Reason3D: Searching and Reasoning 3D Segmentation via Large Language Model

Kuan-Chih Huang[1] Xiangtai Li[2♦] Lu Qi[1♠] Shuicheng Yan[2,3] Ming-Hsuan Yang[1]
[1]University of California, Merced [2]Skywork AI, Singapore [3]Nanyang Technological University

https://reason3d.github.io/

Figure 1. **Overview.** We propose **Reason3D**, a novel LLM-based 3D point cloud searching and reasoning framework that can output dense segmentation masks based on textural descriptions. Our Reason3D can handle four tasks involving 1) 3D Reasoning, 2) 3D Hierarchical Searching, 3) 3D Express Referring, and 4) 3D QA with responding dense segmentation masks.

Abstract

Recent advancements in multimodal large language models (LLMs) have demonstrated significant potential across various domains, particularly in concept reasoning. However, their applications in understanding 3D environments remain limited, primarily offering textual or numerical outputs without generating dense, informative segmentation masks. This paper introduces Reason3D, a novel LLM designed for comprehensive 3D understanding. Reason3D processes point cloud data and text prompts to produce textual responses and segmentation masks, enabling

♦Project lead, ♠Corresponding author

advanced tasks such as 3D reasoning segmentation, hierarchical searching, express referring, and question answering with detailed mask outputs. We propose a hierarchical mask decoder that employs a coarse-to-fine approach to segment objects within expansive scenes. It begins with a coarse location estimation, followed by object mask estimation, using two unique tokens predicted by LLMs based on the textual query. Experimental results on large-scale ScanNet and Matterport3D datasets validate the effectiveness of our Reason3D across various tasks.

1. Introduction

Recently, large language models (LLMs) [18, 39, 40] have significantly advanced their capabilities in sophisticated reasoning within natural language processing. Building on these advancements, a new class of models known as Multimodal Large Language Models (MLLMs) [3, 3, 12, 23, 24, 24, 28, 33, 34, 47, 48, 50] has emerged, thereby enhancing LLMs' ability to interpret and understand visual inputs.

To advance the capabilities of Multimodal Large Language Models (MLLMs) in complex 3D environments, several studies have made significant strides by using point clouds as input tokens. Some research efforts [29, 30, 38, 46] have primarily focused on 3D object-level understanding. In addition, 3D-LLM [15] aggregates multi-view features to enrich 3D feature comprehension and employs an LLM for subsequent 3D reasoning. LL3DA [8] directly encodes 3D point clouds for scene representation, facilitating human interaction to enhance understanding.

These methods integrate large language models (LLMs) with point cloud inputs to enhance 3D reasoning capabilities; however, they face certain limitations. First, their outputs are restricted to textual or numerical forms, which are insufficient for predicting dense data types such as segmentation masks. Second, these models struggle to locate or identify objects in 3D scenes based on complex or abstract concepts, as they primarily rely on spatial relationships within the scenes.

To enable 3D-based LLM models to produce segmentation masks, we can direct the LLM to generate a `[SEG]` token. The embedding of this token is then utilized to guide the decoder in learning to predict 3D segmentation masks, similar to approaches used in 2D reasoning models [21]. However, unlike structured image data, this straightforward adaptation may encounter difficulties due to the inherent sparsity and unstructured nature of point clouds. These challenges are particularly pronounced when segmenting small objects within large-scale 3D scenes.

This paper introduces Reason3D, a framework that enables reasoning and searching within 3D scenes using large language models with only point cloud inputs. Unlike other methods that generate only textual and numerical outputs, Reason3D also produces 3D segmentation masks from textual queries. We first group point features into superpoints to reduce complexity and then utilize a transformer to align point features with textual instructions. These aligned features, along with query tokens, serve as the input for a pre-trained LLM. To address the challenges of segmenting small objects within extensive point clouds, such as searching for a ball in a large house, we develop a hierarchical mask decoder that employs a coarse-to-fine strategy. This strategy begins by instructing the LLM to output the `[LOC]` and `[SEG]` token embeddings. The `[LOC]` token guides the learning of a region mask to identify likely object-containing areas. This region mask then serves as a prior, along with the segmentation token `[SEG]`, for generating a precise object mask, facilitating effective localization in complex 3D environments.

Figure 1 illustrates Reason3D's capability to handle diverse tasks, such as reasoning, searching, referring, and question answering. To validate the effectiveness of our approach, we collect a dataset for 3D reasoning segmentation, comprising over one thousand point-instruction pairs. These pairs are annotated from Matterport3D [5] and ScanNetv2 [11] with implicit text queries that demand complex reasoning knowledge.

The main contributions of this work are:

- We introduce Reason3D, a comprehensive framework for reasoning and searching within 3D scenes using extensive language prompts. Reason3D processes 3D point clouds and language inputs to generate both textual outputs and detailed 3D segmentation masks. It supports a wide range of tasks, including expressive 3D referring segmentation, 3D reasoning segmentation, hierarchical 3D searching, and 3D question answering.
- We establish the novel task of 3D reasoning segmentation, which involves interpreting implicit human instructions within 3D scenes, and we have built a dataset to evaluate this task.
- We develop a hierarchical mask decoder to effectively address the challenges posed by the sparsity and extensive scale of 3D point clouds. This approach first identifies a coarse region likely containing the object and uses this region's probability as a prior to guide the refinement of the final mask prediction.

2. Related Work

3D Point Cloud Segmentation. Recent advancements in point cloud segmentation [20, 35, 36, 41, 45] have led to improved class-aware prediction techniques, predominantly employing UNet-like models that process data as either 3D points or voxels. Point-based methods [7, 51] enhance features with aggregation mechanisms or transformers, while voxel-based methods [9, 16] transform irregular point clouds into regular voxel grids for process-

ing with 3D convolutional networks. Another line of work [14, 26, 27, 37] focuses on understanding 3D scenes with open-vocabulary inquiry. As 3D segmentation tasks mature, developing more advanced interactions with these systems using complex instruction has become essential.

The 3D express referring segmentation task [1, 6] enhances interaction through human language by segmenting 3D objects based on specific textual descriptions. TGNN [17] uses a two-stage approach to integrate instance and textual features, computing a matching score to identify targets. X-RefSeg3D [31] combines linguistic and visual features to create a cross-modal scene graph for interactions based on textual and spatial relations. Similarly, 3D-STMN [42] aligns superpoints with textual inputs to enhance multimodal representation. However, while these studies make significant strides in object identification using spatial relation cues, they do not fully explore deeper reasoning capabilities. In this work, we introduce Reason3D, a novel approach that extends beyond traditional identification to incorporate advanced reasoning with 3D segmentation models, addressing complex interactions not yet tackled by existing methodologies.

Large Language Model. Recent advancements in large language models (LLMs) [18, 19, 39, 40, 53] have showcased their broad generalization across diverse language tasks, thanks to training on extensive textual datasets. Through self-supervised learning techniques such as token prediction and masked token reconstruction, as well as further refinements via instruction tuning and specialized datasets, researchers have significantly enhanced the adaptability of these models to new tasks. Building on this, the remarkable reasoning capabilities of LLMs are increasingly applied in multimodal contexts. Modern models incorporate advanced architectures that integrate visual data [2, 3, 12, 21, 23, 24, 28, 43, 44, 47, 49, 54], utilizing mechanisms such as cross-attention and image-text feature alignment to enable comprehensive multimodal understanding. This has paved the way for models that engage in visual question answering and perform complex reasoning tasks. Notably, LISA [21] introduces a specialized segmentation token into its vocabulary, decoded to generate a segmentation mask, enabling more precise reasoning capabilities.

Recent efforts have extended large language models (LLMs) to include 3D data for understanding point clouds. Point-LLM [46] interprets object-level points using LLMs, while 3D-LLM [15] enhances understanding by integrating multi-view image features with LLMs. LL3DA [8] combines textual instructions with visual interactions to improve feature extraction for more effective instruction-following. Unlike existing methods, which are limited to bounding box-level grounding, textual responses, or lack contextual reasoning, our approach enables fine-grained segmentation of precisely searched objects within 3D data.

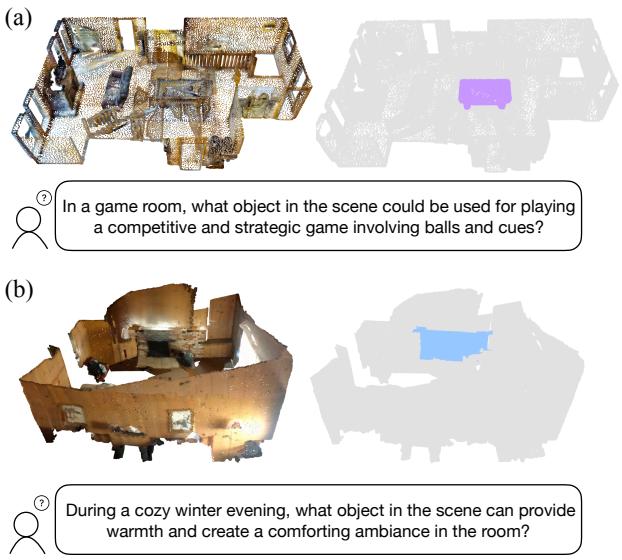

Figure 2. **Annotated Sample Examples.** (a) shows a sample from the Matterport3D dataset with the answer pool table. (b) presents a sample from the ScannetV2 dataset with the answer fireplace.

3. 3D Reasoning Segmentation

Problem Definition. 3D reasoning segmentation task involves generating a 3D segmentation map $\mathbf{M}$ from a given 3D scene point cloud $\mathbf{P}$ alongside a complex textual instruction $\mathbf{X}_{\text{txt}}$. This instruction often demands sophisticated linguistic comprehension, extending beyond mere identification tasks, like 3D referring segmentation task [17]. For instance, rather than processing simple directives like "the red chair," the textual queries might involve intricate descriptions or scenarios, such as "an object usually situated in a living room that can accommodate multiple people sitting together comfortably." which requires in-depth world knowledge and reasoning understanding.

Dataset Collection. Given the absence of a standardized dataset for evaluating 3D reasoning segmentation, we have collected the 3D scans from indoor datasets, Matterport3D [5] and ScanNetv2 [11] and annotated them with complex text instructions and detailed 3D segmentation masks. The dataset consists of 1339 samples for training and 1145 samples for validation. Two sample data are shown in Figure 2. More details can be found in the supplementary materials.

3D Hierarchical Searching. Building on the foundation of 3D reasoning segmentation, we can extend the task to include searching for an object within a specified location in a large-scale 3D scene based on an abstract query. For example, instead of merely finding an object to sit on, we can specify that the object should be inside a bedroom, thus precisely limiting the object's location.

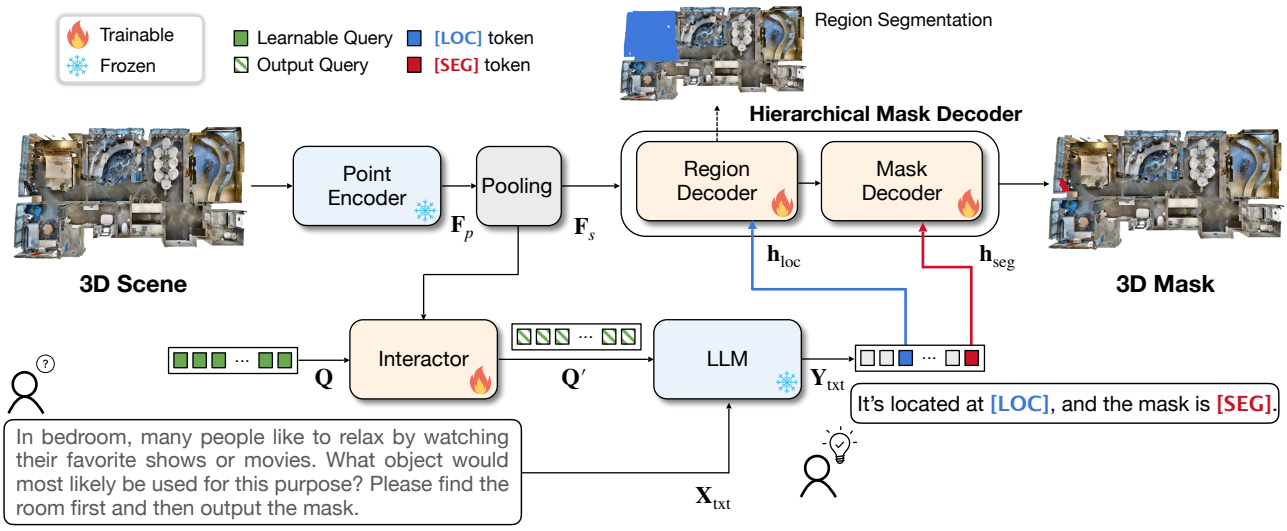

Figure 3. **Overview of our Reason3D framework.** Initially, we utilize a point encoder to extract point features from the input scene, which are simplified by a superpoint pooling layer to reduce complexity. An interactor merges these superpoint features with a learnable query, input into a frozen LLM along with instructions to generate an output containing specifical tokens, `[LOC]` and `[SEG]`. A hierarchical mask decoder then utilizes the `[LOC]` embedding to estimate a coarse location that likely covers the target object. Finally, this estimated location prior is integrated with the `[SEG]` embedding to enable the prediction of the final segmentation masks.

4. Reason3D

We introduce Reason3D, a novel LLM-based framework for searching and reasoning within 3D point clouds, as illustrated in Figure 3. Given a 3D point cloud and a textual query describing an object of interest, our method leverages an LLM model to align point features and predict dense object segmentation masks. Section 4.1 discusses the alignment of point clouds with LLMs in the feature space. Section 4.2 introduces the proposed hierarchical mask decoder, which employs a coarse-to-fine approach for generating dense segmentation masks. Finally, Section 4.3 details the training loss of our Reason3D framework.

4.1. Alignment between LLMs and Point Cloud

Given a point cloud $\mathbf{P} \in \mathbb{R}^{N\times 6}$ consisting of N points, each characterized by three colors channels (r, g, b) and three coordinates (x, y, z), we aim to extract point features and align them with decoder-only LLM models to facilitate 3D scene understanding based on textual instructions.

Scene Encoder. We employ a voxelization operation on the point cloud and utilize a U-Net style backbone [13] to extract point-wise features $\mathbf{F}_p \in \mathbb{R}^{N\times C}$, where C denotes the channel dimension. To further reduce complexity, we feed these features into a superpoint pooling layer that leverages pre-computed superpoints [22]. This layer aggregates superpoint features $\mathbf{F}_s \in \mathbb{R}^{M\times C}$ by performing average pooling on the point-wise features within each superpoint, effectively reducing the number of points from N to M, where M represents the number of superpoints.

This reduction is crucial for managing large-scale scenes without needing to divide the point cloud into smaller segments. For example, a single Matterport3D scene [5] contains approximately one million points, posing a significant challenge for existing algorithms [27], which typically require data segmentation. Our approach addresses this challenge by utilizing superpoints, enabling us to handle extensive data in a single pass.

Alignment with LLM. To align the superpoint features $\mathbf{F}_s$ with existing decoder-only LLM models, we employ an Interactor $\mathcal{F}$ following Q-Former [23] to facilitate dynamic interaction between the point cloud features $\mathbf{F}_s$ and the learnable query $\mathbf{Q}$, resulting in an output query $\mathbf{Q}' = \mathcal{F}(\mathbf{Q}, \mathbf{F}_s)$. Subsequently, the output query $\mathbf{Q}'$ and textual instructions $\mathbf{X}_{\text{txt}}$ are fed into a frozen decoder-only language model (LLM) to generate targeted responses:

$$\mathbf{Y}_{\text{txt}} = \text{LLM}(\mathbf{Q}', \mathbf{X}_{\text{txt}}). \tag{1}$$

We freeze the point cloud encoders and the LLM, allowing updates only to the interactor module. This setup focuses on learning interactions between 3D and linguistic data, enhancing the model's ability to produce accurate, contextually relevant responses to textual commands about 3D data.

4.2. Hierarchical Mask Decoder

Current LLM-based methods for 3D scene understanding [8, 15] are limited to producing textual or numerical outputs and cannot thus predict dense 3D masks. To overcome these limitations, we propose a Hierarchical Mask

Decoder (HMD) that utilizes a coarse-to-fine approach to predict segmentation tasks guided by the output of the LLM. Specifically, we first utilize a location prompt $\mathbf{P}_{\text{loc}}$ to learn the coarse location of the mask that potentially covers the target object. This coarse location then serves as a prior for learning the object segmentation mask, guided by another segmentation prompt $\mathbf{P}_{\text{seg}}$. We will provide a more detailed explanation later.

The Hierarchical Mask Decoder predicts the segmentation masks $\mathbf{M}_{\text{seg}}$ by utilizing the superpoint features $\mathbf{F}_s$ and two specific prompts $\langle \mathbf{P}_{\text{loc}}, \mathbf{P}_{\text{seg}} \rangle$, which is based on the instruction $\mathbf{X}_{\text{txt}}$:

$$\mathbf{M}_{\text{seg}} = \text{HMDecoder}(\mathbf{F}_s; \langle \mathbf{P}_{\text{loc}}, \mathbf{P}_{\text{seg}} \rangle | \langle \mathbf{F}_s, \mathbf{X}_{\text{txt}} \rangle). \quad (2)$$

To generate these two prompt features for guidance, we direct the LLM to output specific embedding tokens inspired by LISA [21]. Considering the segmentation prompt, the LLM is instructed to generate a `[SEG]` token. The last-layer embedding $\mathbf{h}_{\text{seg}}$ associated with the `[SEG]` token is transformed through an MLP projection layer $\mathcal{G}$, resulting in the segmentation prompt $\mathbf{P}_{\text{seg}} = \mathcal{G}(\mathbf{h}_{\text{seg}})$. This prompt encapsulates the features of the target object, derived from the textual instructions, to guide the final mask prediction.

Moreover, to effectively target small objects within large scenes, we introduce a location token, `[LOC]`, which the LLM is instructed to generate. This token learns a coarse location that may potentially encompass the object mask, serving as a prior feature to enhance the accuracy of the final segmentation results. Similar to the `[SEG]` token process, we refine the embedding $\mathbf{h}_{\text{loc}}$ of the `[LOC]` token using an MLP layer $\mathcal{G}$, resulting in the location prompts $\mathbf{P}_{\text{loc}} = \mathcal{G}(\mathbf{h}_{\text{loc}})$. To this end, we can use these two prompts to guide the final mask prediction. In practice, we first exploit a region decoder, $\mathcal{F}_{\text{loc}}$, built on a transformer decoder architecture [35, 36]. The location prompt $\mathbf{P}_{\text{loc}}$ serves as the query, while the superpoint features $\mathbf{F_s}$ act as key and value to generate the location mask $\mathbf{M}_{\text{loc}}$:

$$\mathbf{M}_{\text{loc}} = \mathcal{F}_{\text{loc}}(\mathbf{P}_{\text{loc}}, \mathbf{F_s}), \quad (3)$$

where the location mask indicates the probability of the region potentially covering the target object.

After that, we use an MLP layer $\mathcal{H}_{\text{loc}}$ to encode the location mask $\mathbf{M}_{\text{loc}}$, serving as a feature prior, which is then combined with the point features. Subsequently, we make the final mask generation based on the segmentation prompt and the integrated point features:

$$\mathbf{M}_{\text{seg}} = \mathcal{F}_{\text{seg}}(\mathbf{P}_{\text{seg}}, \mathbf{F_s} + \mathcal{H}_{\text{loc}}(\mathbf{M}_{\text{loc}})), \quad (4)$$

where $\mathbf{M}_{\text{seg}}$ is the final mask, and $\mathcal{F}_{\text{seg}}$ shares the same architectural framework as $\mathcal{F}_{\text{loc}}$.

4.3. Training Reason3D

The loss function for Reason3D comprises two essential components: the LLM loss $\mathcal{L}_{\text{llm}}$ and the segmentation mask loss $\mathcal{L}_{\text{mask}}$. The overall combination is represented as:

$$\mathcal{L} = \mathcal{L}_{\text{llm}} + \mathcal{L}_{\text{mask}}. \quad (5)$$

In particular, The LLM loss, $\mathcal{L}_{\text{llm}}$, embodies the linguistic aspects through an auto-regressive cross-entropy loss for text generation, incorporating cross-entropy loss CE for each token:

$$\mathcal{L}_{\text{llm}} = \text{CE}(\mathbf{Y}_{\text{txt}}, \hat{\mathbf{Y}}_{\text{txt}}) \quad (6)$$

where $\hat{\mathbf{Y}}_{\text{txt}}$ represents the ground truth word token. In addition, the mask loss $\mathcal{L}_{\text{mask}}$ aims at encouraging the model to generate high-quality segmentation masks. This loss is computed using a binary cross-entropy (BCE) loss and DICE loss for all superpoints, which is represented as:

$$\mathcal{L}_{\text{mask}_*} = \text{BCE}(\mathbf{M}_*, \hat{\mathbf{M}}_*) + \text{DICE}(\mathbf{M}_*, \hat{\mathbf{M}}_*), * \in [\text{loc}, \text{seg}]. \quad (7)$$

where $\hat{\mathbf{M}}_*$ means the ground truth segmentation mask for region-level and object-level superpoints. For the object-level mask $\hat{\mathbf{M}}_{\text{seg}}$, we use the mask corresponding to the specific object we are targeting. For the region-level mask $\hat{\mathbf{M}}_{\text{loc}}$, we designate the points as foreground points if the distance between any point and the object's center is smaller than threshold τ or we select the points within the specific room for hierarchical searching task.

5. Experiments

5.1. Experimental Setting

Datasets. Our training data includes three main types of datasets: (1) For the 3D expressive referring segmentation task, we use ScanRefer [6] and Sr3D datasets [1]. (2) For the 3D question answering task, we utilize ScanQA dataset [4]. (3) For the 3D reasoning segmentation task, we construct the Reason3D dataset from ScanNetV2 and Matterport3D datasets. The results of 3D QA and more details are included in the supplementary materials.

Model Architecture. We use a pre-trained Sparse 3D U-Net [36] to extract point-wise features. For the language learning model, we employ FlanT5 [10], maintaining most of its pre-trained weights frozen, except for adapting the weights for the newly-added location and segmentation tokens. Our Interactor is constructed following BLIP-2 [23], incorporating 1408-dimensional features.

Evaluation Metrics. For the 3D expressive referring segmentation and 3D reasoning segmentation tasks, the primary evaluation metrics are Mean Intersection over Union

Method	Venue	ScanNet			Matterport3D		
		Acc@0.25	Acc@0.50	mIoU	Acc@0.25	Acc@0.50	mIoU
OpenScene [27]	CVPR'23	4.22	0.97	5.03	4.07	0.57	6.36
OpenScene [27]+FlanT5 [10]	CVPR'23+ArXiv'22	24.68	7.14	15.03	19.98	4.02	13.60
OpenMask3D [37]	NeurIPS'23	5.70	3.25	7.14	3.25	0.12	5.96
OpenMask3D [37]+FlanT5 [10]	NeurIPS'23+ArXiv'22	20.78	6.82	13.38	17.46	0.23	9.07
3D-STMN [42]	AAAI'24	25.43	17.78	18.23	20.68	10.81	13.47
Llama2 [18]+CLIP [32]	ArXiv'23+ArXiv'22	39.26	25.93	27.23	28.51	14.86	17.80
Reason3D (Ours)	-	**43.21**	**32.10**	**31.20**	**31.22**	**17.43**	**19.54**

Table 1. **3D Reasoning Segmentation Results.** The evaluation metric is accuracy at IoU 0.25, IoU 0.5 and mIoU.

	Method	Room Num = 1∼2			Room Num = 3∼4			Room Num ≥ 5		
		Acc@0.25	Acc@0.50	mIoU	Acc@0.25	Acc@0.50	mIoU	Acc@0.25	Acc@0.50	mIoU
(a)	FlanT5 [10] + OpenScene [27]	17.65	3.95	12.89	11.27	1.02	7.69	6.22	0.97	2.21
(b)	Reason3D-base	25.23	10.32	15.56	12.84	5.50	8.23	8.26	2.52	5.33
(c)	Region Seg + Reason3D-base	26.98	13.21	17.02	19.21	8.21	11.67	11.96	4.78	7.21
(d)	Reason3D	**29.82**	**16.97**	**18.81**	**22.25**	**11.93**	**14.12**	**16.06**	**7.34**	**10.35**

Table 2. **3D Hierarchical Searching Results** on Matterport3D dataset with different room numbers. Reason3D-base refers to the full Reason3D model without the proposed [LOC] token and region decoder. The evaluation metric is IoU@0.25, IoU@0.5 and mIoU.

(mIoU), which quantifies the average overlap between the predicted and true 3D volumes, and Accuracy at k Intersection over Union (Acc@kIoU). This latter metric measures the proportion of descriptions for which the predicted mask overlaps the ground truth with an IoU greater than k, where k is set at thresholds of 0.25 and 0.5, thus assessing the model's performance at varying levels of precision.

5.2. 3D Reasoning Segmentation Results

Table 1 presents the results of 3D reasoning segmentation, where our model significantly outperforms previous methods, achieving a notable increase in mean Intersection over Union (mIoU). Unlike typical 3D referring segmentation tasks, this task demands not only spatial understanding but also robust reasoning and contextual comprehension.

Our model excels at interpreting long sentence queries and managing 3D reasoning segmentation tasks, outperforming open-vocabulary segmentation methods like OpenScene [27] and OpenMask3D [37], which primarily use vocabulary as the query. We also compare it to the two-stage methods, where FlanT5 [10] generates a short vocabulary output followed by segmentation with OpenScene [27] or OpenMask3D [37]. Our approach surpasses these by leveraging more expressive hidden embeddings, offering a richer representation than relying solely on text as an intermediary.

Compared to leading 3D referring segmentation models like 3D-STMN [42] fine-tuned on the Reason3D dataset, we find that while 3D-STMN excels in direct referencing, it struggles with indirect queries. In contrast, our model, with its integration of large language models, shows superior adaptability and performance in these scenarios.

We also compare our model to a two-stage method that combines Llama2 [18] and CLIP [32], both fine-tuned on the Reason3D dataset. In this approach, Llama2 [18] generates a vocabulary output, which CLIP [32] converts into textual features. These features then interact with the same point features and mask decoder used in our Reason3D to produce segmentation masks. The results show that our approach significantly outperforms this two-stage method, which is fully decoupled and relies solely on the textual outputs from the LLM.

5.3. 3D Hierarchical Searching Results

For the 3D hierarchical searching task, an extension of 3D reasoning segmentation task, our goal is to segment target objects within a larger space (*e.g.*, multiple rooms) rather than a single room. The task involves specifying a target room where the model must locate the object, such as finding the TV in the bedroom, as shown in Figure 3. Given that the Reason3D dataset mainly focuses on single-room scenarios, we extend it to multi-room settings by reusing a subset of annotated Matterport3D [5] data. We chose not to use ScanNet, as it features only single-room scenes.

Table 2 presents different results: (a) a two-stage baseline that uses an LLM [10] to parse the instruction, and applies an open-world segmentation model [27]; (b) Reason3D-base (without Region Decoder), (c) Reason3D combined with a region segmentation model that first segments the target room's region, then applies the Reason3D-base for the segmented region (without location prompt), and (d) Reason3D (full model). Reason3D (d) outperforms the two-stage baseline (a) due to its effective design. Com-

Method	Venue	Unique (∼19%)			Multiple (∼81%)			Overall		
		Acc@0.25	Acc@0.50	mIoU	Acc@0.25	Acc@0.50	mIoU	Acc@0.25	Acc@0.50	mIoU
ScanRefer [6]*	ECCV'20	67.6	44.4	39.9	31.2	20.9	19.5	38.2	25.5	23.5
3DVG-Transformer [52]*	ICCV'21	79.5	58.0	49.9	42.0	30.8	27.0	49.3	36.1	31.4
3D-SPS [25]*	CVPR'22	84.8	65.6	54.7	41.7	30.8	26.7	50.1	37.6	32.1
3D-LLM [15]*	NeurIPS'23	57.8	30.6	32.5	24.7	12.8	14.0	31.1	16.3	17.6
TGNN [17]	AAAI'21	69.3	57.8	50.7	31.2	26.6	23.6	38.6	32.7	28.8
X-RefSeg3D [31]	AAAI'24	-	-	-	-	-	-	40.3	33.8	29.9
3D-STMN [42]	AAAI'24	**89.3**	84.0	74.5	46.2	29.2	31.1	54.6	39.8	39.5
Reason3D (Ours)	-	88.4	**84.2**	**74.6**	**50.5**	**31.7**	**34.1**	**57.9**	**41.9**	**42.0**

Table 3. **3D Referring Expression Segmentation Results** on ScanRefer dataset with the accuracy evaluated by IoU 0.25, IoU 0.5 and mIoU. For the first block methods * that only output 3D bounding boxes, we reproduce the results based on their official codes by extracting the points inside the boxes as the segmentation mask predictions.

Ablation	Acc@0.25	Acc@0.50
(a) w/o region sup. ($\mathcal{L}_{\text{mask}_{\text{loc}}}$)	14.27	6.33
(b) hard thresholding for $\mathbf{M}_{\text{loc}}$	20.35	9.98
(c) Eq. 4 $\rightarrow$ concat. $\mathcal{F}_{\text{seg}}$ and $\mathcal{H}_{\text{loc}}(\mathbf{M}_{\text{loc}})$	21.53	10.98
(d) Reason3D (full model)	22.25	11.93

Table 4. **The effect of different designs in Hierarchical Mask Decoder (HMD)** on Matterport dataset. We use room number = 3∼4 as the main experiments.

paring (b) and (d) shows that the Hierarchical Mask Decoder (HMD) significantly boosts performance by using an additional token to guide coarse region learning, effectively managing segmentation masks, especially as point cloud complexity increases. Additionally, baseline (c) improves upon (b) but faces optimization challenges in the two-stage training pipeline compared to the full model (d).

5.4. 3D Referring Expression Segmentation Results

To demonstrate the effectiveness of our model in the 3D express referring segmentation task, we compare Reason3D against state-of-the-art methods on the ScanRefer validation set, as shown in Table 3. Our approach significantly outperforms 3D-STMN [42] in overall performance. Given the limited focus on 3D referring segmentation in the literature, we also compare our model to several 3D grounding approaches that predict only 3D bounding boxes, which can generate segmentation masks by extracting points within the predicted boxes. Notably, our approach vastly outperforms the LLM-based method, 3D-LLM [15], which struggles with accurately locating 3D boxes and effectively extracting segmentation masks.

5.5. Ablation Study

Effectiveness of different design in HMD. Besides the Reason3D-base (without the proposed coarse-to-fine approach) in Table 2, we also conduct an ablation study on the Matterport3D dataset for the 3D hierarchical searching task to validate the impact of various designs in our Hierarchical Mask Decoder (HMD), as shown in Table 4. When the region supervision term ($\mathcal{L}_{\text{mask}_{\text{loc}}}$) is removed (a), the model's performance drops significantly, achieving only 14.27% accuracy at a 0.25 IoU, which indicates that region supervision plays a crucial role in guiding the HMD decoder to learn the priors for segmentation predictions. Comparing (b) and (d), we observe that hard thresholding performs worse than probability-based region mask prediction. Hard thresholding discards uncertainty, while using probability as a prior retains valuable information, enabling more nuanced and accurate segmentation decisions. Furthermore, comparing (c) and (d) suggests that summation is more effective than concatenation for combining point features with learned location priors in Equation 4, likely due to its more integrated and simpler feature representation.

Effectiveness of superpoints. Table 5(a) proves that the superpoints pooling operation is essential for our pipeline since it helps to reduce the training complexity and enables the effective training of the pipeline. Also, the average pooling for the superpoints can achieve better performance.

Effectiveness of different segmentation loss. Table 5(b) presents the performance impact of various components of segmentation loss. Using either Binary Cross-Entropy (BCE) loss or Dice loss alone leads to significantly reduced performance. In contrast, combining Dice loss and BCE loss results in the most favorable outcomes.

Effectiveness of different decoder layers. Table 5(c) presents the impact of different numbers of decoder layers. We use six layers for the decoder as the default number.

5.6. Visualization Results

Figure 4 displays the visualization results of our Reason3D model for the 3D reasoning segmentation task, highlighting our model's proficiency in accurately generating segmentation masks based on the query. Additional visualization results are included in the supplementary materials.

Superpoint	Pool	Acc@0.25	Time (ms)
✗	-	37.55	486.1
✓	max	42.97	271.3
✓	Avg	43.21	268.5

(a) Superpoint Pooling.

DICE	BCE	Acc@0.25	Acc@0.50
✓		41.73	31.36
	✓	30.86	22.47
✓	✓	43.21	32.10

(b) Segmentation Loss.

Layer	Acc@0.25	Acc@0.50
1	40.25	27.65
3	42.78	30.62
6	43.21	32.10

(c) Layer of Decoders.

Table 5. **Ablation experiments** for different design on the scannetV2 dataset for the 3D reasoning segmentation task.

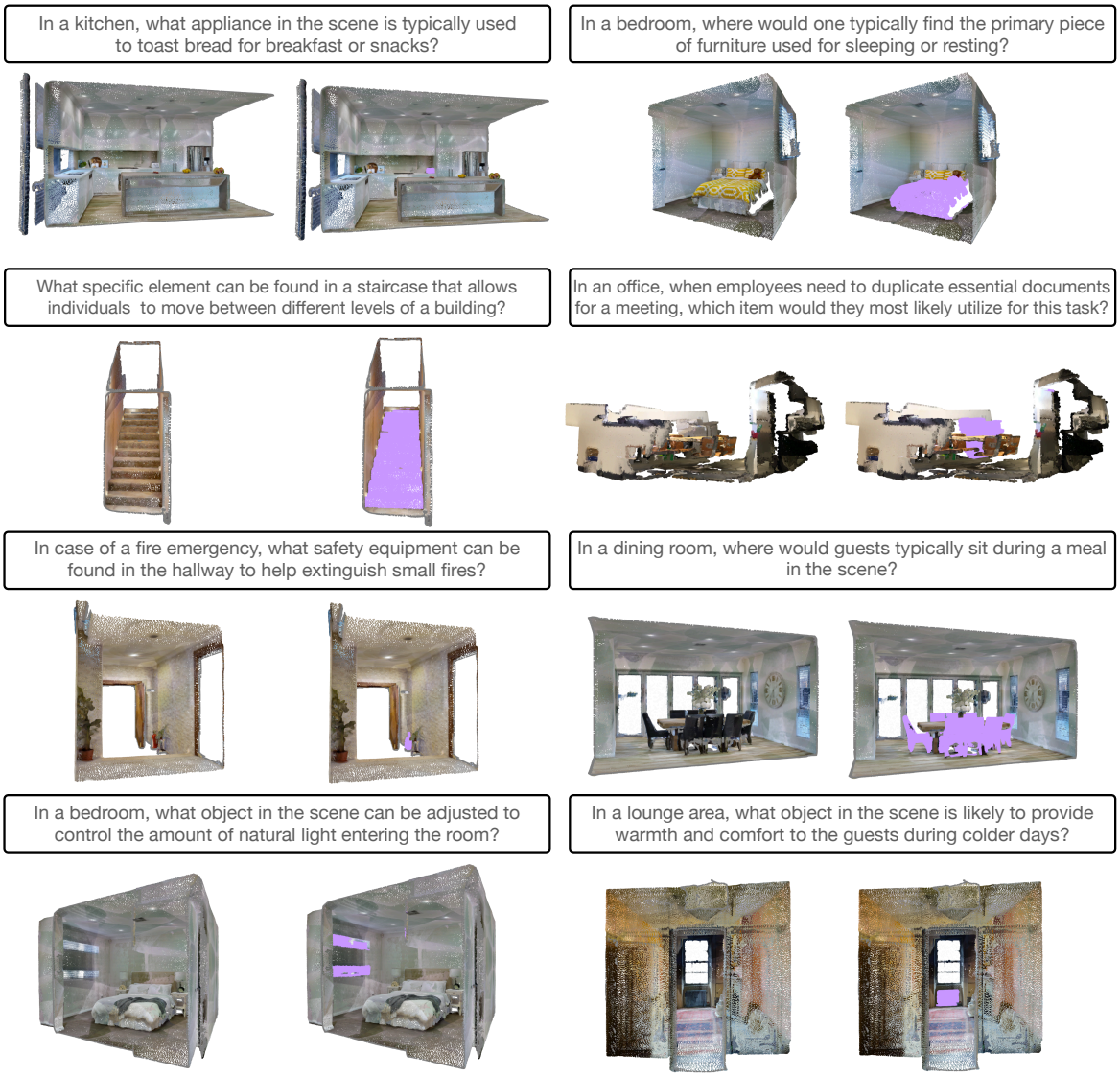

Figure 4. **Visualization Results for 3D Reasoning Segmentation Tasks.** Each sub-figure presents a textual query alongside the input point cloud. The purple regions highlight the predicted segmentation masks generated by our model.

6. Conclusion

This paper presents Reason3D, a framework that leverages Large Language Models (LLMs) for enhanced scene understanding, generating textual responses and segmentation predictions. We introduce the novel task of 3D reasoning segmentation, requiring interpreting implicit human instructions within three-dimensional scenes. A hierarchical mask decoder is proposed to enhance mask prediction by first identifying a broad region likely to contain the target object, which then serves as feature priors for further refinement. Extensive experiments on the ScanNetV2 and Matterport3D datasets demonstrate outstanding performance across tasks like 3D reasoning segmentation, 3D hierarchical searching, 3D referring segmentation, and question answering.

References

[1] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *ECCV*, 2020. 3, 5

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022. 3

[3] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. 2, 3

[4] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *CVPR*, 2022. 5

[5] 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. In *3DV*, 2017. 2, 3, 4, 6

[6] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *ECCV*, 2020. 3, 5, 7

[7] Shaoyu Chen, Jiemin Fang, Qian Zhang, Wenyu Liu, and Xinggang Wang. Hierarchical aggregation for 3d instance segmentation. In *ICCV*, 2021. 2

[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 *CVPR*, 2024. 2, 3, 4

[9] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *CVPR*, 2019. 2

[10] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, and et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. 5, 6

[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*, 2017. 2, 3

[12] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023. 2, 3

[13] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *CVPR*, 2018. 4

[14] Qiao Gu, Alihusein Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, Chuang Gan, Celso Miguel de Melo, Joshua B. Tenenbaum, Antonio Torralba, Florian Shkurti, and Liam Paull. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. *arXiv preprint arXiv:2309.16650*, 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. In *NeurIPS*, 2023. 2, 3, 4, 7

[16] Ji Hou, Angela Dai, and Matthias Nießner. 3d-sis: 3d semantic instance segmentation of rgb-d scans. In *CVPR*, 2019. 2

[17] Pin-Hao Huang, Han-Hung Lee, Hwann-Tzong Chen, and Tyng-Luh Liu. Text-guided graph neural networks for referring 3d instance segmentation. In *AAAI*, 2021. 3, 7

[18] Louis Martin Hugo Touvron, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288*, 2023. 2, 3, 6

[19] Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, Xian Li, Brian O'Horo, Gabriel Pereyra, Jeff Wang, Christopher Dewan, Asli Celikyilmaz, Luke Zettlemoyer, and Ves Stoyanov. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212.12017*, 2022. 3

[20] Maxim Kolodiazhnyi, Anna Vorontsova, Anton Konushin, and Danila Rukhovich. Oneformer3d: One transformer for unified point cloud segmentation. *arXiv preprint arXiv:2311.14405*, 2023. 2

[21] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *CVPR*, 2024. 2, 3, 5

[22] Loic Landrieu and Martin Simonovski. Large-scale point cloud semantic segmentation with superpoint graphs. In *CVPR*, 2018. 4

[23] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 2, 3, 4, 5

[24] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 2, 3

[25] Junyu Luo, Jiahui Fu, Xianghao Kong, Chen Gao, Haibing Ren, Hao Shen, Huaxia Xia, and Si Liu. 3d-sps: Single-stage 3d visual grounding via referred point progressive selection. *arXiv preprint arXiv:2204.06272*, 2022. 7

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

[27] Songyou Peng, Kyle Genova, Chiyu "Max" Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser.

Openscene: 3d scene understanding with open vocabularies. In *CVPR*, 2023. 3, 4, 6

[28] Lu Qi, Yi-Wen Chen, Lehan Yang, Tiancheng Shen, Xiangtai Li, Weidong Guo, Yu Xu, and Ming-Hsuan Yang. Generalizable entity grounding via assistance of large language model. *arXiv preprint arXiv:2402.02555*, 2024. 2, 3

[29] Zekun Qi, Runpei Dong, Shaochen Zhang, Haoran Geng, Chunrui Han, Zheng Ge, He Wang, Li Yi, and Kaisheng Ma. Shapellm: Universal 3d object understanding for embodied interaction. *arXiv preprint arXiv:2402.17766*, 2024. 2

[30] 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 *CVPR*, 2024. 2

[31] Zhipeng Qian, Yiwei Ma, Jiayi Ji, and Xiaoshuai Sun. X-refseg3d: Enhancing referring 3d instance segmentation via structured cross-modal graph neural networks. In *AAAI*, 2024. 3, 7

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 6

[33] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S. Khan. Glamm: Pixel grounding large multimodal model. In *CVPR*, 2024. 2

[34] Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. *arXiv preprint arXiv:2312.02228*, 2023. 2

[35] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3D: Mask Transformer for 3D Semantic Instance Segmentation. In *ICRA*, 2023. 2, 5

[36] Jiahao Sun, Chunmei Qing, Junpeng Tan, and Xiangmin Xu. Superpoint transformer for 3d scene instance segmentation. *arXiv preprint arXiv:2211.15766*, 2022. 2, 5

[37] Ayça Takmaz, Elisabetta Fedele, Robert W. Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. OpenMask3D: Open-Vocabulary 3D Instance Segmentation. In *NeurIPS*, 2023. 3, 6

[38] Yuan Tang, Xu Han, Xianzhi Li, Qiao Yu, Yixue Hao, Long Hu, and Min Chen. Minigpt-3d: Efficiently aligning 3d point clouds with large language models using 2d priors. *arXiv preprint arXiv:2405.01413*, 2024. 2

[39] OpenAI teams. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2, 3

[40] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv:2302.13971*, 2023. 2, 3

[41] Peng-Shuai Wang. Octformer: Octree-based transformers for 3D point clouds. In *SIGGRAPH*, 2023. 2

[42] Changli Wu, Yiwei Ma, Qi Chen, Haowei Wang, Gen Luo, Jiayi Ji, and Xiaoshuai Sun. 3d-stmn: Dependency-driven superpoint-text matching network for end-to-end 3d referring expression segmentation. In *AAAI*, 2024. 3, 6, 7

[43] Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. *arXiv preprint arXiv:2312.14135*, 2023. 3

[44] Tsung-Han Wu, Giscard Biamby, David Chan, Lisa Dunlap, Ritwik Gupta, Xudong Wang, Joseph E. Gonzalez, and Trevor Darrell. See, say, and segment: Teaching lmms to overcome false premises. *arXiv preprint arXiv:2312.08366*, 2023. 3

[45] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. In *NeurIPS*, 2022. 2

[46] Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. *arXiv preprint arXiv:2308.16911*, 2023. 2, 3

[47] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 2, 3

[48] Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, et al. Anygpt: Unified multimodal llm with discrete sequence modeling. *arXiv preprint arXiv:2402.12226*, 2024. 2

[49] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*, 2023. 3

[50] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023. 2

[51] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *ICCV*, 2021. 2

[52] Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3DVG-Transformer: Relation modeling for visual grounding on point clouds. In *ICCV*, 2021. 7

[53] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023. 3

[54] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 3