

LL3DA: Visual Interactive Instruction Tuning for Omni-3D Understanding, Reasoning, and Planning

Sijin Chen¹ Xin Chen^{2,*} Chi Zhang² Mingsheng Li¹ Gang YU²

Hao Fei³ Hongyuan Zhu⁴ Jiayuan Fan¹ Tao Chen^{1,†}

¹Fudan University ²Tencent PCG ³National University of Singapore

⁴Institute for Infocomm Research (I²R) & Centre for Frontier AI Research (CFAR), A*STAR, Singapore

<https://github.com/Open3DA/LL3DA>

* project lead † corresponding author

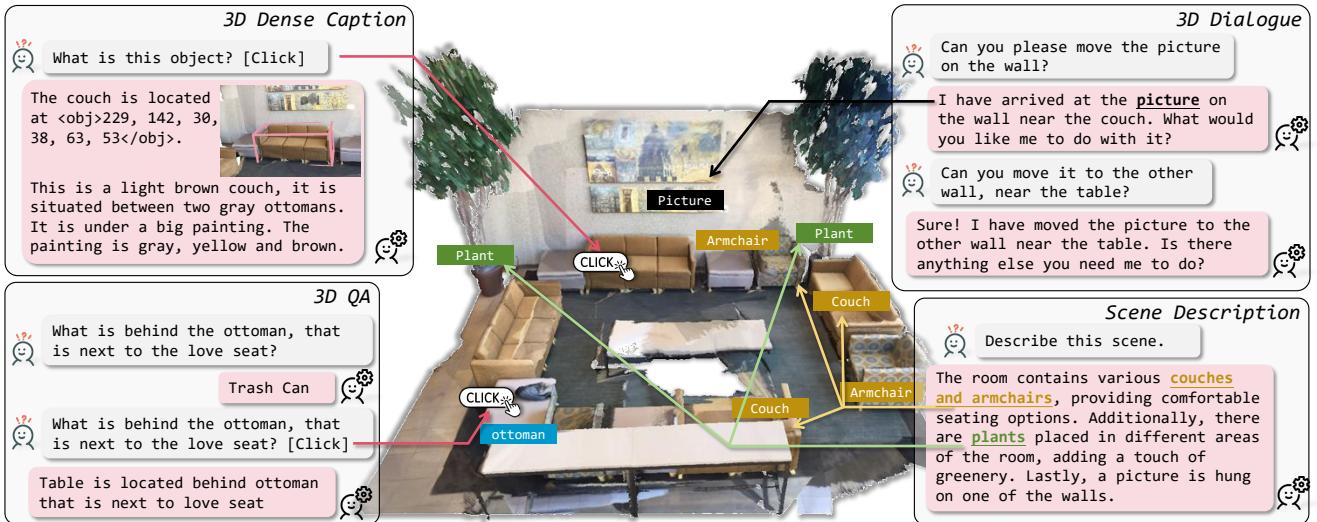


Figure 1. We propose LL3DA, a Large Language 3D Assistant that demonstrates mighty instruction-following capacities of understanding, reasoning, and planning in complex 3D environments. LL3DA takes both the textual instructions and potential visual interactions into consideration to help remove ambiguities when addressing various tasks in diverse 3D scenes.

Abstract

Recent advances in Large Multimodal Models (LMM) have made it possible for various applications in human-machine interactions. However, developing LMMs that can comprehend, reason, and plan in complex and diverse 3D environments remains a challenging topic, especially considering the demand for understanding permutation-invariant point cloud 3D representations of the 3D scene. Existing works seek help from multi-view images, and project 2D features to 3D space as 3D scene representations. This, however, leads to huge computational overhead and performance degradation. In this paper, we present LL3DA, a Large Language 3D Assistant that takes point cloud as direct input and respond to both textual-

instructions and visual-prompts. This help LMMs better comprehend human interactions and further help to remove the ambiguities in cluttered 3D scenes. Experiments show that LL3DA achieves remarkable results, and surpasses various 3D vision-language models on both 3D Dense Captioning and 3D Question Answering.

1. Introduction

The recent surge in Large Language Model (LLM) families [13, 27, 41, 49, 58] opens up great opportunities for solving various machine learning tasks in a generalized way [16, 28, 32, 34]. During this LLM carnival, researchers are also seeking generalized LLM solutions to various vision language tasks [16, 54, 59]. Among these, LLM-based 3D scene understanding is a valuable topic that would benefit

the development of autonomous driving [8, 22] and embodied AI agents [20, 47]. However, it is also challenging given 1) the diversity and complication of 3D environments and 2) the demands for understanding sparse 3D points.

Prior works have made initial success addressing various 3D vision and language tasks. The mainstream of researches build 3D specialists aiming at solving one specific down-stream task, including **3D Question Answering** (3D-QA) [2, 37], **3D Visual Grounding** (3D-VG) [6, 52], and **3D Dense Captioning** (3D-DC) [9, 11]. Meanwhile, other works [4, 12, 30, 65] study the mutual promotion of different 3D vision and language tasks with shared structure modelling relations among objects. Recently, researchers have also introduced LLMs for general purpose 3D understanding, where Point-Bind and Point-LLMs [23, 54] mainly focus on the understanding of 3D objects. Concurrently, 3D-LLM [26] proposes an LLM-driven solution that aggregates multi-view features for 3D features, presenting mighty capacities a machine could understand various 3D object and scenes, and follow textual instructions produced by human.

Though these methods have achieved remarkable success addressing different challenges in understanding 3D worlds with natural language, there are certain limitations. With limited supervision, 3D specialists could hardly scale-up for better performance, while the joint pre-training still requires separate heads for specific tasks. Extracting multi-view features results in huge computational overhead, and ignores the essential geometry properties. Additionally, plain texts often lead to ambiguities especially in cluttered and complex 3D environments.

To address the above issues, we propose the LL3DA, a **Large Language 3D Assistant** that could respond to both textual and visual interactions from human, and understand, reason, and plan in complex 3D environments (Fig. 1). We adopt a multi-modal transformer that aggregates information from textual instructions, visual prompts, and 3D scene into a fixed length of learnable querying tokens via the attention mechanism. The querying tokens are projected and used as the prefix of the textual instructions, serving as the input to a pre-trained and frozen LLM. This design not only helps to address the contradiction between the permutation-invariant 3D scene embeddings with the LLM embedding space, but also extracts interaction-aware 3D scene embeddings for efficient instruction following.

We conduct extensive experiments to explore the capacities of LLMs in understanding, reasoning, and planning within complex and diverse 3D environments. Our model achieves state-of-the-art results on two widely used datasets for 3D Dense Captioning [1, 6], and 3D Question Answering [2]. Additionally, by introducing additional visual interactions, our method could further remove the ambiguities within the vague textual instructions.

To summarize, our key contributions lie in:

- We present a LLM-based solution for understanding, reasoning, and planning in complex 3D environments.
- Our model takes both the textual instructions and visual interactions as inputs, and extracts interaction-aware features for effective instruction-following.
- Extensive experiments show that our method surpasses various state-of-the-art 3D vision language models.

2. Related Work

3D Vision and Language alignment, pre-training, and understanding [6, 19, 65] covers a bunch of tasks requiring a model to adopt its understanding towards a complex 3D scene answering to, or answering with natural language. Among those, **3D Dense Captioning** (3D-DC) [9, 11, 51] expects a model to translate an input 3D scene into a set of instance coordinates and natural language descriptions. Existing methods could be categorized into “detect-then-describe” models [4, 11, 51] and the “set-to-set” prediction approaches [9, 10]. The former builds explicit relations on the instance coordinate predictions, while the latter direct learns the location and description of instances from the input 3D scene. **3D Visual Grounding** (3D-VG) [1, 6, 52] demands a model to respond the natural language queries with the instance coordinates in the 3D scene. The mainstream of existing methods [4, 60, 65] address 3D-VG via selecting a candidate from a 3D detector’s prediction. **3D Question Answering** (3D-QA) [2, 37, 56, 61] requires a model to answer the questions with natural language based on the input 3D scene. The majority of existing methods [2, 17, 43] directly select the desired response from a given answer set. Researchers have also studied the mutual promotion of various 3D vision language tasks via training their shareable architectures simultaneously on different tasks [4, 12, 30, 65]. UniT3D [12] and 3DJCG [4] focus on the joint promotion between 3D-DC and 3D-VG in the relation modelling, while 3D-VLP [30] further includes 3D-QA. Recently, 3D-LLM [26] introduce a family of LLM-driven 3D generalists that could handle diverse textual instructions with the reconstructed 3D feature from multi-view images [25]. In this paper, we present LL3DA, an LLM solution that directly extracts features from the 3D scene, and could handle both visual prompts and textual instructions to diversify the possible interactions human could make with the complex 3D environment.

Large Multimodal Models (LMM). Along with the rapid development of **Large Language Models** (LLM) [14, 58], researchers have made great recent efforts adapting LLMs to visual understanding and reasoning tasks [53, 57]. Some project or compress global image features as prefix for instruction following [32, 34, 55, 64], while others extract ROI features as LLM tokens for region-oriented instruction reasoning [5, 59]. Meanwhile, InstructBLIP [16] proposes

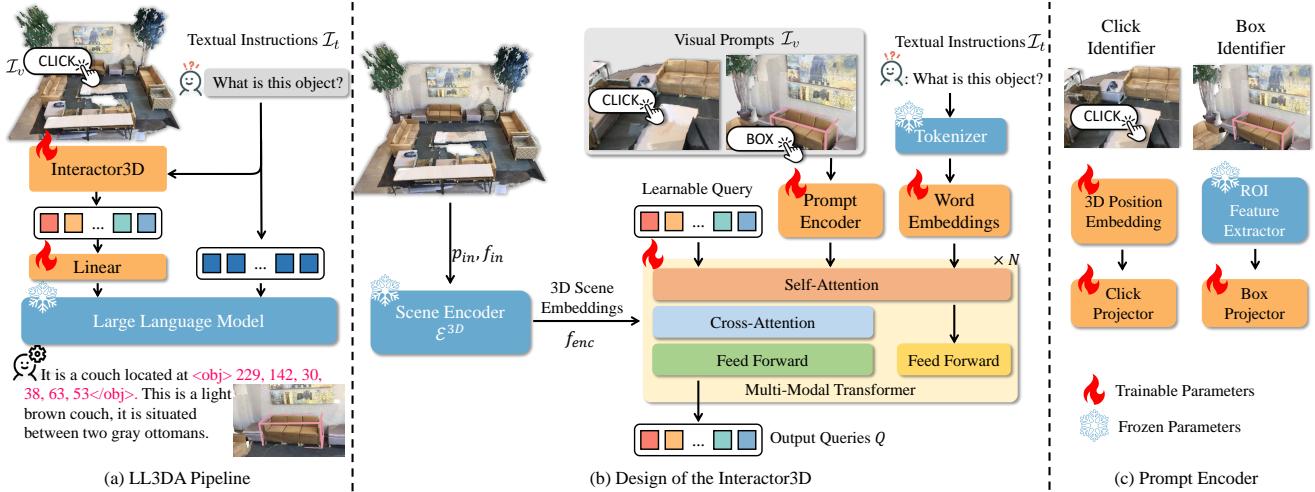


Figure 2. **Overview of the Proposed Approach.** (a) The overall pipeline of our proposed LL3DA first extracts interaction-aware 3D scene embeddings, which are later projected to the prefix of textual instructions as the input of a frozen LLM. (b) The detailed design of the Interactor3D, which aggregates visual prompts, textual instructions, and 3D scene embeddings into a fixed length querying tokens. (c) The prompt encoder encodes the user clicks and box coordinates with the positional embeddings and ROI features, respectively.

to extract textual instruction-aware visual features, and has achieved remarkable success addressing complex and unseen instructions. Concurrently, researchers have also made great attempts solving various 3D tasks using LLMs. Notably, [23, 36, 54, 63] demonstrate remarkable success in understanding and reasoning about 3D objects. In this paper, we present an LLM-driven solution that could handle both interactions in forms of visual prompts and textual instructions, and propose to extract interaction-aware 3D scene representations for better instruction following.

3. Methodology

To build a general purpose agent that could handle both visual and textual interactions in complex 3D environments, we propose LL3DA, an LLM driven auto-regressive approach to 3D vision language tasks. In this section, we first introduce the problem formatting in Sec. 3.1. Next, we introduce our model design in details (Sec. 3.2).

3.1. Problem Formatting

Model I/O. As shown in Fig. 2 (a), the **input** of our model consists of a 3D scene represented by a set of points PC , the textual instruction I_t , and potential visual interactions I_v that serve as supplementary spatial identifiers. Here, point cloud $PC = [p_{in}, f_{in}] \in \mathbb{R}^{N \times (3+F)}$, where $p_{in} \in \mathbb{R}^{N \times 3}$ and $f_{in} \in \mathbb{R}^{N \times F}$ are the point coordinates and the additional point features, including *color*, *normal*, and *height*. The **output** of our model is free-form natural language, part of whom could be interpreted into 3D coordinates.

Instruction Formatting. Following existing LMMs [54], we begin the textual instructions I_t with the “**### human.**”

identifier, and ask the model to generate responses after the “**### assistant:**” identifier. This endows the model with the ability to distinguish information from the context, and further engage in multi-turn conversations.

Coordinate Representations. To provide LLMs with the capacity to perceive and respond with 3D coordinates, we convert the 3D points and 3D bounding boxes to plain texts. Specifically, a point is represented by “**<loc>x, y, z</loc>**”, and a bounding box is represented by its center point and size, *i.e.* “**<obj>c_x, c_y, c_z, w, h, l</obj>**”. Here, all the numerical data is discretized into unsigned integers within a range of [0, 255] with respect to the boundary of the input 3D scene. This design could naturally fit in the vocabulary of existing pre-trained LLMs [49, 58]. Without introducing any additional learnable tokens, we could save the effort of tuning the whole LLM.

3.2. Model Design

As shown in Fig. 2 (a), our model first aggregates a fixed-length scene embeddings through the Interactor3D, which takes the visual prompts, the textual instructions, and the 3D scene as input. Next, the aggregated scene embeddings are projected to the prefix of textual instructions as inputs of a frozen LLM. The detailed design of Interactor3D is shown in Fig. 2 (b), which consists of a frozen 3D scene encoder E^{3D} , a visual prompt encoder, and a Q-Former to transform the permutation-invariant 3D embeddings into a fixed-length interaction-aware scene embedding, serving as the prefix of the LLM’s input.

Scene Encoder. We adopt the masked transformer encoder pre-trained on ScanNet detection [9] as the scene encoder,

\mathcal{E}^{3D} , which takes PC as its input, and outputs the 3D scene embeddings:

$$f_{enc} = \mathcal{E}^{3D}(PC) = \mathcal{E}^{3D}(p_{in}; f_{in}) \in \mathbb{R}^{M \times d}. \quad (1)$$

Here, f_{enc} consists of d -dimensioned features for M points uniformly down-sampled from the input 3D scene through the Farthest Point Sampling (FPS) algorithm. In practice, we choose to keep the scene encoder frozen to save the memory cost during training.

Visual Prompt Encoder. We mainly take two common types of visual interactions into consideration, user clicks and 3D box annotations [31]. Each user click is first normalized within a range of $[0, 1]$ by the size of the input 3D scene $p_{click} \in \mathbb{R}^3$. Then, we encode p_{click} with the 3D Fourier positional embeddings [48] function:

$$\text{pos}(p_{click}) = [\sin(2\pi p_{click} \cdot B); \cos(2\pi p_{click} \cdot B)]. \quad (2)$$

Here, $B \in \mathbb{R}^{3 \times (d/2)}$ is a learnable matrix. The box annotation is represented by the ROI feature $f_{box} \in \mathbb{R}^d$ extracted by a pre-trained 3D object detector [9]. The two types of the visual prompts are then projected with separate and identical Feed Forward Networks (FFN).

$$\begin{aligned} f_{click} &= FFN_{click}(\text{pos}(p_{click})) \\ f_{box} &= FFN_{box}(f_{box}) \end{aligned} \quad (3)$$

In practice, we represent each visual prompt with 8 tokens.

Multi-Modal Transformer(MMT) serves as a role to 1) address the contradiction between permutation-invariant 3D scene embeddings and the position-sensitive causal LLMs, 2) bridge the gap between frozen unimodal experts, and 3) fill the needs for interaction-aware feature extraction. Inspired by the Q-Former architecture [16, 32], MMT aggregates the visual information within a fixed number of 32 learnable querying tokens. In each layer, the queries interact with the encoded visual prompts $[f_{click}; f_{box}]$ and the textual instructions \mathcal{I}_t through a shared self-attention. Then, we allow the learnable querying tokens and the visual prompts to interact with the task-agnostic 3D scene embeddings f_{enc} via cross-attention. The output of MMT is 32 queries noted as $Q \in \mathbb{R}^{32 \times 768}$, and are finally projected to the embedding space of LLM with a simple linear projector. In practice, we notice that initializing Q-Former with pre-trained BERT [18, 32] weights will lead to repetitive outputs, thus we only choose to initialize the pre-trained word and position embeddings from BERT.

LLM. We consider the decoder-only generative pre-trained transformers [49, 58] as our large language model backbone, which are sensitive to the input orders because of the position embeddings, and the causal attention mask. The parameters and the embedding layers of the LLM are kept

frozen to save memory cost. During inference, we generate the responses via searching for the response s^* that satisfies:

$$s^* = \arg \max_s P(s|PC, \mathcal{I}_t, \mathcal{I}_v). \quad (4)$$

In practice, we use beam search with a beam size of 4.

4. Multi-modal Instruction Tuning

A general purpose agent is meant to deal with various tasks in complex 3D scenes. Apart from introducing proper training data, it is important to direct the model to generate the desired outputs. Thus, Sec. 4.1 will first introduce how we identify each task. After that, Sec. 4.2 will present details for training objective.

4.1. Tasks and Instructions.

All the following tasks will be modelled as auto-regressive generation after the “**### assistant:**” identifier.

3D Dense Captioning requires the localization and description of instances in diverse 3D environments. We adopt either user clicks and box annotations as the visual prompt to identify the object to be described. Additionally, we design two types of textual instructions that ask the model to either “describe” or “describe and localize” the object, which diversifies the tasks, and leads to better performance.

3D Question Answering requires the model to generate response to the questions based on the global knowledge of a 3D scene. To help the model better understand the 3D environment, we also design two types of textual instructions that ask the model to either “answer” or “answer and localize the related objects”. The latter serves as an auxiliary task widely adopted in various 3D-QA methods [2, 43]. To diversify the tasks during training, we randomly adopt visual prompts to the objects mentioned in the questions.

Scene Description requires the model to translate its global knowledge of the 3D scene into natural languages, thus we simply ask the “describe” this 3D scene.

Embodied Conversation and Planning could be treated as multi-turn conversation, which we use the “**### human:**” and “**### assistant:**” identifier to distinguish the source of information as introduced in Sec. 3.1.

4.2. Instruction Following Tuning

During training, for tasks that requires additional visual interactions, *i.e.* 3D-DC and 3D-QA, we randomly select the user clicks and box annotations.

Training Objective. Our training objective is to optimize the trainable parameters θ , so that the likelihood of the target response sequence s is maximized given the input point cloud PC , and the interactions \mathcal{I}_v and \mathcal{I}_t :

$$\theta^* = \arg \max_\theta P(s|PC; \mathcal{I}_v; \mathcal{I}_t; \theta). \quad (5)$$

In practice, this is accomplished by adopting the token-wise cross-entropy loss that trains the model to predict the i th token $s_{[i]}$ given the previous $(i - 1)$ tokens, $s_{[1, \dots, i-1]}$.

$$\mathcal{L}(\theta) = -\sum_{i=1}^{|s|} \log P(s_{[i]} | PC; \mathcal{I}_v; \mathcal{I}_t; \theta; s_{[1, \dots, i-1]}). \quad (6)$$

Here, $|s|$ is the number of tokens in the desired response.

5. Experiments

To examine the capacity of our proposed model, we provide numerous evaluations. To begin with, we introduce the datasets, metrics, and implementation details (Sec. 5.1). Then, we compare our model’s capacity of understanding and reasoning in complex 3D environments with previous 3D specialists on 3D Dense Captioning and 3D Question Answering (Sec. 5.2), and conduct quantitative ablation studies on the model design and training strategy (Sec. 5.3). Finally, Sec. 5.4 showcases several qualitative results.

5.1. Datasets, Metrics and Implementation Details

Datasets. In this paper, we experiment with 3D data from ScanNet [15], a 3D dataset covering 1,201 and 312 diverse and complex indoor 3D scenes for training and validation. The language annotations used in this study are sourced from ScanRefer [6], Nr3D [1], ScanQA [2], and the ScanNet subset of 3D-LLM [26]. This combination covers a variety of tasks, including instance and scene descriptions, conversations, embodied planning and question answering. Please refer to the supplementary materials for more details on the statistics of data.

Metrics. Here, we adopt **C**, **B-4**, **M**, **R** as abbreviations for CiDER [50], BLEU-4 [42], METEOR [3], and Rouge-L [33] to evaluate the quality of the generated textual responses.

Implementation Details. Following previous works on 3D vision language tasks [9, 11], we randomly sample 40k points from each 3D scene as the 3D input. We adopt the pre-trained OPT-1.3B [58] as our causal LLM backbone, which is frozen and loaded in float16 to save memory cost. We adopt the AdamW [35] optimizer with a weight decay of 0.1 and a learning rate decaying from 10^{-4} to 10^{-6} with a cosine annealing scheduler for about 100k iterations. For all the training tasks, we train with a total batch size of 16, and evaluate our method every 4k iterations. Each training process consumes no more than eight Nvidia RTX3090 (24G) GPUs within a day.

5.2. Comparison with SoTA Specialists

We evaluate the model’s capacity to understand and reason in 3D environments via 3D-DC and 3D-QA. For each evaluation task, we fine-tune the trainable parameters in our model on each task for ~ 30 k iterations.

3D Dense Captioning demands a model to localize and describe any instance in a 3D scene. We benchmark state-of-the-art methods on the widely-used ScanRefer [6] and Nr3D [1] dataset in Tab. 1 under the $m@k$ IoU metric [11]. Here, $m \in \{C, B-4, M, R\}$, and the m score of a caption is set to 0 if the IoU between the predicted box and the object is less than the given threshold k . Following existing works [9, 11], we consider C@0.25 and C@0.5 as the main metric for ScanRefer, and C@0.5 for Nr3D. Among the listed methods, UniT3D [12], 3DJCG [4], and 3D-VLP [30] are pre-trained on multiple 3D vision and language tasks annotated on ScanNet scenes. Additionally, UniT3D [12] adopts off-the-shelf image caption models [40] and multi-view images to generate additional instance-captions for pre-training. It is worth mentioning that the results of 3D-ViSTA [65] come from their reported version that is not trained on additional 3D scenes. To evaluate our model, we adopt the box predictions produced by Vote2Cap-DETR [9] as the visual prompt. Results show that our method consistently outperforms existing methods on all both datasets. For example, our method achieves 65.19% C@0.5 on ScanRefer and 51.18% C@0.5 on Nr3D, which is (+3.38% and +7.34%) higher than the current state-of-the-art 3D vision and language model, Vote2Cap-DETR.

3D Question Answering requires a model to generate responses to the natural language queries questioning towards an 3D scene. We benchmark state-of-the-art methods on the ScanQA [2] validation set as well as two test benchmarks in Tab. 2, and consider CiDER as the main metric. The majority of the listed methods are based on classification (marked “CLS”), *i.e.*, selecting responses from a pre-defined answer set. Meanwhile, 3D-LLM [26] tries to address 3D-QA via generating texts (marked “GEN”), and we list their fine-tuned version for comparison. Results show that our method consistently outperforms existing methods on all the evaluation sets, and surpasses the generation based method, 3D-LLM, by a large margin (+7.39% CiDER score on the validation set).

5.3. Ablation Studies

In this section, we provide ablation studies on model designs and training strategies. We evaluate on ScanRefer and ScanQA to quantize the effectiveness.

Effectiveness of the Q-Former Design. We list two ways of utilizing visual prompts in Fig. 3, where (a) is our proposed method that adopts a unified transformer that aggregates information from both the textual instructions and visual prompts, and (b) is the “direct injection” version, which only extract instruction-aware 3D feature with visual prompts concatenated after the scene embeddings. We train both models from scratch and evaluate their performance on ScanRefer 3D Dense Captioning. The results (Fig. 3) show that the method we use (Fig. 3 (a)) could better capture fea-

Table 1. **Quantitative Comparisons for 3D Dense Captioning on ScanRefer[6] and Nr3D[1].** For fair comparison, we list methods that are trained under the standard per-word cross-entropy loss without additional 3D scenes. We use the box estimations from Vote2Cap-DETR to simulate the box annotations as the visual prompts. Our proposed LL3DA surpasses previous 3D specialists on both datasets.

Method	ScanRefer								Nr3D			
	C@0.25↑	B-4@0.25↑	M@0.25↑	R@0.25↑	C@0.5↑	B-4@0.5↑	M@0.5↑	R@0.5↑	C@0.5↑	B-4@0.5↑	M@0.5↑	R@0.5↑
Scan2Cap[11]	56.82	34.18	26.29	55.27	39.08	23.32	21.97	44.78	27.47	17.24	21.80	49.06
MOREI[29]	62.91	36.25	26.75	56.33	40.94	22.93	21.66	44.42	-	-	-	-
SpaCap3D[51]	-	-	-	-	44.02	25.26	22.33	45.36	33.71	19.92	22.61	50.50
REMAN[38]	62.01	36.37	26.76	56.25	45.00	26.31	22.67	46.96	34.81	20.37	23.01	50.99
D3Net[7]	-	-	-	-	46.07	30.29	24.35	51.67	33.85	20.70	23.13	53.38
Contextual[62]	-	-	-	-	46.11	25.47	22.64	45.96	35.26	20.42	22.77	50.78
Unit3D[12]	-	-	-	-	46.69	27.22	21.91	45.98	-	-	-	-
3DJCG[4]	64.70	40.17	27.66	59.23	49.48	31.03	24.22	50.80	38.06	22.82	23.77	52.99
3D-VLP[30]	70.73	41.03	28.14	59.72	54.94	32.31	24.83	51.51	-	-	-	-
3D-VisTA*[65]	-	-	-	-	61.60	34.10	26.80	55.00	-	-	-	-
Vote2Cap-DETR[9]	71.45	39.34	28.25	59.33	61.81	34.46	26.22	54.40	43.84	26.68	25.41	54.43
LL3DA (Ours)	74.17	41.41	27.76	59.53	65.19	36.79	25.97	55.06	51.18	28.75	25.91	56.61

Table 2. **Quantitative Comparisons for 3D Question Answering on ScanQA[2].** We categorize previous works into classification based (“CLS”) and generation based (“GEN”) methods. The results from 3D-LLM* come from their fine-tuned version. LL3DA out-performs previous methods on the validation set and two test sets.

Method	Answer Type	Validation				Test w/ object				Test w/o object			
		C↑	B-4↑	M↑	R↑	C↑	B-4↑	M↑	R↑	C↑	B-4↑	M↑	R↑
ScanQA[2]	CLS	64.86	10.08	13.14	33.33	67.29	12.04	13.55	34.34	60.24	10.75	12.59	31.09
Clip-Guided[43]		-	-	-	-	69.53	14.64	13.94	35.15	62.83	11.73	13.28	32.41
Multi-CLIP[17]		-	-	-	-	68.70	12.65	13.97	35.46	63.20	12.87	13.36	32.61
3D-VLP[30]		66.97	11.15	13.53	34.51	70.18	11.23	14.16	35.97	63.40	15.84	13.13	31.79
3D-VisTA[65]		-	-	-	-	68.60	10.50	13.80	35.50	55.70	8.70	11.69	29.60
3D-LLM*[26]	GEN	69.40	12.00	14.50	35.70	69.60	11.60	14.90	35.30	-	-	-	-
LL3DA (Ours)		76.79	13.53	15.88	37.31	78.16	13.97	16.38	38.15	70.29	12.19	14.85	35.17

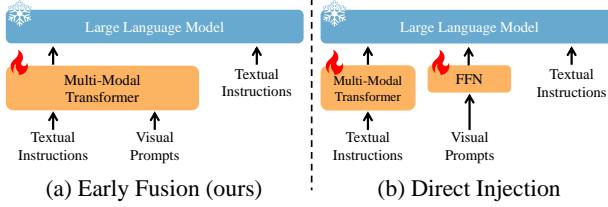


Figure 3. **Different Ways of Encoding Visual Prompts.** We listed two ways of encoding visual prompts, (a) adopting a unified transformer to aggregate features from all kinds of interactions, and (b) directly concatenate the visual prompts to the scene embeddings. Experiments (Tab. 3) show that early fusion(a) leads to a better performance.

ture related to the visual prompts, leading to better instance caption generation performance (+3.45% C@0.5).

Table 3. **Effectiveness of Q-Former Design on ScanRefer[6].** We design two different ways of utilizing visual prompts. The “early fusion” enables direct interaction with the 3D scene, thus it achieves a better performance.

Visual Prompt	C@0.5↑	B-4@0.5↑	M@0.5↑	R@0.5↑
direct	59.39	33.27	25.19	53.39
ours	62.84	35.81	25.81	54.45

Instructions as Auxiliary Tasks for 3D Dense Captioning. We have introduced two types of textual instructions in Sec. 4.1 for 3D-DC, *i.e.* the “describe” only instructions and “detect and localize” instructions. Additionally, we have introduced two types of visual prompts (Fig. 2 & Sec. 4.2). In this study, we show how they serve as auxiliary tasks for 3D-DC by evaluating on ScanRefer in Tab. 4. All the methods listed are trained from scratch. In Tab. 4, “Aux.Loc” identifies whether we train the model with the “detect and localize” instructions, and “Clicks” identifies whether we train the model with clicks as additional visual prompts. Results show that either way can serve as good auxiliary tasks for 3D Dense Captioning.

Table 4. **Effectiveness of Instructions as 3D Dense Captioning Auxiliary Tasks.** We train the models from scratch and evaluate on ScanRefer[6]. “Aux.Loc” identifies whether we train with the “describe and localize” instructions. “Clicks” identifies whether we train with clicks as additional visual prompts.

Aux.Loc	Clicks	C@0.5↑	B-4@0.5↑	M@0.5↑	R@0.5↑
-	-	60.85	34.09	25.53	53.48
✓	-	61.81	34.15	25.49	53.83
-	✓	62.20	34.26	25.67	53.87
✓	✓	62.84	35.81	25.81	54.45

Instructions as Auxiliary Tasks for 3D Question Answering.

Table 5. Evaluation as a Generalist. The first three rows list the performance of models trained from scratch as experts on each dataset. The results in the following three rows belong to the model fine-tuned from the generalist weights. The last row evaluates the model trained as a generalist. ScanRefer[6] and Nr3D[1] are used to evaluate the dense captioning performance, and ScanQA[2] is used to evaluate the question answering performance. Serving as a generalist, our method can differentiate each task, and produce strong results based on textual instructions and visual prompts.

Method	ScanRefer				Nr3D				ScanQA			
	C@0.5↑	B-4@0.5↑	M@0.5↑	R@0.5↑	C@0.5↑	B-4@0.5↑	M@0.5↑	R@0.5↑	C↑	B-4↑	M↑	R↑
ScanRefer(scratch)	62.84	35.81	25.81	54.45	-	-	-	-	-	-	-	-
Nr3D(scratch)	-	-	-	-	44.95	27.67	25.67	55.79	-	-	-	-
ScanQA(scratch)	-	-	-	-	-	-	-	-	74.80	13.68	15.40	36.25
ScanRefer(fine-tuned)	65.19	36.79	25.97	55.06	-	-	-	-	-	-	-	-
Nr3D(fine-tuned)	-	-	-	-	51.18	28.75	25.91	56.61	-	-	-	-
ScanQA(fine-tuned)	-	-	-	-	-	-	-	-	76.79	13.53	15.88	37.31
w/o fine-tuning	62.98	35.97	25.66	54.65	23.94	13.37	22.31	45.78	75.67	13.33	15.37	37.02

Table 6. Effectiveness of Interactions as 3D Question Answering Auxiliary Tasks. We train the model from scratch and evaluate all the models from scratch on ScanQA[2] validation set. “Aux.Loc” identifies whether we train with the “answer and localize” instructions, and “Visual Prompts” identifies whether we train with visual prompts.

Aux.Loc	Visual Prompts	CiDER↑	BLEU-4↑	METEOR↑	Rouge-L↑
-	-	67.85	11.87	13.96	33.87
✓	-	72.73	13.27	14.90	35.87
-	✓	68.09	12.59	14.20	33.71
✓	✓	74.80	13.68	15.40	36.25

ing. We have made similar study on the effectiveness of treating “answer and localize” instructions and additional visual prompts as auxiliary tasks for 3D-QA on ScanQA [2] validation set in Tab. 6. The listed methods are evaluated without any visual interactions for fair comparison. Results show that the additional textual instructions and visual prompts improve the task diversity and further improve the performance on 3D Question Answering.

Performance as a Generalist. To examine whether our method could distinguish different tasks given the textual instructions and visual prompts introduced in Sec. 4.1, we evaluate our model on different tasks with the same set of weights in Tab. 5 with several baseline methods. The first three rows list the performance our proposed models trained on each dataset from scratch as generalists. The following three rows list the performance of the model fine-tuned on each dataset. The last row lists the performance of our model as a generalist. Results show that our model could distinguish 3D-DC and 3D-QA given the textual instructions and visual prompts, providing strong results (62.98% C@0.5 on ScanRefer, 75.67% CiDER on ScanQA). However, the generalist model achieves poor results on Nr3D [1], which is because we did not try to differentiate between Nr3D and ScanRefer during training as these two datasets are used for the same task in the first place. There is also an interesting observation that although we did not differentiate between the two datasets for 3D-DC, and the training

sample sizes of the two models were similar, the model still tend to achieve high scores on ScanRefer (62.98% C@0.5). We are also excited to see that the weights of the generalist model could serve as a strong initialization of weights for fine-tuning. For example, the fine-tuned model on ScanRefer could achieve 65.19% C@0.5, which is +2.35% higher than the model trained from scratch.

Importance of Textual Instructions. We further conduct study to see whether the instructions are necessary in 3D Dense Captioning in Tab. 7. The first row is our baseline method that directly generates the captions based on visual prompts without any textual instructions, and the second row is our method that is trained with textual instructions introduced in Sec. 4.1. Both methods are trained from scratch for fair comparison. We notice that since the LLM is frozen, certain textual instructions are beneficial to generating results in specific domains/tasks.

Table 7. Effectiveness of Instructions on 3D Dense Captioning. We perform experiments on ScanRefer[6]. The baseline method directly generates the captions given the input 3D scene and visual prompts without any textual instructions.

Instructions	C@0.5↑	B-4@0.5↑	M@0.5↑	R@0.5↑
-	60.20	34.79	25.40	54.03
✓	62.84	35.81	25.81	54.45

Clicks for Better Question Answering. One major challenge of answering questions in complex 3D environments is the vague identification of objects with plain texts. Therefore, we try to click on some of the related objects along with the textual instructions during evaluation, and see how it could affect the generated answers on the ScanQA validation set in Tab. 8. Results show that this technique would remove the ambiguities, and further improve the quality of the answers (+6.12% C). This illustrates the importance of visual interaction in complex 3D environments.

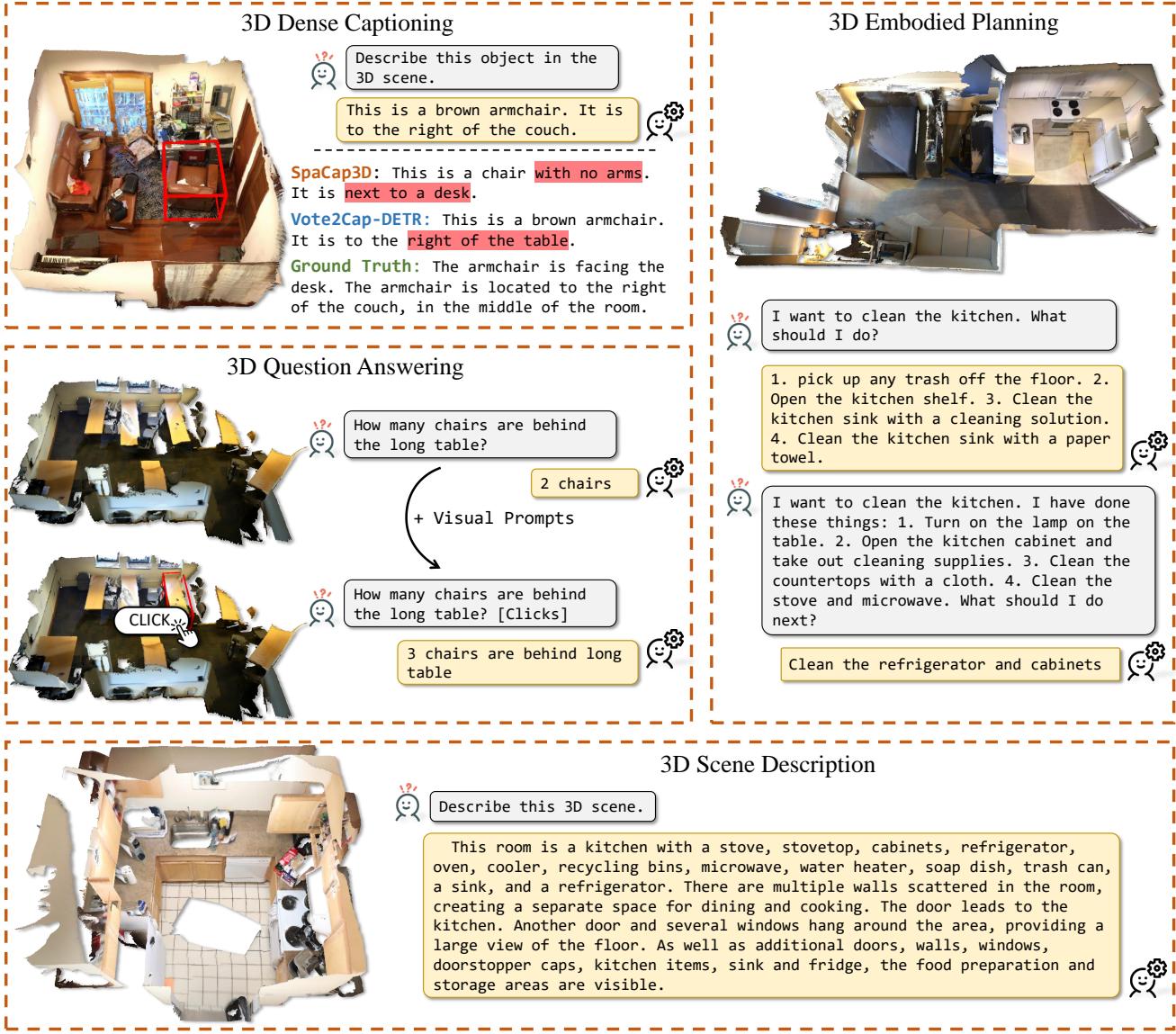


Figure 4. **Qualitative Results.** We provide several visualization results on various 3D vision and language tasks in diverse 3D environments (living room, classroom, kitchen, and bedroom). Red highlights the wrong answer.

Table 8. **Test Time Visual Interactions for Question Answering on ScanQA[2].** The model achieves better performance on the question answering when we add visual prompts to some of the related objects along with the text instructions during evaluation.

Visual Prompts	CiDER \uparrow	BLEU-4 \uparrow	METEOR \uparrow	Rouge-L \uparrow
-	76.79	13.53	15.88	37.31
✓	82.91	11.80	16.74	39.97

5.4. Qualitative Results

We present several visualization results on different tasks in Fig. 4 to show our model’s capacities in understanding, reasoning,

and planning in different 3D environments. To prevent repetition when generating long sequences, we combine the top- k [21] and top- p [24] sampling strategy with $k = 50$ and $p = 0.95$.

6. Conclusions

In this paper, we present LL3DA, a large language 3D assistant that could take both textual- and visual- interactions from human for understanding, reasoning, and planning in complex 3D environments. Our model directly encodes 3D point cloud for scene representations, and aggregates information from scenes and human interactions with the atten-

tion mechanism. The visual interactions could remove the ambiguities in cluttered 3D environments, showing mighty instruction-following capacities. Experimental results show that our method could achieve remarkable results on various 3D vision-language benchmarks. We hope that our approach could inspire further designs and training strategies for large 3D language models. In future studies, we anticipate that the availability of high-quality and diverse 3D vision and language annotations will enhance the model’s reasoning and planning capabilities.

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 *European Conference on Computer Vision*, pages 422–440. Springer, 2020. [2](#), [5](#), [6](#), [7](#), [19](#)
- [2] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19129–19139, 2022. [2](#), [4](#), [5](#), [6](#), [7](#), [8](#), [19](#)
- [3] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. [5](#), [15](#)
- [4] Daigang Cai, Lichen Zhao, Jing Zhang, Lu Sheng, and Dong Xu. 3djcg: A unified framework for joint dense captioning and visual grounding on 3d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16464–16473, 2022. [2](#), [5](#), [6](#)
- [5] Chi Chen, Ruoyu Qin, Fuwen Luo, Xiaoyue Mi, Peng Li, Maosong Sun, and Yang Liu. Position-enhanced visual instruction tuning for multimodal large language models. *arXiv preprint arXiv:2308.13437*, 2023. [2](#)
- [6] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European Conference on Computer Vision*, pages 202–221. Springer, 2020. [2](#), [5](#), [6](#), [7](#), [19](#)
- [7] Dave Zhenyu Chen, Qirui Wu, Matthias Nießner, and Angel X Chang. D3net: A speaker-listener architecture for semi-supervised dense captioning and visual grounding in rgb-d scans. *arXiv preprint arXiv:2112.01551*, 2021. [6](#)
- [8] Long Chen, Oleg Sinavski, Jan Hünermann, Alice Karnsund, Andrew James Willmott, Danny Birch, Daniel Maund, and Jamie Shotton. Driving with llms: Fusing object-level vector modality for explainable autonomous driving. *arXiv preprint arXiv:2310.01957*, 2023. [2](#)
- [9] Sijin Chen, Hongyuan Zhu, Xin Chen, Yinjie Lei, Gang Yu, and Tao Chen. End-to-end 3d dense captioning with vote2cap-detr. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11124–11133, 2023. [2](#), [3](#), [4](#), [5](#), [6](#), [17](#), [18](#)
- [10] Sijin Chen, Hongyuan Zhu, Mingsheng Li, Xin Chen, Peng Guo, Yinjie Lei, Gang Yu, Taihao Li, and Tao Chen. Vote2cap-detr++: Decoupling localization and describing for end-to-end 3d dense captioning. *arXiv preprint arXiv:2309.02999*, 2023. [2](#)
- [11] Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3193–3203, 2021. [2](#), [5](#), [6](#), [17](#)
- [12] Zhenyu Chen, Ronghang Hu, Xinlei Chen, Matthias Nießner, and Angel X Chang. Unit3d: A unified transformer for 3d dense captioning and visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18109–18119, 2023. [2](#), [5](#), [6](#)
- [13] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. [1](#)
- [14] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. [2](#)
- [15] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. [5](#), [17](#), [19](#)
- [16] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructclip: Towards general-purpose vision-language models with instruction tuning, 2023. [1](#), [2](#), [4](#)
- [17] Alexandros Delitzas, Maria Parelli, Nikolas Hars, Georgios Vlassis, Sotirios Anagnostidis, Gregor Bachmann, and Thomas Hofmann. Multi-clip: Contrastive vision-language pre-training for question answering tasks in 3d scenes. *arXiv preprint arXiv:2306.02329*, 2023. [2](#), [6](#)
- [18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. [4](#), [20](#)
- [19] Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. Pla: Language-driven open-vocabulary 3d scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7010–7019, 2023. [2](#)
- [20] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023. [2](#)
- [21] Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*, 2018. [8](#)
- [22] Daocheng Fu, Xin Li, Licheng Wen, Min Dou, Pinlong Cai, Botian Shi, and Yu Qiao. Drive like a human: Rethink-

- ing autonomous driving with large language models. *arXiv preprint arXiv:2307.07162*, 2023. 2
- [23] Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming Han, Kexin Chen, Peng Gao, Xianzhi Li, Hongsheng Li, et al. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. *arXiv preprint arXiv:2309.00615*, 2023. 2, 3
- [24] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019. 8
- [25] Yining Hong, Chunru Lin, Yilun Du, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. 3d concept learning and reasoning from multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9202–9212, 2023. 2
- [26] 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. *arXiv preprint arXiv:2307.12981*, 2023. 2, 5, 6, 15, 16, 19
- [27] Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, et al. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212.12017*, 2022. 1
- [28] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *arXiv preprint arXiv:2306.14795*, 2023. 1
- [29] Yang Jiao, Shaoxiang Chen, Zequn Jie, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. More: Multi-order relation mining for dense captioning in 3d scenes. *arXiv preprint arXiv:2203.05203*, 2022. 6
- [30] Zhao Jin, Munawar Hayat, Yuwei Yang, Yulan Guo, and Yinjie Lei. Context-aware alignment and mutual masking for 3d-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10984–10994, 2023. 2, 5, 6
- [31] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 4
- [32] 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. 1, 2, 4
- [33] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 5, 15
- [34] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 1, 2
- [35] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5
- [36] Tiange Luo, Chris Rockwell, Honglak Lee, and Justin Johnson. Scalable 3d captioning with pretrained models. *arXiv preprint arXiv:2306.07279*, 2023. 3
- [37] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. *arXiv preprint arXiv:2210.07474*, 2022. 2
- [38] Aihua Mao, Zhi Yang, Wanxin Chen, Ran Yi, and Yong-jin Liu. Complete 3d relationships extraction modality alignment network for 3d dense captioning. *IEEE Transactions on Visualization and Computer Graphics*, 2023. 6
- [39] Ishan Misra, Rohit Girdhar, and Armand Joulin. An end-to-end transformer model for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2906–2917, 2021. 20
- [40] Ron Mokady, Amir Hertz, and Amit H Bermano. Clip-cap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 5
- [41] OpenAI. Gpt-4 technical report, 2023. 1
- [42] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 5, 15
- [43] Maria Parelli, Alexandros Delitzas, Nikolas Hars, Georgios Vlassis, Sotirios Anagnostidis, Gregor Bachmann, and Thomas Hofmann. Clip-guided vision-language pre-training for question answering in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5606–5611, 2023. 2, 4, 6
- [44] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 20
- [45] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9277–9286, 2019. 17
- [46] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 17
- [47] Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2998–3009, 2023. 2
- [48] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems*, 33:7537–7547, 2020. 4
- [49] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 1, 3, 4, 16
- [50] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer*

- vision and pattern recognition*, pages 4566–4575, 2015. 5, 15
- [51] Heng Wang, Chaoyi Zhang, Jianhui Yu, and Weidong Cai. Spatiality-guided transformer for 3d dense captioning on point clouds. *arXiv preprint arXiv:2204.10688*, 2022. 2, 6
- [52] Yanmin Wu, Xinhua Cheng, Renrui Zhang, Zesen Cheng, and Jian Zhang. Eda: Explicit text-decoupling and dense alignment for 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19231–19242, 2023. 2
- [53] Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2
- [54] 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. 1, 2, 3
- [55] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 2
- [56] Shuquan Ye, Dongdong Chen, Songfang Han, and Jing Liao. 3d question answering. *IEEE Transactions on Visualization and Computer Graphics*, 2022. 2
- [57] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023. 2
- [58] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuhui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 1, 2, 3, 4, 5, 16
- [59] 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. 1, 2
- [60] Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3dvg-transformer: Relation modeling for visual grounding on point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2928–2937, 2021. 2
- [61] Lichen Zhao, Daigang Cai, Jing Zhang, Lu Sheng, Dong Xu, Rui Zheng, Jinjie Zhao, Lipeng Wang, and Xibo Fan. Towards explainable 3d grounded visual question answering: A new benchmark and strong baseline. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022. 2
- [62] Yufeng Zhong, Long Xu, Jiebo Luo, and Lin Ma. Contextual modeling for 3d dense captioning on point clouds. *arXiv preprint arXiv:2210.03925*, 2022. 6
- [63] Qiang Zhou, Chaohui Yu, Shaofeng Zhang, Sitong Wu, Zhibing Wang, and Fan Wang. Regionblip: A unified multimodal pre-training framework for holistic and regional comprehension, 2023. 3
- [64] 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. 2
- [65] Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3d-vista: Pre-trained transformer for 3d vision and text alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2911–2921, 2023. 2, 5, 6

Supplementary Material

The supplementary material consists of more visualization results (Sec. A), quantitative evaluations on scene descriptions, embodied dialogue, and embodied planning (Sec. B), additional studies on 3D coordinates generated by LL3DA (Sec. C), data statistics as well as training samples (Sec. D), additional implementation details (Sec. E).

Codes and Pre-trained weights will be fully released. We provide example code files along with this file, including the training and evaluation process for different tasks.

Video. We also provide video clips for more illustrative demonstrations.

A. More Visualizations

We present additional visualization results on Scene Descriptions (Fig. 5), 3D Dense Captioning, 3D Question Answering, Embodied Dialogue (Fig. 6), and Embodied Planning (Fig. 7).

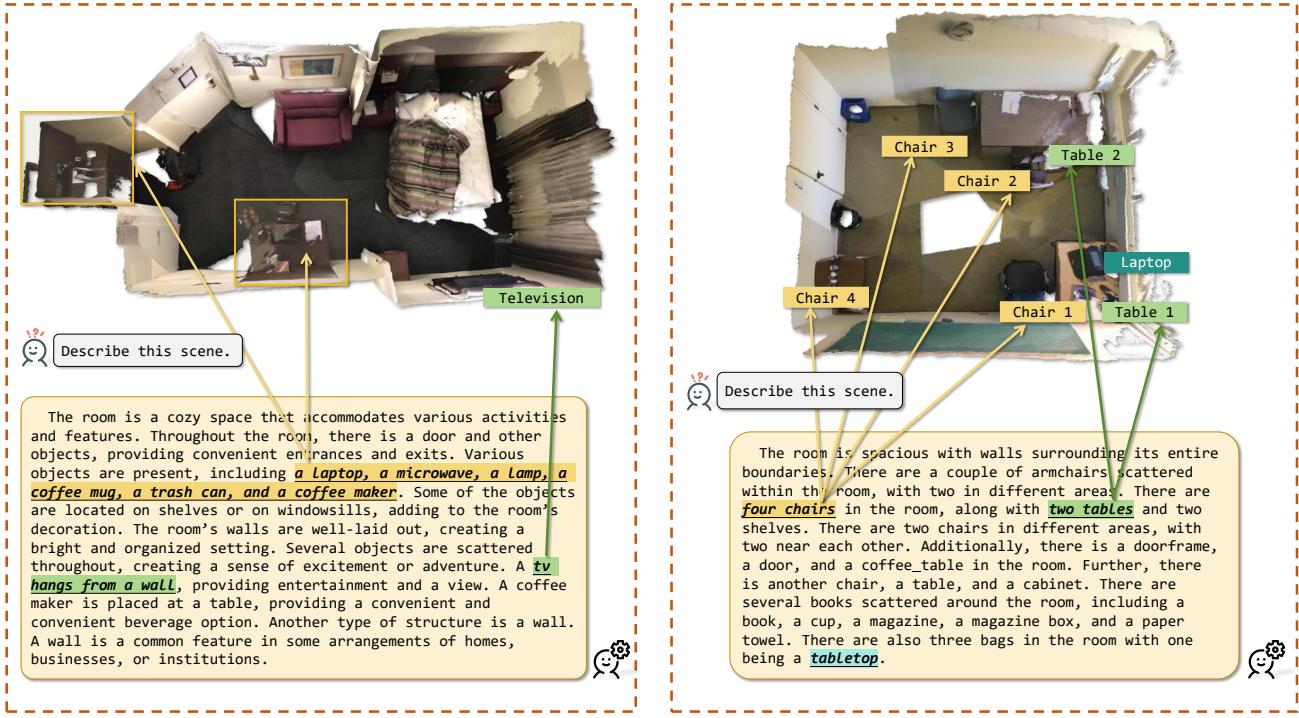


Figure 5. **Qualitative Results on Scene Descriptions.** We highlight some of the phrases in the generated scene descriptions mentioning the instances in the 3D environment.

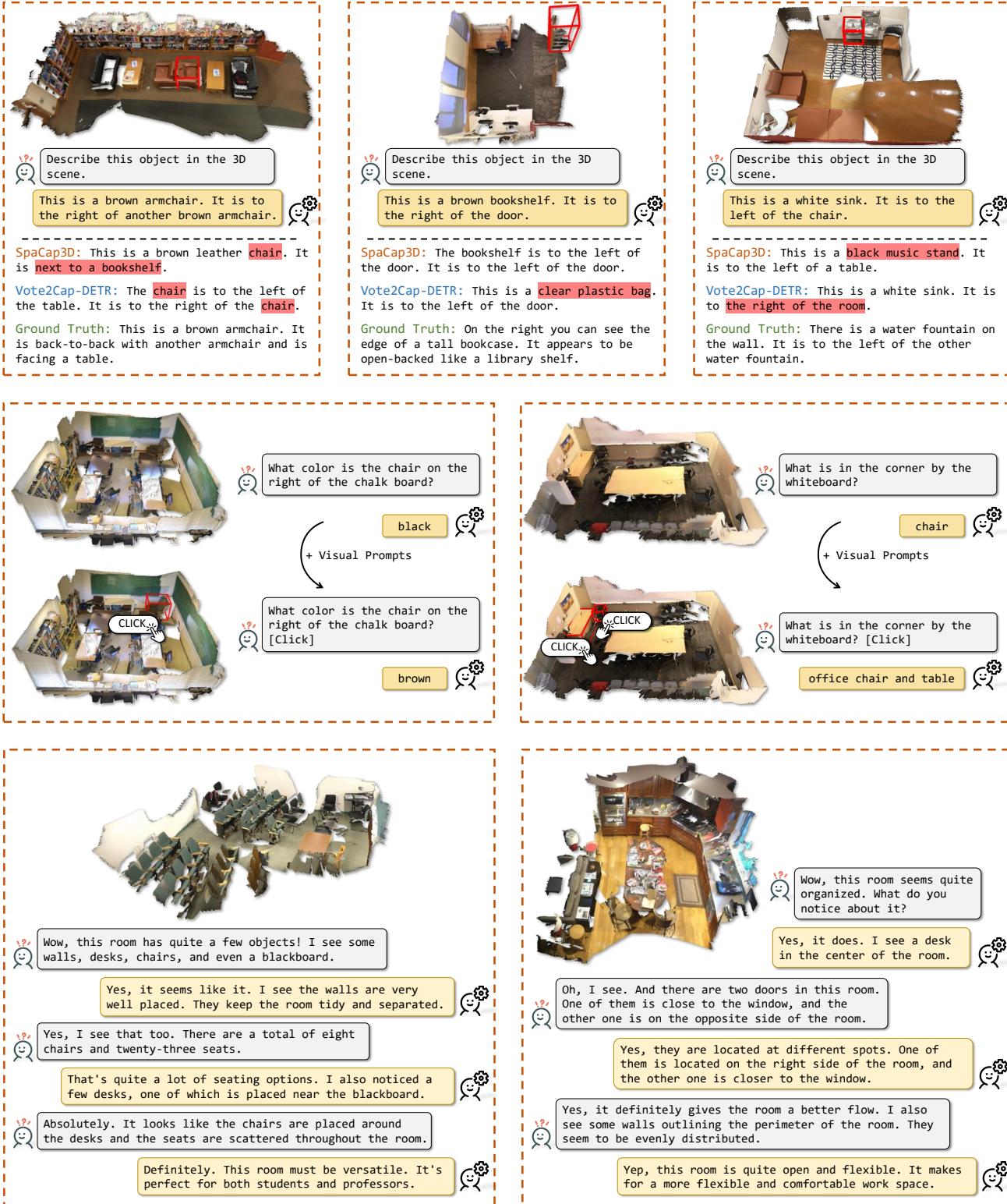


Figure 6. More Qualitative Results on 3D Dense Captioning (upper), 3D Question Answering (middle), and Embodied Dialogue (lower). Red highlights the wrong answer.



Figure 7. **Qualitative Results on Embodied Planning.** The planning results generated by LL3DA are consistent with common sense knowledge. Additionally, LL3DA could not only make plans directly, but also generate feedbacks based on the things have been done.

B. More Evaluations

Since 3D-LLM [26] has not yet released the validation set they used for evaluation, we make our own split to quantify our method’s performance on: 1) scene descriptions, 2) embodied dialogue, and 3) embodied planning.

Data Splits. In practice, we set the scenes with ids less than 600 as the training set, and the rest as the validation set. The data statistics of our split are listed in Tab. 9. It is worth mentioning that:

- The multi-turn embodied dialogues are further decomposed into 8,490 and 1,222 lines of data samples following the procedure in Fig. 8 for training and validation.
- The embodied planning data is also decomposed into 16,972 and 2,282 lines of data samples for training and validation following the procedure in Fig. 9.

As mentioned in Sec. 4.1 in the main paper, we distinguish the source of texts with the “`### human:`” and “`### assistant:`” identifier.

Metrics. Similar to the main paper, we evaluate the generated natural language responses under the CiDER [50], BLEU [42], METEOR [3], and Rouge-L [33] metrics in Tab. 10.

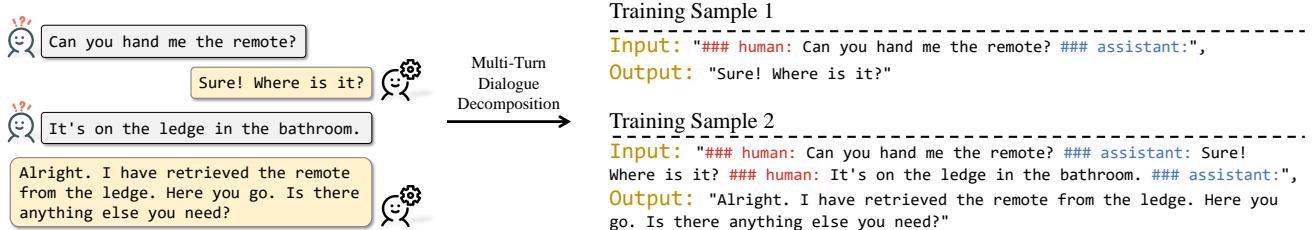


Figure 8. **The Multi-Turn Dialogue Decomposition Pipeline.** For one sample of a n -turn dialogue data ($n \geq 1$), we decompose it into n lines of training samples, and distinguish the source of data with the “`### human:`” and “`### assistant:`” identifier.

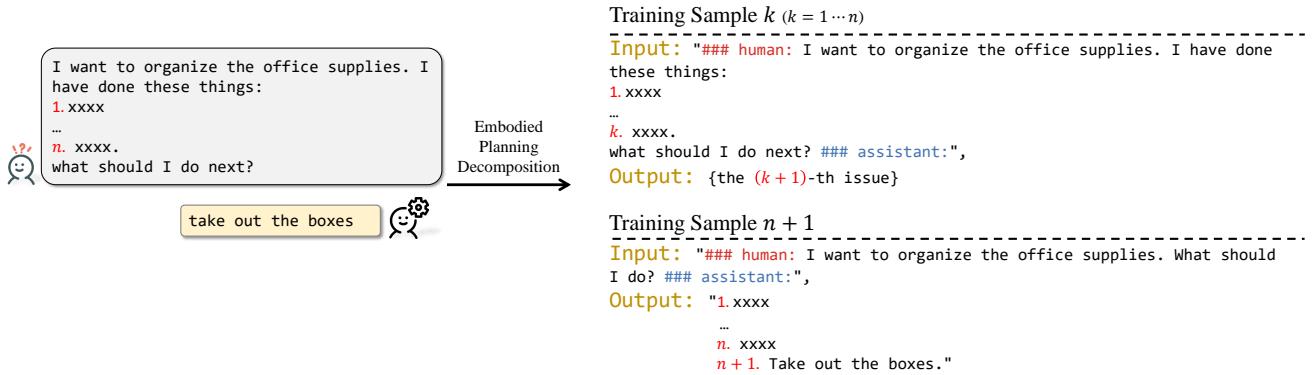


Figure 9. **The Embodied Planning Decomposition Pipeline.** For one sample with given n -tasks ($n \geq 1$), we decompose it into $(n + 1)$ lines of training and validation samples.

Table 9. **Details about the Training-Validation Split of the ScanNet Part of 3D-LLM [26].** We list the number of natural language annotations (“# annotations”) and number of scenes (“# scenes”) in each split for each task.

Split	Scene Description		Embodied Planning		Embodied Dialogue	
	# annotations	# scenes	# annotations	# scenes	# annotations	# scenes
Train	912	456	1,636	449	2,592	456
Validation	121	61	228	61	363	61
Total	1,033	517	1,864	510	2,955	517

Table 10. **Quantitative Comparisons on Scene Description, Embodied Dialogue and Embodied Planning.** We evaluate our method with different sampling strategies on the ScanNet part of 3D-LLM [26]. We manually set the scenes with ids less than 600 as the training set, and the rest as the validation set. To prevent repetition, we adopt the n -gram repetition penalty so that no n -gram appears twice ($n = 4$).

Task	Method	BLEU-1↑	BLEU-2↑	BLEU-3↑	BLEU-4↑	CiDER↑	METEOR↑	Rouge-L↑
Zero-Shot:								
Scene Description	OPT-1.3B [58]	15.79	6.10	2.07	0.84	0.00	8.40	11.70
	OPT-2.7B [58]	19.97	7.59	2.14	0.00	0.11	6.60	12.32
	OPT-6.7B [58]	24.40	9.79	3.62	1.13	0.06	8.99	16.96
	<u>LLAMA-7B [49]</u>	19.26	7.69	2.79	0.92	0.20	7.00	12.31
	Ours:							
	top- k & top- p	43.02	26.70	15.97	8.97	0.96	14.65	24.84
Embodied Dialogue	greedy decoding	29.15	20.51	13.99	9.38	1.44	12.83	24.62
	beam search	29.94	21.56	14.93	10.02	1.32	12.31	27.08
	Zero-Shot:							
	OPT-1.3B [58]	2.44	1.05	0.46	0.23	0.31	5.62	4.83
	OPT-2.7B [58]	3.88	1.56	0.73	0.39	0.38	7.38	6.28
	OPT-6.7B [58]	3.59	1.65	0.81	0.43	0.25	6.88	6.16
Embodied Planning	<u>LLAMA-7B [49]</u>	4.08	1.80	0.90	0.50	0.27	7.81	6.68
	Ours:							
	top- k & top- p	41.38	32.59	27.47	23.95	190.01	23.50	40.61
	greedy decoding	48.08	39.59	34.37	30.70	251.78	27.01	47.57
	beam search	48.14	39.83	34.83	31.32	260.07	27.21	47.69
	Zero-Shot:							
	OPT-1.3B [58]	1.26	0.59	0.26	0.13	0.16	0.24	3.56
	OPT-2.7B [58]	2.02	0.99	0.49	0.26	0.10	3.59	4.35
	OPT-6.7B [58]	2.03	1.06	0.53	0.28	0.00	3.65	3.94
	<u>LLAMA-7B [49]</u>	2.24	1.13	0.55	0.29	0.04	3.53	4.71
	Ours:							
	top- k & top- p	40.72	27.18	18.64	12.95	128.80	17.05	39.25
	greedy decoding	38.81	27.58	20.10	14.66	186.13	19.60	45.34
	beam search	45.07	33.04	24.96	19.15	196.78	19.87	45.58

Quantitative Comparisons. As shown in Tab. 10, we perform zero-shot evaluation on the frozen LLMs [49, 58] with the top- k and top- p sampling strategy as the baseline, and evaluate our method with different sampling strategies. Among all the listed results, we adopt the n -gram penalty that restricts no n -gram appears twice with $n = 4$. We also set $k = 50$ and $p = 0.95$ for top- k and top- p sampling, and a beam size = 4 for beam search. From the table, one could see that our method could generate high-quality responses, and out-performs language-only LLMs.

C. Additional Studies: Generating 3D Bounding Boxes with LLM

Quantitative Evaluation on 3D Open-Vocabulary Detection with LLM-Generated Boxes. We treat 3D Open-Vocabulary Detection as the 3D Dense Captioning problem, and simulate the click prompts using the spatial location of vote queries [9]. The textual instruction is set as “**### human: what is this object? ### assistant:**”. To obtain localization results, we reconstruct 3D bounding boxes from the “`<obj>cx, cy, cz, w, h, l</obj>`” in the generated texts, and also decode the object category prediction from the texts. In Tab. 13, we present performances under the ScanNet vocabulary [15], and the ScanNet200 vocabulary [46]. Per-class mAP results on ScanNet could be found in Tab. 12. Results show that LL3DA could produce high-quality bounding boxes comparable to 3D detectors under the threshold IoU=0.5.

Table 11. **Quantitative Comparisons on Open-Vocabulary Detection with LL3DA.** We treat 3D Detection as the 3D Dense Captioning problem. We simulate the click prompt with the vote queries [9], and decode the “`<obj>cx, cy, cz, w, h, l</obj>`” and category names from the generated texts. Results show that our method is able to produce comparable to the 3D detector specialist when IoU=0.5.

Method	ScanNet Vocabulary				ScanNet200 Vocabulary			
	IoU=0.25		IoU=0.5		IoU=0.25		IoU=0.5	
	mAP↑	AR↑	mAP↑	AR↑	mAP↑	AR↑	mAP↑	AR↑
VoteNet [45]	57.17	81.18	31.50	50.08	-	-	-	-
Ours	48.94	65.15	32.48	49.46	7.40	12.10	5.20	9.04

Table 12. **Per-class AP under IoU threshold of 0.25 and 0.5 on ScanNet validation set.**

IoU	bathtub	bed	bookshelf	cabinet	chair	counter	curtain	desk	door	others	picture	refrigerator	shower curtain	sink	sofa	table	toilet	window
0.25	83.33	62.16	10.63	43.13	80.83	55.63	40.20	42.75	37.17	27.03	17.11	24.96	62.92	56.72	64.30	52.80	87.73	31.46
0.50	68.33	53.24	6.50	19.81	72.15	23.21	23.55	25.81	18.32	13.54	6.13	18.00	35.42	18.02	51.02	45.72	76.93	8.98

Quantitative Evaluation on 3D Dense Captioning with LLM-Generated Boxes. We evaluate the quality of the 3D bounding boxes generated by LL3DA on 3D Dense Captioning. Following existing works [11], we evaluate with the $m@k$ IoU metric, where m could be one of CiDER, BLEU-4, METEOR, Rouge-L:

$$m@k\text{IoU} (b_{\text{pred}}^{3D}, c_{\text{pred}}) = \frac{1}{N} \sum_{i=1}^N m(c_{\text{pred}}, c_{\text{gt}}) \cdot \mathbb{I}\{\text{IoU}(b_{\text{pred}}^{3D}, b_{\text{gt}}^{3D}) \geq k\}. \quad (7)$$

Here, N is the total size of annotated instances, c_{pred} and c_{gt} are the predicted caption and the ground truth human annotations for this instance, b_{pred}^{3D} and b_{gt}^{3D} are the predicted 3D bounding box and the ground truth 3D box annotation, and $\mathbb{I}\{\text{condition}\}$ is the identification function that equals 1 if the condition meets, and 0 otherwise.

Our baseline method is listed in Row 1 (Tab. 13). The baseline method is also the LL3DA model, but it only generates instance captions in response to the box prompts generated by the 3D detector [9] and the textual instruction “**### human: describe this object in the given 3D scene. ### assistant:**”. The results listed in the second row of Tab. 13 come from reconstructing 3D bounding boxes from decoding the “`<obj>cx, cy, cz, w, h, l</obj>`” in the generated texts. We use “**### human: given the 3D scene, localize and describe this object. ### assistant:**” as the textual instruction. Though there is still gap between the bounding boxes generated by LL3DA and 3D localizer specialists, LL3DA could generate reasonable bounding box estimations. One can also refer to Fig. 10 for more visualization details.

Table 13. **3D Dense Captioning Performance with 3D Bounding Boxes Generated by LL3DA.** Though there is still gap between LL3DA and 3D specialists for object localization, LL3DA could generate reasonable bounding box estimations.

Test Time Localization	IoU = 0.25				IoU = 0.5			
	C@0.25↑	B-4@0.25↑	M@0.25↑	R@0.25↑	C@0.5↑	B-4@0.5↑	M@0.5↑	R@0.5↑
3D Detector [9]	74.17	41.41	27.76	59.53	65.19	36.79	25.97	55.06
LL3DA	62.90	35.14	26.70	55.62	51.12	29.27	24.18	49.51

Qualitative Results of LL3DA’s Box Predictions. We adopt the spatial location of vote queries to simulate the user click (256 queries per scene), and ask the model to generate 3D bounding boxes in forms of “`<obj>cx, cy, cz, w, h, l</obj>`” as introduced in the main paper. Then, we reconstruct the 3D bounding boxes, and visualize them in Fig. 10. The visualization results show that LL3DA could produce tight bounding boxes close to the objects in diverse and complex 3D environments.



Figure 10. **Visualization of the Bounding Boxes.** We simulate the user click with the spatial position of vote queries proposed in Vote2Cap-DETR [9]. The bounding boxes from LL3DA are reconstructed via decoding the “`<obj>cx, cy, cz, w, h, l</obj>`” from the generated texts. The reconstructed bounding boxes are tight to the objects in the 3D scene.

D. Data and Instructions

In this section, we provide statistic details on the 3D vision and language datasets we train the model (Sec. D.1) as well as the several textual instruction samples for different tasks (Sec. D.2).

D.1. Data Statistics

Our method is trained on the following datasets:

ScanNet [15] is a 3D indoor dataset covering diverse 3D environments, including apartments, living rooms, kitchens, bedrooms, and so on. ScanNet splits the dataset into 1201, 312, and 100 scenes for training, validation, and testing, respectively.

ScanRefer [6] is a 3D vision-language dataset, which contains 36,665 free-form natural language descriptions on 7,875 objects from 562 scenes for training, and 9,508 descriptions on 2,068 objects from 141 scenes from ScanNet [15] for evaluation.

Nr3D [1] is a 3D vision-language dataset with 32,919 free-form natural language descriptions on 4,664 objects from 511 scenes for training, and 8,584 descriptions on 1,214 objects from 130 scenes from ScanNet [15] for evaluation.

ScanQA [2] is a 3D vision-language dataset. The training set of ScanQA consists of 25,563 question-answer pairs on 562 scenes from ScanNet training set. ScanQA further splits 141 unique 3D scenes from ScanNet validation set into two sets:

- The validation set contains 4,675 question-answer pairs on 71 out of 141 validation scenes.
- The “test w/ object” set (“test set with ScanNet object annotations”) contains 4,976 questions on the rest 70 scenes.

Additionally, ScanQA annotates 6,149 questions on 97 scenes on the ScanNet test set as the “test w/o object” set (“test set without ScanNet object annotations”).

3D-LLM [26]. The ScanNet subset of 3D-LLM covers 1) 1,033 descriptions on 517 scenes, 2) 1,864 lines of embodied task planning on 510 scenes, and 3) 2,955 lines of multi-turn embodied dialogues on 517 scenes. All of the annotated scenes come from the training set of ScanNet. The evaluation details could be found in Sec. B.

D.2. Training Samples

We list a couple of the training samples for different tasks in Tab. 14. The special tokens “[Caption]”, “[Question]”, “[Answer]”, “[Box]”, and “[Response]” will be replaced with the language annotations and 3D coordinates.

Table 14. **Data Samples Used to Train LL3DA.** We list a couple of data samples used to train LL3DA for each task.

Task Name	Text Instructions	Visual Interactions	Expected Output
3D Dense Captioning	“### human: describe this object in the given 3D scene. ### assistant: “### human: given the 3D scene, localize and describe this object. ### assistant.””	[Click] or [Box] [Click] or [Box]	[Caption] the object is localized at [Box], [Caption]
3D Question Answering	“### human: given the 3D scene, answer the question: “[Question]” “### assistant:” “### human: answer the question: “[Question]” with the related object locations in the input 3D scene. ### assistant:””	Optional [Clicks]	[Answer]. the related objects are localized at [Box]. the answer is: [Answer].
Scene Description	“### human: describe this 3D scene ### assistant:” “### human: I want to prepare and cook a meal in the kitchen. I have done these things: 1. go to the refrigerator. What should I do next? ### assistant:””	-	[Caption] open the refrigerator
Embodied Planning	“### human: I want to set up a home office workspace. What should I do? ### assistant:””	-	1. place a desk against the wall 2. position a chair at the desk...”
Embodied Dialogue	“### human: ... ### assistant:” “### human: ... ### assistant: ... ### human: ... ### assistant:””	-	[Response] [Response]

E. Additional Implementation Details

Scene Encoder. As introduced in the main paper, we adopt the pre-trained masked transformer encoder [39] as the scene encoder. The masked transformer encoder first tokenizes the input point cloud PC into 2,048 points tokens uniformly scattered in the 3D scene with a set-abstraction layer [44]. Following that are three cascaded transformer encoder blocks with masking radius of 0.16, 0.64, 1.44, respectively. There is another set-abstraction layer between the first two transformer blocks, which further down-samples the encoded tokens into 1,024 point tokens. The output feature of the scene encoder is $f_{enc} = \mathbb{R}^{1,024 \times 256}$, representing a 256 dimensioned feature for each of the 1,024 point tokens.

Multi-Modal Transformer. The feature dimension of the Multi-Model Transformer is 768. We use up to 12 heads for each attention layer, and a total of six transformer layers in practice. As mentioned in the main paper, we choose to load the pre-trained word embeddings and positional embeddings from BERT [18]. The total vocabulary size is 30,522, and the number of position embeddings is 512.

Trainable Parameters. The total number of trainable parameters is about 111M, which is less than 10% of the parameters in the frozen LLM backbone (OPT-1.3B).