# ScanReason: Empowering 3D Visual Grounding with Reasoning Capabilities

Chenming Zhu[1,2], Tai Wang[2], Wenwei Zhang[2], Kai Chen[2], and Xihui Liu[1†]

[1] The University of Hong Kong
[2] Shanghai AI Laboratory
chaimzhu@connect.hku.hk
https://zcmax.github.io/projects/ScanReason

**Abstract.** Although great progress has been made in 3D visual grounding, current models still rely on explicit textual descriptions for grounding and lack the ability to reason human intentions from implicit instructions. We propose a new task called 3D reasoning grounding and introduce a new benchmark ScanReason which provides over 10K question-answer-location pairs from five reasoning types that require the synerization of reasoning and grounding. We further design our approach, ReGround3D, composed of the visual-centric reasoning module empowered by Multi-modal Large Language Model (MLLM) and the 3D grounding module to obtain accurate object locations by looking back to the enhanced geometry and fine-grained details from the 3D scenes. A chain-of-grounding mechanism is proposed to further boost the performance with interleaved reasoning and grounding steps during inference. Extensive experiments on the proposed benchmark validate the effectiveness of our proposed approach.

**Keywords:** 3D reasoning grounding · 3D visual grounding · Multi-modal large language models

## 1 Introduction

Understanding and reasoning in the 3D visual world is critical for applications such as robotics and AR, where embodied agents are expected to understand the 3D layout and predict the 3D locations of objects based on human instructions. The example in Fig. 1 demonstrates a scenario where the question can only be solved with a comprehensive understanding of the 3D scene and joint reasoning of the question and the 3D environment. However, current 3D visual grounding models [24, 39, 43, 50] trained on [1, 7] localize objects based on explicit descriptions of the object category, attribute, and 3D spatial relationships, and lack the ability to reason the user intentions and predict object locations with implicit human instructions such as "I am thirsty, can I have something to drink?".

To bridge the aforementioned gap and to push the boundaries of what embodied agents can understand and how they can interact with the 3D world, we

† Corresponding author: xihuiliu@eee.hku.hk

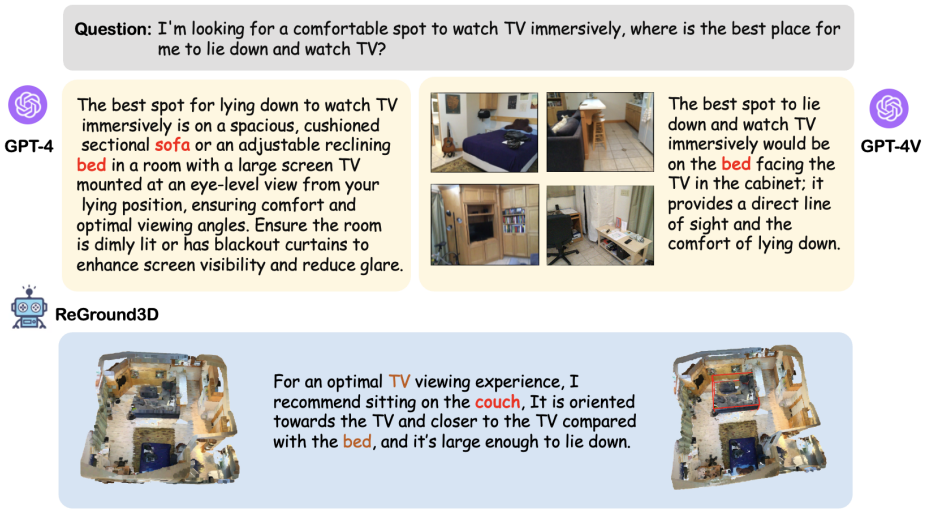


**Fig. 1:** For an embodied agent, they not only need to be able to understand the 3D environment and complex human instructions but also localize the target objects for interaction and navigation. Although GPT-4 (GPT-4V) have strong text (multi-modal) reasoning abilities, they lack the ability to directly perceive the 3D scene, understand the 3D spatial relationships, and output corresponding target object locations. Instead, our proposed method ReGround3D has the 3D perception, reasoning, and grounding capabilities in the real-world 3D environment.

propose a new task of 3D reasoning grounding and introduce a new benchmark named **ScanReason**. The task requires the model to conduct joint reasoning on the question and the 3D environment before predicting the 3D locations of target objects. We define five categories of 3D reasoning: spatial reasoning, functional reasoning, logical reasoning, emotional reasoning, and safety reasoning. The first two categories focus on the fundamental understanding of the 3D physical world, while the last three categories are built upon fundamental abilities to address user-centric real-world challenges. The benchmark comprises more than 10K question-answer-3D bounding box pairs from 2K scenes belonging to the five reasoning types mentioned above. The GPT-4-assisted data annotation process largely increases the efficiency of curating such a dataset.

We propose ReGround3D as an initial attempt to the new task of 3D Reasoning Grounding. Intuitively, for 3D grounding with implicit instructions, we need to first conduct reasoning on the language instructions and the coarse visual environment. Then with the idea of which object we want to find in mind, we look back to the 3D scene and ground the target object. For complex instructions, we may need to alternate the reasoning and look-back process for multiple iterations. Inspired by this intuition, our framework is composed of a visual-centric reasoning module and 3D grounding with geometry-enhanced look-back module, with a Chain-of-Grounding mechanism during inference to

alternately conduct reasoning and grounding for multiple rounds. Specifically, the visual-centric reasoning module conducts joint reasoning of the 3D scene and instructions with an MLLM. This module predicts a special token representing the semantic and location information of the target object, which is used for the grounding module. The 3D grounding module uses the output token embedding from the previous reasoning module to locate the target object by looking back at the fine-grained 3D scene representation. Unlike previous MLLMs attempting to directly predict bounding box coordinates, our look-back mechanism enables the model to capture more comprehensive 3D geometry and fine-grained object details for accurate 3D grounding. The Chain-of-Grounding mechanism is proposed to synergize reasoning and grounding, which allows multiple rounds alternating between reasoning and grounding during inference.

In summary, our contributions are threefold: 1) We propose a new task of 3D reasoning grounding which requires the model to synergize reasoning and grounding. We further introduce a new benchmark ScanReason, which comprises five reasoning types (spatial reasoning, functional reasoning, logical reasoning, emotional reasoning, and safety reasoning) for the task of 3D reasoning grounding in 3D scenes. 2) We design a new framework ReGround3D with a visual-centric reasoning module and a 3D grounding module with geometry-enhanced look-back. We further introduce a Chain-of-Grounding mechanism to boost the 3D reasoning grounding ability with a chain of interleaved reasoning and grounding steps. 3) Extensive experiments demonstrate the effectiveness of the our ReGround3D on the ScanReason benchmark for 3D reasoning grounding.

## 2 Related Work

**3D Vision and Language Learning** 3D Vision-language learning (3D-VL) is garnering increasing attention, with many 3D-VL tasks focusing on how to connect the 3D world with natural language. Among them, 3D Question Answering (3D QA) [2,45] aims to enable models to provide text answers based on natural language questions. Situation Question Answering in 3D Scenes (SQA3D) [29] requires an agent to first understand its location based on a text description, then provide a reasonable answer based on the surrounding environment, which can be seen as an extension of 3D QA in the embodied AI area. 3D Visual Grounding [1, 4, 7, 28, 31, 39] demands that models identify and locate target objects in a 3D scene based on given descriptions, outputting the objects' coordinates and 3D bounding boxes. These descriptions usually explicitly rely on the objects' attributes and their spatial relationships. 3D Dense Captioning [5, 8, 14, 16, 25, 36, 46, 49] requires models to output a series of object coordinates and corresponding scene-based descriptions based on a given scene. Different from these 3D-VL tasks, the questions in 3D reasoning grounding could be more implicit and complex.

**3D Visual Grounding** The task of 3D visual grounding is aimed at localizing the objects that are explicitly referred to by free-form guided language

expressions in the 3D scene. Inspired by the success of transformers in natural language processing, recent 3D visual grounding approaches [12,17,24,39,50,53] have started to adopt transformer [35] architectures for handling the complex relationships between language descriptions and 3D visual data. These methods leverage the self-attention mechanism of transformers to dynamically weigh the importance of different parts of the input data, facilitating a more effective grounding of textual descriptions in the 3D environment. Recent method [3] proposes a Chain-of-Thoughts module that predicts a chain of anchor objects that are subsequently utilized to localize the final target object. Compared with 3D visual grounding, our proposed 3D reasoning grounding requires the model to reason the complex question, ground target objects, and give the explanation at the same time.

**Multi-modal Large Language Models** Recently, there has been an increasing effort to extend the powerful complex reasoning and world knowledge capabilities of LLMs [18,34,47] to other modalities [6,9,10,19,26,44,51,52]. Among these works, some have aimed to enable LLMs to understand the 3D world. [20, 40] focus on delving into LLMs' ability to comprehend 3D objects, which can not be directly applied to 3D scenes. 3D-LLM [21] is the pioneering work that incorporates the 3D scene into LLM to carry out general 3D understanding tasks. However, by using 3D features constructed through projecting the 2D features of multi-view 2D images extracted by pre-trained 2D Vision-Language Models into 3D space, 3D-LLM struggles to directly capture the complex spatial relationships between objects and the structure of 3D scenes. [13,27] directly extract 3D features from the reconstructed 3D point cloud and support multi-modal visual prompts (text, images, 3D objects) in an MLLM. To alleviate the difficulty LLMs face in understanding complex 3D scenes, [22] choose to first explicitly segment the objects in the 3D scene and then perform multi-stage object-aware scene-text alignment to achieve 3D scene understanding. However, due to the lack of large-scale 3D-language alignment data and the intricate content of 3D scenes, although current MLLMs can achieve favorable performance in 3D scene understanding tasks, their localization performance is still significantly behind the 3D localization specialists. Our approach seeks to address this issue by introducing a 3D grounding model to enhance the localization capability of MLLMs.

## 3 ScanReason Benchmark

### 3.1 3D Reasoning Grounding Task

Given a 3D scene and an implicit free-form query, 3D reasoning grounding requires the model to predict the 3D bounding boxes of the target objects as well as the textual answers and explanations. As shown in Fig. 2, different from traditional 3D visual grounding, the queries of 3D reasoning grounding are implicit and complex, requiring strong reasoning, commonsense, and world knowledge. The number of target objects in 3D reasoning grounding is flexible and any object satisfying the requirements should be considered as the target object.

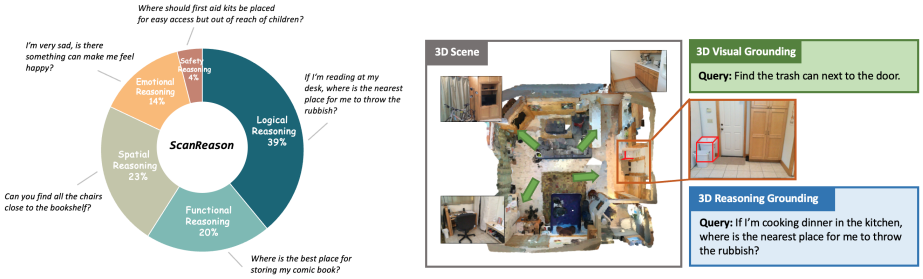


**Fig. 2:** The left side figure shows the overall of our ScanReason dataset. For each reasoning category, we designed different prompts to generate corresponding questions. And the right side figure shows the differences between the traditional 3D visual grounding task and our proposed 3D reasoning grounding task.

### 3.2 Question Types

To comprehensively evaluate the 3D reasoning grounding abilities, we define 5 types of questions depending on which type of reasoning is required. *Spatial reasoning* and *functional reasoning* require a fundamental understanding of the 3D physical world, while *logical reasoning*, *emotional reasoning*, and *safety reasoning* are high-level reasoning skills built upon the two fundamental reasoning abilities to address user-centric real-world applications, as shown in Fig. 2.

**Spatial Reasoning** measures models' understanding of 3D spatial relationships among objects in 3D scene. It encompasses the ability to comprehend the layout of the 3D scene and the 3D location of objects within it, which could serve as the foundation for navigating or planning movements in the 3D environment.

**Functional Reasoning** involves understanding and inferring the purpose, function, or affordance of objects within the 3D scene. For example, functional reasoning allows an embodied agent to recognize that a chair is for sitting, a lamp is for lighting, and a refrigerator is for storing food at low temperatures. Such understanding enables the embodied agent to assist users and to perform complex tasks more effectively (e.g., turning on a lamp when the room gets dark, or navigating to a refrigerator to fetch a drink).

**Logical Reasoning** allows an embodied agent to not only understand its environment but also to interact with it in a goal-directed manner. For example, given a question shown in Fig. 2: "If I'm cooking dinner in the kitchen, where is the nearest place to throw the rubbish?", an agent needs to use such reasoning ability to infer the location of objects (in this question, a rubbish bin) based on their function and spatial relationships under the specific setting (the kitchen).

**Emotional Reasoning** plays a critical role in human-robot interaction, where the target objects are determined by understanding human emotions, preferences, and behavioral patterns. This ability makes the embodied agents more attuned to the emotional and psychological states of humans, allowing them to provide more personalized, empathetic, and contextually appropriate responses

and solutions, such as: "I'm very sad, is there something that can make me feel happy?" shown in 2.

**Safety Reasoning** focuses on preventing harm and ensuring the well-being of humans in the 3D environment. It requires the embodied agent to identify and assess the risk and make safety-aware decisions, such as: "Where should first aid kits be placed for easy access but out of reach of children?".

### 3.3 Automatic Data Annotation with GPT-4

We leverage the 3D scenes and bounding box annotations from the EmbodiedScan dataset [37] and apply GPT-4 [30] to automatically generate question-answer-location pairs for the five question types respectively. Specifically, we provide GPT-4 with the categories and bounding box locations of all objects in the scene, and ask GPT-4 to generate questions and answers with the target object ids of the provided objects. The details can be found in the Appendix.

In total, our ScanReason dataset consists of 12929 complex reasoning question-answer-location pairs from 1456 scenes, which are split into 11455 training pairs and 1474 validation pairs. All 1474 validation question-answer pairs have been manually verified, including 342 spatial reasoning questions, 287 functional reasoning questions, 581 logical reasoning questions, 211 emotional questions, and 53 safety reasoning questions. We provide detailed statistics and more examples of our dataset in the Appendix.

### 3.4 Evaluation Metric

To evaluate the accuracy of predicted objects and their locations for a flexible number of ground-truth objects, we follow the evaluation metric of 3D object detection tasks. Specifically, we adopt Acc@$k$IoU as our metric, where $k$ is the threshold for the Intersection of Union (IoU) between positive predictions and ground truths. We evaluate the performance under $k = 0.25$ in our experiments.

## 4 Method

Solving the task of 3D reasoning grounding requires the synergization of the perception, reasoning, and grounding capability of the embodied agent. Intuitively, we can first conduct reasoning based on implicit instruction such as "where should first aid kits be placed?" and the visual environment. The reasoning process provides us with information about the rough location and semantics of the object we are looking for. Then keeping that information in mind, we look back to the 3D environment to precisely locate the object. For complex scenarios, alternate reasoning and looking back are required for multiple rounds until we obtain the final answer.

Inspired by this intuition, we propose ReGround3D, consisting of a visual-centric reasoning module and a 3D grounding module with geometry-enhanced look-back, as illustrated in Fig. 3. The visual-centric reasoning module (Sec. 4.1)

performs joint reasoning of language instruction and visual scene, and predicts a special <LOC> token representing the grounding information. The 3D grounding module (Sec. 4.2) looks back to the original 3D scene with comprehensive geometry information and fine-grained details. It takes the hidden embedding of the <LOC> token containing grounding-related information from the 3D features and eventually predicts the 3D locations of the target objects. Furthermore, we propose a Chain-of-Grounding mechanism (CoG) (Sec. 4.3), *i.e.*, a chain of interleaved reasoning and grounding steps, to further synergize the grounding and reasoning capability for the 3D reasoning grounding task, as illustrated in Fig. 4.

### 4.1 Visual-Centric Reasoning

Due to the complexity of the 3D scene and user instructions, particularly given the implicit intention behind the human instruction, ReGround3D starts with a visual-centric reasoning module that can perceive the scene, comprehend the human instructions, and conduct joint reasoning of 3D scene and instructions. We believe the reasoning process eventually implies the grounding intention, *i.e.*, implicitly encodes the information indicating the target object to solve the task. Thus, we design the visual-centric reasoning module to predict grounding queries for localizing the target objects in the following 3D grounding module.

Specifically, for simplicity, we leverage 3D-LLM to serve as the visual-centric reasoning module because of its strong reasoning abilities inherited from the LLM. Based on the BLIP2 architecture [26], 3D-LLM uses pre-trained image encoders to extract multi-view 2D image features and back-project them into 3D spaces. The visual features are encoded by the Q-Former to produce 32 tokens as the visual input to the LLM. By leveraging the pre-trained image encoders and the Q-Former, the visual tokens encode rich semantics but lack the 3D structures, spatial interactions, and fine-grained details. Therefore, instead of directly predicting the object locations by the 3D-LLM, we ask the 3D-LLM to predict the feature representation as the output of the reasoning process, and the predicted feature is further used to ground the target object in the grounding module.

To enable the prediction of the grounding feature, we expand the original vocabulary of the 3D-LLM with a special <LOC> token. The <LOC> token is laden with the contextual scene and the target object information which can guide the 3D grounding module to accurately localize target objects.

### 4.2 3D Grounding with Geometry-Enhanced Look-Back

Once obtaining the <LOC> token, ReGround3D extracts the last-layer embedding $h_{loc}$ of the <LOC> token and sends it into the 3D grounding module to predict the 3D bounding boxes. The 3D grounding module is devised with a "look-back" mechanism which allows the model to access the 3D geometry and fine-grained details from a 3D point cloud encoder. The fine-grained geometry-enhanced 3D visual features and $h_{loc}$ are sent into a query selection module to retrieve the most

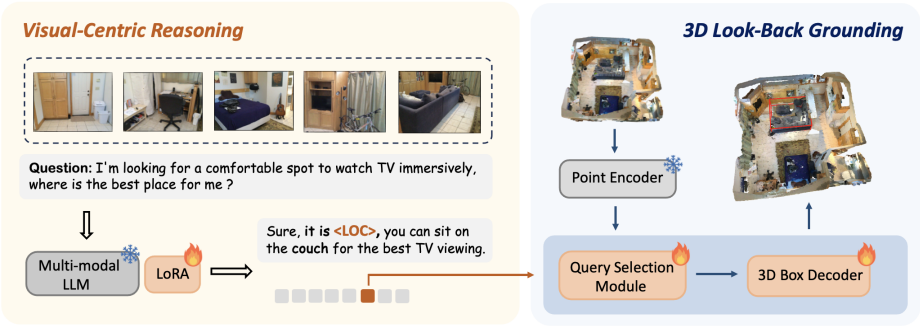


**Fig. 3:** The pipeline of ReGround3D. Given the 3D scene and human instruction, the visual reasoning module first performs joint 3D scene and instruction reasoning, and then guide the 3D grounding module to look-back the 3D scene and perform target object location.

relevant object features. Those features are further decoded into 3D bounding boxes with the 3D box decoder.

**3D Visual Encoder** Unlike the 2D image encoder used for 3D-LLM, the 3D visual encoder directly extracts features from 3D point clouds to capture more geometric and spatial information about the 3D structures and fine-grained details. The powerful 3D visual encoder which captures comprehensive geometry, structure, layout, and fine-grained information is critical to accurate 3D grounding. Subsequently, the 3D features $f_{scene}$ produced by the 3D visual encoder and the grounding feature $h_{loc}$ from 3D-LLM are sent to the query selection module.

**Query Selection Module** We adopt a cross-attention mechanism, where we treat $f_{scene}$ as $Q$ (Query), and $h_{loc}$ as both $K$ (Key) and $V$ (Value), to implicitly obtain a feature-level reasoning activation heatmap. During this scene look-back process, this module roughly locates scene features that have a high response to the `<LOC>` token. We then select the $k$ most relevant features as the object query $f_{query}$.

**3D Box Decoder** is a classical transformer decoder, which consists of M transformer decoder layers, In each decoder layer, the object queries $f_{query}$ go through the text feature $h_{loc}$ cross-attention layer and scene feature $f_{scene}$ cross-attention layer. Finally, the prediction head takes the updated object queries as input and predicts the final 3D locations and matching score.

**Discussion** In comparison to previous works [11,13,15,21] that directly predicts the bounding boxes by the LLM, the extra grounding module has the following advantages: 1) Based on the MLLM reasoning results, it allows the grounding module to perceive the scene again and focus on the region under the implicit guidance of MLLM, adapted to the user queries and the reasoning results. 2) The scene representation perceived in the grounding module can be more precise and fine-grained, which is complementary to the visual-centric reasoning module. 3) The two-step reasoning-grounding pipeline is flexible and can generalized

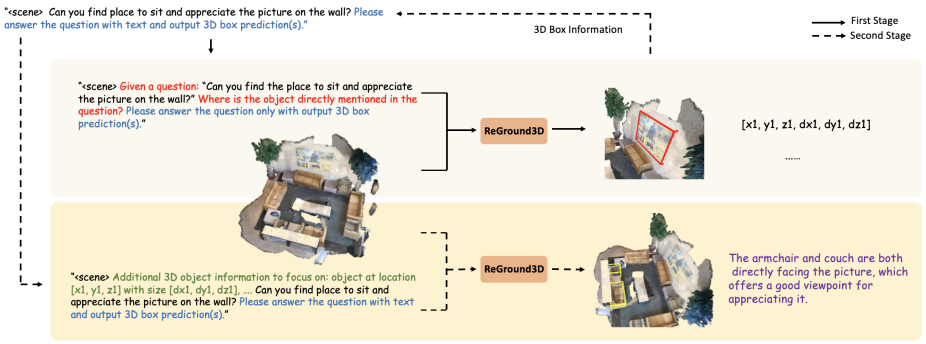


**Fig. 4:** Illustration of Chain-of-Grounding (CoG) Mechanism

to other types of predicting formats such as segmentation masks (by simply replacing the 3D box decoder with a 3D mask decoder).

### 4.3 Chain-of-Grounding Mechanism

The existing design conducts the reasoning and grounding process sequentially, *i.e.*, the reasoning process is finished before grounding. We argue that the grounding results can also facilitate the reasoning process, especially for those requiring spatial information. Thus, to further synergize the reasoning and grounding process, inspired by chain-of-thought (CoT) [38], we propose Chain-of-Grounding (CoG), which introduces a chain of interleaved steps of reasoning and grounding to find the targeted objects during inference, as shown in Fig. 4. Such a process allows the model to actively find relevant objects that help solve the problem, and then conduct reasoning with the assistance of the additional information of these relevant objects so that the model can more precisely find the target objects.

Specifically, given a question provided by users, CoG translates it into another question of finding the explicitly mentioned objects in the original question. The generated new question is sent into ReGround3D to ground the objects mentioned in the original question in the 3D scene with corresponding confidence scores. An object can be seen as a successfully located object when its confidence score is above the threshold, and the located object information could serve as explicit guidance for 3D-LLM in the next reasoning stage. As shown in Fig. 4, after obtaining the 3D information of objects explicitly mentioned in the original question, the located object information is inserted to update the question. The updated question is then sent to ReGround3D to perform reasoning and grounding to output the target object locations.

### 4.4 Instruction Tuning

**Training Objective** We use the pre-trained weights of 3D-LLM as the initialization for the visual-centric reasoning module. Except for freezing the 3D

**Table 1:** We list part of data templates used to train ReGround3D for each task.

| Task Name | Text Instructions | Output Type Templates | Expected Output |
|---|---|---|---|
| 3D Visual Grounding | <scene> Here is a description about an object: "<expr>", where is the object in the 3D scene? | Please answer the question only with output 3D box prediction(s). | It is <LOC>. |
| 3D/Spatial Question Answering | <scene> Answer the question: "<question>" | Please answer the question only with text, do not output 3D box prediction(s). | <answer>. |
| | <scene> <situation>, <question> | Please answer the question only with text, do not output 3D box prediction(s). | <answer>. |
| | <scene><question> | Please answer the question only with text, do not output 3D box prediction(s). | <answer>. |
| 3D Object Detection | <scene> Where is the <category> in this 3D scene? | Please answer the question only with output 3D box prediction(s). | Sure, <LOC>. |
| 3D Reasoning Grounding | <scene> Answer the question: "<question>" | Please answer the question with text and output 3D box prediction(s). | Sure, <LOC>, <reason> |

visual encoder pre-trained on [37], the rest of the parameters in the visual-centric reasoning module and 3D grounding module in our framework are trained in an end-to-end manner. The training supervision is a weighted sum of the next token prediction loss from 3D-LLM and the 3D detection loss from the 3D grounding module.

$$\mathcal{L} = \lambda_{text}\mathcal{L}_{text} + \lambda_{det}\mathcal{L}_{det} \tag{1}$$

The 3D detection loss is defined following:

$$\mathcal{L}_{det} = \lambda_{IOU}\mathcal{L}_{IOU} + \lambda_{contrast}\mathcal{L}_{contrast} \tag{2}$$

**Instruction Tuning Dataset** We load the pre-trained weights of 3D-LLM and the 3D visual encoder, and finetune the LoRA of 3D-LLM, the query selection module, and the 3D box decoder with an instruction tuning dataset. To construct the instruction tuning dataset, we reformulate the data annotations from existing 3D datasets into question-answer or question-answer-bounding-box pairs. Specifically, the 3D visual grounding data from ScanRefer [7], SR3D, NR3D [1] and the 3D object detection data from EmbodiedScan [37] are formulated into question-answer-bbox pairs, and the 3D/spatial question answering data from SR3D [1], CLEVR3D [41], SQA3D [29] are formulated into question-answer pairs without bounding boxes. The information shown in Tab. 1 illustrates how we unify the instruction and output with task-specific templates. More details can be found in the Appendix. The reformulated data combined with our proposed ScanReason dataset, serve as the instruction tuning dataset.

## 5 Experiment

### 5.1 Implementation Details

**Network Architecture** For the 3D grounding module, we adopt the pre-trained point cloud encoder as the 3D visual encoder. During the training stage, We use LoRA to efficiently finetune the 3D-LLM to preserve the original 3D scene understanding capability and reduce the computation costs. The number of object queries $k$ in the query selection module is set to 256.

**Training Parameters** The training is done on 8 NVIDIA A100 GPUs. We adopt the AdamW optimizer with a learning rate of 3e-4 and use a learning rate scheduler WarmupDecayLR with the warmup steps of 100. The total batch size is set to be 16. The loss weight parameters $\lambda_{text}$ and $\lambda_{det}$ in total loss $\mathcal{L}$ are set to 1.0 and 1.0, respectively, and the weight $\lambda_{IOU}$ and $\lambda_{contrast}$ in $\mathcal{L}_{det}$ are set to 1.0 and 1.0.

### 5.2 Results Comparison

**Evaluation on 3D Visual Grounding** In order to verify the superiority of our model in grounding ability and facilitate comparison between current models, we report the explicit grounding performance on the existing 3D visual grounding task. Since the evaluation settings of Nr3D and Sr3D [1] are based on ground-truth object proposals, while ScanRefer [7] requires models to output 3D bounding boxes, we choose ScanRefer as our benchmark for comparison. We divide the existing methods into two categories, one is the grounding model designed specifically for the 3D visual grounding task, and the other is generalist MLLMs which can understand a variety of 3D vision-language tasks. The original 3D-LLM embeds 3D locations in the vocabularies and represents the grounded 3D bounding boxes by a sequence of discrete location tokens. However, since the fine-tuned model of 3D-LLM on ScanRefer and related location tokens implementation are not accessible, we adapt 3D-LLM to directly output 3D numerical coordinates representing 3D bounding boxes by fine-tuning the pre-trained model on our reformulated 3D visual grounding data, denotes as 3D-LLM (vg). As shown in Tab. 2, current generalist MLLM models still lag behind the specialist models in terms of grounding ability. By incorporating the 3D grounding module into MLLM, ReGround3D shows the SOTA performance on the traditional 3D visual grounding task.

**Table 2:** Results on 3D visual grounding task among ReGround3D (ours) and existing methods.

| Type | Methods | Acc@0.25 | Acc@0.5 |
|---|---|---|---|
| Specialists | ScanRefer [7] | 37.3 | 24.3 |
| | MVT [23] | 40.8 | 33.3 |
| | 3DVG-Trans [48] | 45.9 | 34.5 |
| | ViL3DRel [12] | 47.9 | 37.7 |
| | BUTD-DETR [24] | 52.2 | 39.8 |
| | L3Det [50] | 52.8 | 40.2 |
| Generalized MLLMs | LLM-Grounder [42] | 17.1 | 5.3 |
| | 3D-LLM [21] | 30.3 | - |
| | 3D-LLM(vg) [21] | 33.1 | 28.7 |
| | Chat3D-v2 [22] | 35.9 | 30.4 |
| ours | ReGround3D | **53.1** | **41.1** |

**Table 3:** Results (Acc@0.25) on 3D reasoning grounding task among ReGround3D (ours) and existing methods.

| Methods | LLM | Spatial | Functional | Logical | Emotional | Safety | Overall |
|---|---|---|---|---|---|---|---|
| Mask3D [32] + InternLM2-7B [33] | InternLM2-7B | 10.34 | 36.12 | 9.98 | 8.21 | 8.99 | 14.86 |
| 3D-LLM(vg) [21] | FlanT5$_{XL}$-3B | 18.31 | 17.42 | 10.97 | 8.12 | 6.33 | 13.29 |
| Chat3D-v2 [22] | Vicuna-7B | 20.21 | 18.39 | 11.32 | 7.98 | 9.88 | 14.98 |
| ReGround3D* | FlanT5$_{XL}$-3B | 30.76 | 29.8 | 18.67 | 19.22 | 17.12 | 23.27 |
| ReGround3D | FlanT5$_{XL}$-3B | 32.98 | 36.23 | 26.99 | 23.12 | 22.98 | 28.98 |
| ReGround3D (CoG) | FlanT5$_{XL}$-3B | **34.71** | **36.79** | **29.11** | **24.03** | **23.21** | **30.62** |

**Evaluation on 3D Reasoning Grounding** The various visual grounding models rely on explicit text-object alignment in the input object expression to achieve localization, which falls to be applied to 3D reasoning grounding task. We performed a comparison between our proposed method ReGround3D and existing MLLM methods, including 3D-LLM(vg) and Chat3D-v2 [22]. Besides, inspired by Chat-3D v2, which first segments objects, then equips them with unique object identifiers to conduct effective object grounding, we set up a LLM-based 3D reasoning grounding baseline: We first segment the objects from the scene using a 3D instance segmentor [32], then convert the segmented object information including their categories and 3D bounding boxes into text as input of LLM (InternLM2-7B [33]). Besides, to better validate the performance of our model and ensure a fair comparison, we remove the ScanReason dataset from the training data, denoted as ReGround3D*. As shown in Tab. 3, the LLM-based reasoning method (Mask3D [32] + InternLM2-7B [33]) possesses a very strong functional reasoning ability, but struggles to understand of 3D spatial relationship. Additionally, irrespective of whether ScanReason is used in training, our model significantly outperforms the existing MLLMs. By synergizing the reasoning and grounding process utilizing the Chain-of-Grounding (CoG) mechanism, the 3D reasoning grounding performance of ReGround3D can be further improved (from 28.98 to 30.62), especially on the spatial reasoning and logical reasoning questions. The qualitative results comparison shown in Fig. 5 demonstrates the superiority of our method in reasoning human complex instruction based on the 3D scene.

### 5.3 Ablation Study

In this section, we conduct an extensive ablation study to verify the effectiveness of each component.

**Effectiveness of 3D Grounding Module** In order to more comprehensively verify the effectiveness of the 3D grounding module we proposed, we step by step verify the performance changes from 3D-LLM to ReGround3D. Apart from the 3D-LLM(vg), we fine-tuned the 3D-LLM model respectively on full reformulated existing data and all instruction tuning data, including ScanReason, denoted as

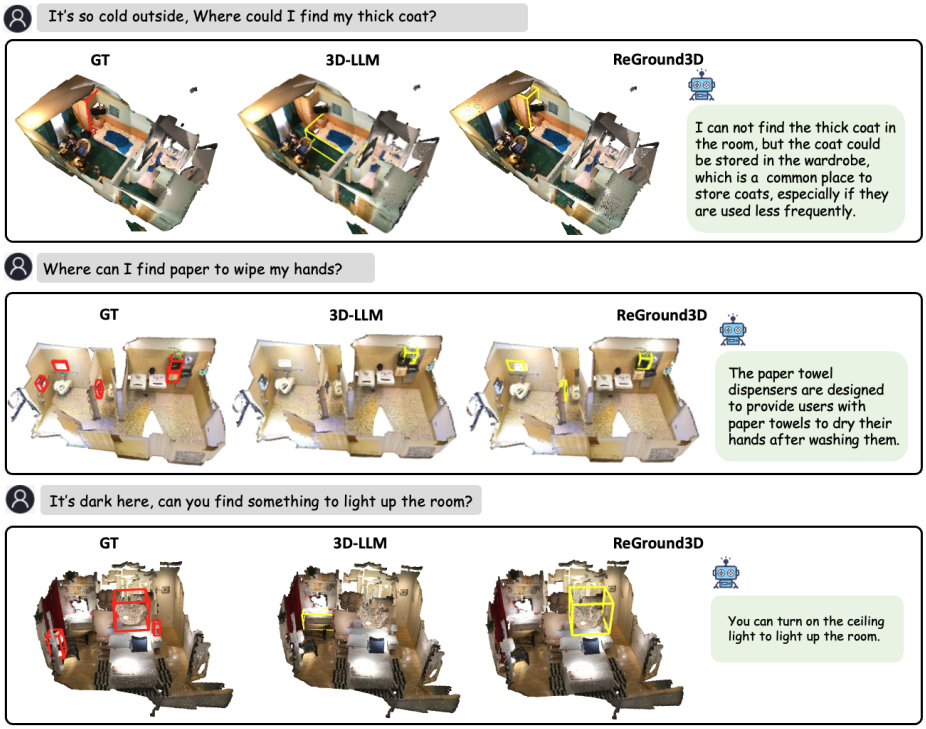


**Fig. 5:** Visualization comparison of 3D reasoning grounding capability between ReGround3D and 3D-LLM. Our method could achieve much more accurate grounding results which satisfy the implicit question intention, and give the corresponding explanation at the same time. More illustrations are given in the Appendix.

3D-LLM (full) and 3D-LLM (full+sr). All the <LOC> tokens in answers of the data used for fine-tuning 3D-LLM have been converted to the numerical coordinates of the corresponding boxes. The results in Tab. 4a show that with the same tuning dataset, our ReGround3D achieves far superior performance to 3D-LLM (28.98 vs. 19.21) by introducing the 3D grounding module.

**Effectiveness of Instruction Tuning Dataset** Tab. 4b showcases the impact of different training data types on 3D visual grounding performance. 3D Object Detection (3D OD) provides the explicit 3D semantic category and visual alignment whereas 3D Question Answering (3D QA) data injects the basic 3D scene understanding ability into the model, which has certain benefits for the visual grounding ability of the model. In addition, we find that training with 3D reasoning grounding dataset can further improve the performance on 3D visual grounding.

**Table 4:** Ablation study on effectiveness of 3D grounding module and training data

**(a)** When using the same tuning dataset, ReGround3D achieves far better performance on ScanReason than 3D-LLM.

| Methods | Acc@0.25 |
|---|---|
| 3D-LLM(vg) [21] | 13.29 |
| 3D-LLM(full) [21] | 15.31 |
| 3D-LLM(full+sr) [21] | 19.21 |
| ReGround3D(full+sr) | **28.98** |

**(b)** Ablation study on training data. We evaluate through the metric of accuracy on the val set of the ScanRefer dataset.

| Dataset | | | | Acc@0.25 | Acc@0.5 |
|---|---|---|---|---|---|
| VG | OD | VQA | RD | | |
| | ✓ | ✓ | | 19.3 | 14.2 |
| ✓ | | ✓ | | 48.7 | 37.6 |
| ✓ | ✓ | | | 49.2 | 38.1 |
| ✓ | ✓ | ✓ | | 51.8 | 39.4 |
| ✓ | ✓ | ✓ | ✓ | **53.1** | **41.1** |

**Discussion of CoG Mechanism** While the CoG Mechanism boosts the performance with interleaved reasoning and grounding steps during inference, it uses the relevant object information explicitly presented in the question to help find the target objects. A natural question arises: if we input the information of all the objects instead of relevant objects into ReGround3D during CoG, will this make the model more accurately find the target objects? We first use the existing 3D Segmentor [32] to segment all objects in the scene, then update the original question using all the object 3D bounding boxes information according to the template in Sec. 4.3 and send into ReGround3D. However, experiments show that using all the object information will instead reduce the 3D reasoning grounding performance from 28.98 to 27.67. One possible reason could be the model's attention is dispersed by too many irrelevant objects.

## 6 Conclusion

This paper introduces a new 3D vision language learning task: 3D reasoning grounding, which requires the model to perform active reasoning over complex and implicit human instruction, localize the target objects, and give corresponding explanations. Furthermore, we propose ScanReason, a new dataset and benchmark to further unlock and thoroughly evaluate the 3D reasoning grounding capability. Based on this dataset, we propose a novel approach: ReGround3D, which utilizes the strong reasoning capability of MLLM guiding the 3D grounding module to obtain accurate object locations, and a Chain of Grounding (CoG) mechanism is presented to further boost the performance with interleaved reasoning and grounding steps during inference. We believe that our work will further the natural interaction between embodied agents and humans in open 3D environments. For the current ScanReason benchmark, we find that the questions in three high-level 3D reasoning categories may have overlaps. For a certain reasoning question, similar questions may appear in one or two other categories. We leave the problem as a future challenge for better reasoning grounding ability evaluation.

## Acknowledgements

This work is supported in part by HKU Startup Fund, HKU Seed Fund for Basic Research, HKU Seed Fund for Translational and Applied Research, HKU IDS research Seed Fund, and HKU Fintech Academy R&D Funding.

## References

1. Achlioptas, P., Abdelreheem, A., Xia, F., Elhoseiny, M., Guibas, L.: Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. pp. 422–440. Springer (2020)
2. Azuma, D., Miyanishi, T., Kurita, S., Kawanabe, M.: Scanqa: 3d question answering for spatial scene understanding. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19129–19139 (2022)
3. Bakr, E.M., Ayman, M., Ahmed, M., Slim, H., Elhoseiny, M.: Cot3dref: Chain-of-thoughts data-efficient 3d visual grounding. arXiv preprint arXiv:2310.06214 (2023)
4. Baroni, M.: Linguistic generalization and compositionality in modern artificial neural networks. Philosophical Transactions of the Royal Society B **375**(1791), 20190307 (2020)
5. Cai, D., Zhao, L., Zhang, J., Sheng, L., Xu, D.: 3djcg: A unified framework for joint dense captioning and visual grounding on 3d point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16464–16473 (2022)
6. Chen, C., Qin, R., Luo, F., Mi, X., Li, P., Sun, M., Liu, Y.: Position-enhanced visual instruction tuning for multimodal large language models. arXiv preprint arXiv:2308.13437 (2023)
7. Chen, D.Z., Chang, A.X., Nießner, M.: Scanrefer: 3d object localization in rgb-d scans using natural language. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX. pp. 202–221. Springer (2020)
8. Chen, D.Z., Wu, Q., Nießner, M., Chang, A.X.: D3net: A speaker-listener architecture for semi-supervised dense captioning and visual grounding in rgb-d scans. arXiv preprint arXiv:2112.01551 (2021)
9. Chen, F., Han, M., Zhao, H., Zhang, Q., Shi, J., Xu, S., Xu, B.: X-LLM: Bootstrapping advanced large language models by treating multi-modalities as foreign languages. arXiv preprint arXiv:2305.04160 (2023)
10. Chen, J., Zhu, D., Shen, X., Li, X., Liu, Z., Zhang, P., Krishnamoorthi, R., Chandra, V., Xiong, Y., Elhoseiny, M.: Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. arXiv preprint arXiv:2310.09478 (2023)
11. Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., Zhao, R.: Shikra: Unleashing multimodal llm's referential dialogue magic. arXiv preprint arXiv:2306.15195 (2023)
12. Chen, S., Guhur, P.L., Tapaswi, M., Schmid, C., Laptev, I.: Language conditioned spatial relation reasoning for 3d object grounding. Advances in Neural Information Processing Systems **35**, 20522–20535 (2022)

13. Chen, S., Chen, X., Zhang, C., Li, M., Yu, G., Fei, H., Zhu, H., Fan, J., Chen, T.: Ll3da: Visual interactive instruction tuning for omni-3d understanding, reasoning, and planning. arXiv preprint arXiv:2311.18651 (2023)
14. Chen, S., Zhu, H., Chen, X., Lei, Y., Yu, G., Chen, T.: End-to-end 3d dense captioning with vote2cap-detr. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11124–11133 (2023)
15. Chen, T., Saxena, S., Li, L., Fleet, D.J., Hinton, G.: Pix2seq: A language modeling framework for object detection. arXiv preprint arXiv:2109.10852 (2021)
16. Chen, Z., Gholami, A., Nießner, M., Chang, A.X.: Scan2cap: Context-aware dense captioning in rgb-d scans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3193–3203 (2021)
17. Chen, Z., Hu, R., Chen, X., Nießner, M., Chang, A.X.: Unit3d: A unified transformer for 3d dense captioning and visual grounding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18109–18119 (2023)
18. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., et al.: Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416 (2022)
19. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning (2023)
20. Han, J., Zhang, R., Shao, W., Gao, P., Xu, P., Xiao, H., Zhang, K., Liu, C., Wen, S., Guo, Z., et al.: Imagebind-llm: Multi-modality instruction tuning. arXiv preprint arXiv:2309.03905 (2023)
21. Hong, Y., Zhen, H., Chen, P., Zheng, S., Du, Y., Chen, Z., Gan, C.: 3d-llm: Injecting the 3d world into large language models. Advances in Neural Information Processing Systems **36** (2024)
22. Huang, H., Wang, Z., Huang, R., Liu, L., Cheng, X., Zhao, Y., Jin, T., Zhao, Z.: Chat-3d v2: Bridging 3d scene and large language models with object identifiers. arXiv preprint arXiv:2312.08168 (2023)
23. Huang, S., Chen, Y., Jia, J., Wang, L.: Multi-view transformer for 3d visual grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15524–15533 (2022)
24. Jain, A., Gkanatsios, N., Mediratta, I., Fragkiadaki, K.: Bottom up top down detection transformers for language grounding in images and point clouds. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI. pp. 417–433. Springer (2022)
25. Jiao, Y., Chen, S., Jie, Z., Chen, J., Ma, L., Jiang, Y.G.: More: Multi-order relation mining for dense captioning in 3d scenes. arXiv preprint arXiv:2203.05203 (2022)
26. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
27. Li, M., Chen, X., Zhang, C., Chen, S., Zhu, H., Yin, F., Yu, G., Chen, T.: M3dbench: Let's instruct large models with multi-modal 3d prompts. arXiv preprint arXiv:2312.10763 (2023)
28. Luo, J., Fu, J., Kong, X., Gao, C., Ren, H., Shen, H., Xia, H., Liu, S.: 3d-sps: Single-stage 3d visual grounding via referred point progressive selection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16454–16463 (2022)
29. Ma, X., Yong, S., Zheng, Z., Li, Q., Liang, Y., Zhu, S.C., Huang, S.: Sqa3d: Situated question answering in 3d scenes. arXiv preprint arXiv:2210.07474 (2022)

30. OpenAI: Gpt-4 technical report (2023)
31. Roh, J., Desingh, K., Farhadi, A., Fox, D.: Languagerefer: Spatial-language model for 3d visual grounding. In: Conference on Robot Learning. pp. 1046–1056. PMLR (2022)
32. Schult, J., Engelmann, F., Hermans, A., Litany, O., Tang, S., Leibe, B.: Mask3d: Mask transformer for 3d semantic instance segmentation. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 8216–8223. IEEE (2023)
33. Team, I.: Internlm: A multilingual language model with progressively enhanced capabilities (2023)
34. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
35. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
36. Wang, H., Zhang, C., Yu, J., Cai, W.: Spatiality-guided transformer for 3d dense captioning on point clouds. arXiv preprint arXiv:2204.10688 (2022)
37. Wang, T., Mao, X., Zhu, C., Xu, R., Lyu, R., Li, P., Chen, X., Zhang, W., Chen, K., Xue, T., et al.: Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. arXiv preprint arXiv:2312.16170 (2023)
38. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems **35**, 24824–24837 (2022)
39. Wu, Y., Cheng, X., Zhang, R., Cheng, Z., Zhang, J.: Eda: Explicit text-decoupling and dense alignment for 3d visual and language learning. arXiv preprint arXiv:2209.14941 (2022)
40. Xu, R., Wang, X., Wang, T., Chen, Y., Pang, J., Lin, D.: Pointllm: Empowering large language models to understand point clouds. arXiv preprint arXiv:2308.16911 (2023)
41. Yan, X., Yuan, Z., Du, Y., Liao, Y., Guo, Y., Li, Z., Cui, S.: Clevr3d: Compositional language and elementary visual reasoning for question answering in 3d real-world scenes. arXiv preprint arXiv:2112.11691 (2021)
42. Yang, J., Chen, X., Qian, S., Madaan, N., Iyengar, M., Fouhey, D.F., Chai, J.: Llm-grounder: Open-vocabulary 3d visual grounding with large language model as an agent. arXiv preprint arXiv:2309.12311 (2023)
43. Yang, Z., Zhang, S., Wang, L., Luo, J.: Sat: 2d semantics assisted training for 3d visual grounding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1856–1866 (2021)
44. Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al.: mplug-owl: Modularization empowers large language models with multimodality. arXiv preprint arXiv:2304.14178 (2023)
45. Ye, S., Chen, D., Han, S., Liao, J.: 3d question answering. IEEE Transactions on Visualization and Computer Graphics (2022)
46. Yuan, Z., Yan, X., Liao, Y., Guo, Y., Li, G., Cui, S., Li, Z.: 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. pp. 8563–8573 (2022)
47. Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., et al.: Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068 (2022)

48. Zhao, L., Cai, D., Sheng, L., Xu, D.: 3dvg-transformer: Relation modeling for visual grounding on point clouds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2928–2937 (2021)
49. Zhong, Y., Xu, L., Luo, J., Ma, L.: Contextual modeling for 3d dense captioning on point clouds. arXiv preprint arXiv:2210.03925 (2022)
50. Zhu, C., Zhang, W., Wang, T., Liu, X., Chen, K.: Object2scene: Putting objects in context for open-vocabulary 3d detection. arXiv preprint arXiv:2309.09456 (2023)
51. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)
52. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)
53. Zhu, Z., Ma, X., Chen, Y., Deng, Z., Huang, S., Li, Q.: 3d-vista: Pre-trained transformer for 3d vision and text alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2911–2921 (2023)