positioned to the right of the cart, while the cart is to the left of the cabinet. The chair is also to the left of the cabinet, and to the right of the cart. The objects in this room suggest a functional and organized space, suitable for dining or social gatherings. The chairs provide seating, the cabinet offers storage, and the carts can be used for serving or transportation.

Simple: I am in a modern living room. In front of me, there is a long black table with a clear vase of green plants and a stack of book on it.

Detail: I am standing in a well-lit room with a modern interior. In front of me is a sleek, black rectangular table with a glass vase containing vibrant green plants on the left side. To the right of the table, there are two gray pillows with black stripes neatly placed against the wall. On the table, an open book lies flat, and in front of it, a clear glass with a green rim and a white tag stands next to another glass. To the left, there's a glass container with a green lid and a white price tag attached to it. The overall setting appears to be a modern living room or a showroom with a minimalist aesthetic.





Simple: I am in a living room. There is a black table with papers and three chairs next it.

Detail: I am in a room with a desk in the foreground. On the desk, there's a black surface with various items scattered on it, including papers, a red backpack hanging on a chair, and a plastic bag. To the left of the desk is a wooden shelf with some objects on it. In the background, there's a television set and a yellow bag on a stand. The room has a dark ambiance with some bright spots, possibly from a light source outside the frame.

Figure 4. Examples of our View2Cap situation captions against SceneVerse. We mark facts in green and spatial relations in blue .

	Localization		Orientation	
	Acc@0.5m	Acc@1.0m	Acc@15°	Acc@30°
LEO w. SG	8.3	30.4	10.9	19.5
+ Anchor	13.7	32.2	16.9	21.8
+ Discrect bins	13.6	32.3	21.6	25.0
+ View2Cap	**17.4**	**36.9**	**24.1**	**28.5**

Table 5. Ablations of our designs on situation grounding.

and utilizing the View2Cap data for region-text alignment significantly improves performance on our ViewQA task.

4.5. Qualitative Analysis

Fig. 4 presents the qualitative example of our View2Cap dataset complete against scene graph based generation method SceneVerse [20]. The first column shows the SceneVerse caption, it just summarizes the objects in the scene and describes the relation of objects in predefined rules. While our method gives more detailed object descriptions and accurate spatial relationships. For instance, the caption from SceneVerse ignores the glass vase and opened books on the table as our View2Cap describes.

4.6. Ablation Study

Design of Situation Grounding Module. We test different designs of situation grounding modules as shown in Table 5. Compared with the model without anchors, the localization Acc@1.0m improves 5.4%, indicating that the anchor mechanism can help the model narrow down the situation in a smaller range. Additionally, using the discrete bins to predict the rotation provides more accurate angles. Pretraining using View2Cap on the captioning task can improve both the position and rotation performance.

	ViewQA		SQA3D		ScanRefer	
	EM	EM-R	EM	EM-R	Acc@0.25	Acc@0.5
LEO	39.3	44.1	62.8	52.4	36.1	30.8
+ SG module	40.2	45.3	50.8	53.2	38.3	32.9
+ View2Cap	**42.0**	**46.6**	**54.0**	**56.0**	**42.8**	**38.4**

Table 6. Ablations of our designs on situation-aware reasoning.

Effectiveness of Situation Data. To evaluate the effectiveness of situation data, we test the performance of the model on SQA3D, as it requires both the situation grounding and QA ability. We also test the influence of situation data on 3D visual grounding, as it also needs to infer the situation to distinguish distractors. The result in Table 6 shows that pretraining on our proposed View2Cap and situation grounding can effectively improve the performance on those tasks that need situation awareness.

5. Conclusion

In this paper, we presented a novel approach to enhance 3D LLMs with situational awareness. Recognizing the limitations of existing methods that overlook the egocentric perspective inherent in 3D environments, we proposed the automatic generation of a situation-aware dataset called View2Cap. By leveraging the scanning trajectories from RGB-D video data and utilizing powerful VLMs, we produced high-quality captions and QA pairs that capture the dynamic viewpoints of an observer moving through a 3D scene. Furthermore, we introduced a situation grounding module, enabling LLMs to ground textual descriptions to situations in 3D space explicitly. We hope our work will advance the first-person 3D understanding of embodied tasks.

6. Acknowledgements

This work was supported by NSFC with Grant No. 62293482, by the Basic Research Project No. HZQB-KCZYZ-2021067 of Hetao Shenzhen HK S&T Cooperation Zone, by Shenzhen General Program No. JCYJ20220530143600001, by Shenzhen-Hong Kong Joint Funding No. SGDX20211123112401002, by the Shenzhen Outstanding Talents Training Fund 202002, by Guangdong Research Project No. 2017ZT07X152 and No. 2019CX01X104, by the Guangdong Provincial Key Laboratory of Future Networks of Intelligence (Grant No. 2022B1212010001), by the Guangdong Provincial Key Laboratory of Big Data Computing, CHUK-Shenzhen, by the NSFC 61931024&12326610, by the Key Area R&D Program of Guangdong Province with grant No. 2018B030338001, by the Shenzhen Key Laboratory of Big Data and Artificial Intelligence (Grant No. ZDSYS201707251409055), by Shaanxi Mathematical Basic Science Research Project (No.23JSY047), and by Tencent & Huawei Open Fund, by China Association for Science and Technology Youth Care Program.

References

[1] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19129–19139, 2022. 6, 7

[2] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 6

[3] Daigang Cai, Lichen Zhao, Jing Zhang, Lu Sheng, and Dong Xu. 3djcg: A unified framework for joint dense captioning and visual grounding on 3d point clouds. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16464–16473, 2022. 6, 7

[4] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 2, 4

[5] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European Conference on Computer Vision*, pages 202–221. Springer, 2020. 2

[6] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Language conditioned spatial relation reasoning for 3d object grounding. *Advances in Neural Information Processing Systems*, 35:20522–20535, 2022. 4

[7] Sijin Chen, Hongyuan Zhu, Xin Chen, Yinjie Lei, Gang Yu, and Tao Chen. End-to-end 3d dense captioning with vote2cap-detr. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11124–11133, 2023. 7

[8] Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26428–26438, 2024. 1, 3, 6, 7

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

[10] Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3193–3203, 2021. 2, 4, 6, 7

[11] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, pages 5828–5839, 2017. 2

[12] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 4

[13] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[14] Zoey Guo, Yiwen Tang, Ray Zhang, Dong Wang, Zhigang Wang, Bin Zhao, and Xuelong Li. Viewrefer: Grasp the multi-view knowledge for 3d visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15372–15383, 2023. 3

[15] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems*, 36:20482–20494, 2023. 1, 2, 6

[16] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 5

[17] Haifeng Huang, Zehan Wang, Rongjie Huang, Luping Liu, Xize Cheng, Yang Zhao, Tao Jin, and Zhou Zhao. Chat-3d v2: Bridging 3d scene and large language models with object identifiers. *arXiv preprint arXiv:2312.08168*, 2023. 3, 4

[18] 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. *arXiv preprint arXiv:2311.12871*, 2023. 1, 2, 3, 4, 6, 7

[19] Shijia Huang, Yilun Chen, Jiaya Jia, and Liwei Wang. Multi-view transformer for 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15524–15533, 2022. 3

[20] Baoxiong Jia, Yixin Chen, Huangyue Yu, Yan Wang, Xuesong Niu, Tengyu Liu, Qing Li, and Siyuan Huang. Sceneverse: Scaling 3d vision-language learning for grounded scene understanding. In *European Conference on Computer Vision*, pages 289–310. Springer, 2024. 1, 2, 8

[21] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2

[22] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 1

[23] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 3

[24] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 6

[25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1

[26] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. In *International Conference on Learning Representations*, 2023. 2, 3, 4, 6, 7

[27] Yunze Man, Liang-Yan Gui, and Yu-Xiong Wang. Situational awareness matters in 3d vision language reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13678–13688, 2024. 6, 7

[28] OpenAI. Gpt-4 technical report. 2023. 1

[29] OpenAI. Hello gpt-4o. In *URL https://openai.com/index/hello-gpt-4o/*, 2024. 1

[30] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 6

[31] Zhangyang Qi, Ye Fang, Zeyi Sun, Xiaoyang Wu, Tong Wu, Jiaqi Wang, Dahua Lin, and Hengshuang Zhao. Gpt4point: A unified framework for point-language understanding and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26417–26427, 2024. 2

[32] N Reimers. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019. 6

[33] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d: Mask transformer for 3d semantic instance segmentation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8216–8223. IEEE, 2023. 2, 6

[34] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024. 1

[35] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1, 5

[36] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 6

[37] Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Nießner. Rio: 3d object instance relocalization in changing indoor environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7658–7667, 2019. 4

[38] Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, et al. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19757–19767, 2024. 2, 3

[39] Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. In *ECCV*, 2024. 1, 2, 3

[40] Xu Yan, Zhihao Yuan, Yuhao Du, Yinghong Liao, Yao Guo, Shuguang Cui, and Zhen Li. Comprehensive visual question answering on point clouds through compositional scene manipulation. *IEEE Transactions on Visualization and Computer Graphics*, 2023. 2

[41] Zhihao Yuan, Xu Yan, Yinghong Liao, Ruimao Zhang, Zhen Li, and Shuguang Cui. Instancerefer: Cooperative holistic understanding for visual grounding on point clouds through instance multi-level contextual referring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1791–1800, 2021. 2

[42] 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. 2

[43] Zhihao Yuan, Xu Yan, Yinghong Liao, Yao Guo, Guanbin Li, Zhen Li, and Shuguang Cui. X-trans2cap: Cross-modal knowledge transfer using transformer for 3d dense captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[44] Zhihao Yuan, Jinke Ren, Chun-Mei Feng, Hengshuang Zhao, Shuguang Cui, and Zhen Li. Visual programming for zero-shot open-vocabulary 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20623–20633, 2024. 2

[45] 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. 4

[46] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2911–2921, 2023. 6, 7