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

Sijin Chen[1] Xin Chen[2,*] Chi Zhang[2] Mingsheng Li[1] Gang YU[2]
Hao Fei[3] Hongyuan Zhu[4] Jiayuan Fan[1] Tao Chen[1,†]
[1]Fudan University [2]Tencent PCG [3]National University of Singapore
[4]Institute for Infocomm Research (I[2]R) & Centre for Frontier AI Research (CFAR), A*STAR, Singapore
https://github.com/Open3DA/LL3DA
[*] project lead [†] corresponding author

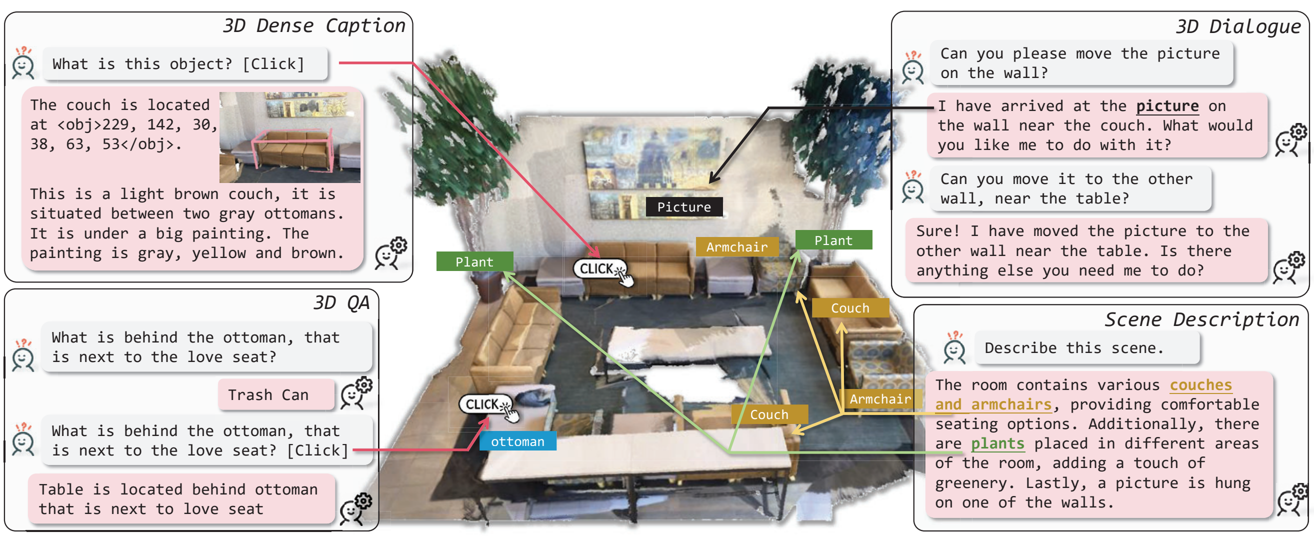

Figure 1. **We propose LL3DA, a Large Language 3D Assistant that demonstrates mighty instruction-following capacities in understanding, reasoning, and planning within complex 3D environments.** LL3DA takes both the textual instructions and potential visual interactions into consideration to help remove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

Abstract

*Recent progress in **L**arge **M**ultimodal **M**odels (LMM) has opened up great possibilities for various applications in the field of 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 representations of the 3D scene. Existing works seek help from multi-view images by projecting 2D features to 3D space, which inevitably leads to huge computational overhead and performance degradation. In this paper, we present LL3DA, a **L**arge **L**anguage **3D A**ssistant that takes point cloud as the direct input and responds to both text instructions and visual interactions. The additional visual interaction enables LMMs to better comprehend human interactions with the 3D environment and further remove the ambiguities within plain texts. 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 **L**arge **L**anguage **M**odel (LLM) families [14, 31, 46, 52, 63] opens up great opportunities for addressing various machine learning tasks in a generalized way [30, 32, 40, 60]. During this LLM carnival, researchers are also seeking generalized LLM solutions for various vision language tasks [37, 50, 62]. Among these, LLM-based 3D scene understanding is a valuable topic that would benefit the development of autonomous driving [9, 23] and embodied AI agents [21, 49]. 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 in various 3D vision and language tasks. The majority of these research build 3D specialists to solve one specific down-stream task, including **3D** **Q**uestion **A**nswering (3D-QA) [2, 43], **3D** **V**isual **G**rounding (3D-VG) [7, 26, 55], and **3D** **D**ense **C**aptioning (3D-DC) [10–12]. There are also several works [4, 13, 34, 70] 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 [24, 57] mainly focus on the understanding of 3D objects. Concurrently, 3D-LLM [29] proposes an LLM-driven solution that aggregates multi-view features for 3D perception, presenting mighty capacities in understanding 3D object and scenes and following text instructions produced by human.

Though these methods have achieved remarkable success in addressing various challenges in understanding the 3D world 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 and depth information. Additionally, plain texts often lead to ambiguities especially in cluttered and complex 3D environments.

To address the above issues, we propose LL3DA, a **L**arge **L**anguage **3D** **A**ssistant that responds to both textual and visual interactions from human, with the ability to understand, reason, and plan in complex 3D environments (Fig. 1). We adopt a multi-modal transformer to aggregate information from textual instructions, visual prompts, and the 3D scene into a fixed length of learnable querying tokens via the attention mechanism. The querying tokens are then projected and used as the prefix for 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 representations and the LLM embedding space, but also extracts interaction-aware 3D scene representations for efficient instruction following.

We conduct extensive experiments to explore the capacities of LL3DA 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, 7], 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 [5, 7, 20, 70] cover tasks requiring a model to adopt its understanding towards a complex 3D scene answering to, or answering with natural language. Among those, **3D** **D**ense **C**aptioning (3D-DC) [10, 12, 54] 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, 12, 54] and the "set-to-set" prediction approaches [10, 11]. The former builds explicit relations on the instance coordinate estimations, while the latter directly learns the locations and descriptions for instances from the input 3D scene. **3D** **V**isual **G**rounding (3D-VG) [1, 7, 55] demands a model to respond the natural language queries with the instance coordinates in the 3D scene. The mainstream of existing methods [4, 65, 70] address 3D-VG via selecting a candidate from a 3D detector's prediction. **3D** **Q**uestion **A**nswering (3D-QA) [2, 43, 59, 66] requires a model to answer the questions with natural language based on the input 3D scene. The majority of existing methods [2, 18, 48] 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 multiple tasks [4, 13, 34, 70]. UniT3D [13] and 3DJCG [4] focus on the joint promotion between 3D-DC and 3D-VG in the relation modelling, while 3D-VLP [34] further includes 3D-QA. Recently, 3D-LLM [29] introduces a family of LLM-driven 3D generalists that could handle diverse textual instructions with reconstructed 3D features from multi-view images [28]. In this paper, we present LL3DA, an LLM solution that directly extracts features from the 3D scene, and handles both visual prompts and textual instructions to diversify the possible interactions human could make with the complex 3D environments.

Large Multimodal Models (LMM). Along with the rapid development of **L**arge **L**anguage **M**odels (LLM) [15, 63], researchers have made great recent efforts adapting LLMs to visual understanding and reasoning tasks [25, 38, 56, 61]. Some methods project or compress global image features as the prefix for text instructions [36, 40, 58, 69], while others extract ROI features as LLM tokens for region-oriented instruction reasoning [6, 64]. Meanwhile, InstructBLIP [17] proposes to extract textual instruction-aware visual features, and has achieved remarkable success in addressing complex

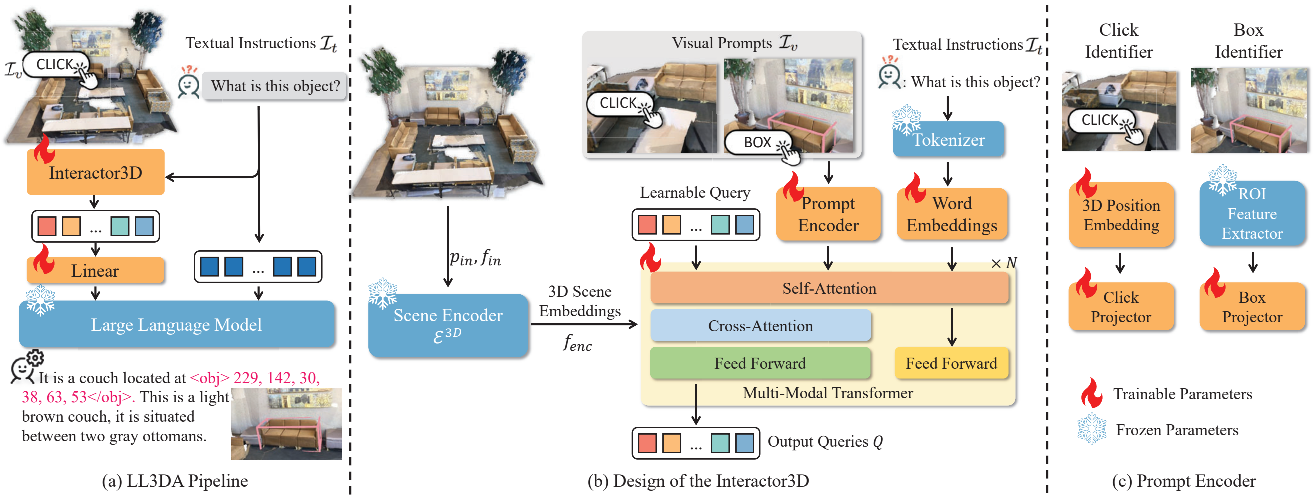

Figure 2. **Overview of the Proposed Approach.** (a) The overall pipeline of our proposed LL3DA first extracts interaction-aware 3D scene embeddings, which are later projected to the prefix of textual instructions as the input of a frozen LLM. (b) The detailed design of the Interactor3D, which aggregates 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.

and unseen instructions. Concurrently, researchers have also made great attempts solving various 3D tasks using LLMs. Notably, [24, 42, 57, 68] 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 text instructions. We also 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 with 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 designs 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 $\mathcal{I}_t$, and potential visual interactions $\mathcal{I}_v$ 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 [57], we begin the textual instructions $\mathcal{I}_t$ 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 [52, 63]. Without the introduction of any additional learnable tokens, we could save the effort of tuning the LLMs.

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 the input. Next, the aggregated 3D scene embeddings are projected to be the prefix of text instructions, and serve as the inputs of a frozen LLM. The detailed design of Interactor3D is shown in Fig. 2 (b), which consists of a frozen 3D scene encoder $\mathcal{E}^{3D}$, a visual prompt encoder, and a multi-modal transformer.

Scene Encoder. We adopt the masked transformer encoder pre-trained on ScanNet detection [10] as the scene encoder, $\mathcal{E}^{3D}$. The scene encoder takes PC as its input, and outputs the 3D scene embeddings:

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

Here, f_{enc} consists of d-dimensioned features for M points uniformly down-sampled from the input 3D scene through the **F**arthest **P**oint **S**ampling (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 [35]. Each user click is first normalized within a range of $[0, 1]$ by the size of the input 3D scene $p_{\text{click}} \in \mathbb{R}^3$. Then, we encode p_{click} with the 3D Fourier positional embeddings [51] function:

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

Here, $B \in \mathbb{R}^{3\times(d/2)}$ is a learnable matrix. The box annotation is represented by the ROI feature $f_{\text{box}} \in \mathbb{R}^d$ extracted by a pre-trained 3D object detector [10]. The two types of the visual prompts are then projected with separate and identical **F**eed **F**orward **N**etworks (FFN).

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

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

Multi-Modal Transformer (MMT) serves as a role to 1) address the contradiction between the permutation-invariant 3D scene embeddings and 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 [17, 36], 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 $[f_{\text{click}}; f_{\text{box}}]$ 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 written as $Q \in \mathbb{R}^{32\times768}$, which are finally projected to the word embedding space of LLMs through a simple linear projector. In practice, we notice that initializing Q-Former with pre-trained BERT [19, 36] 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 [52, 63] as our large language model backbone. The decoder-only LLMs are sensitive to the input orders because of the position embeddings and the causal attention mask. The parameters and the embedding layers of the LLMs are kept frozen to save memory cost. During inference, we generate the responses via searching for the optimal sequence s^* that satisfies:

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

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

4. Multi-modal Instruction Tuning

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

4.1. Tasks and Instructions.

As introduced in Sec. 3.1, LL3DA generates text responses auto-regressively 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, 48]. To diversify the tasks during training, we randomly include additional clicks on the objects related to the questions.

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

Embodied Conversation and Planning could be treated as multi-turn conversations, where we use "*### human:*" and "*### assistant:*" as identifiers to distinguish the source of information as introduced in Sec. 3.1.

4.2. Instruction Following Tuning

During training, for tasks requiring additional visual interactions, *i.e.* 3D-DC and 3D-QA, we randomly choose between clicks or boxes as means of object identification.

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

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

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

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

Table 1. **Quantitative Comparisons for 3D Dense Captioning on ScanRefer[7] 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[12]	56.82	34.18	26.29	55.27	39.08	23.32	21.97	44.78	27.47	17.24	21.80	49.06
MORE[33]	62.91	36.25	26.75	56.33	40.94	22.93	21.66	44.42	-	-	-	-
SpaCap3D[54]	-	-	-	-	44.02	25.26	22.33	45.36	33.71	19.92	22.61	50.50
REMAN[44]	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[8]	-	-	-	-	46.07	30.29	24.35	51.67	33.85	20.70	23.13	53.38
Contextual[67]	-	-	-	-	46.11	25.47	22.64	45.96	35.26	20.42	22.77	50.78
UniT3D[13]	-	-	-	-	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[34]	70.73	41.03	28.14	**59.72**	54.94	32.31	24.83	51.51	-	-	-	-
3D-VisTA*[70]	-	-	-	-	61.60	34.10	**26.80**	55.00	-	-	-	-
Vote2Cap-DETR[10]	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[48]	CLS	-	-	-	-	69.53	**14.64**	13.94	35.15	62.83	11.73	13.28	32.41
Multi-CLIP[18]	CLS	-	-	-	-	68.70	12.65	13.97	35.46	63.20	12.87	13.36	32.61
3D-VLP[34]	CLS	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[70]	CLS	-	-	-	-	68.60	10.50	13.80	35.50	55.70	8.70	11.69	29.60
3D-LLM*[29]	GEN	69.40	12.00	14.50	35.70	69.60	11.60	14.90	35.30	-	-	-	-
LL3DA (Ours)	GEN	**76.79**	**13.53**	**15.88**	**37.31**	**78.16**	13.97	**16.38**	**38.15**	**70.29**	12.19	**14.85**	**35.17**

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

5. Experiments

To test the capacities of LL3DA, we provide numerous evaluations. To begin with, we introduce the datasets, metrics, and implementation details (Sec. 5.1). Then, we compare how our model understands and reasons in complex 3D environments with previous 3D specialists on 3D Dense Captioning and 3D Question Answering (Sec. 5.2). Additionally, we 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 [16], 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 [7], Nr3D [1], ScanQA [2], and the ScanNet subset of 3D-LLM [29]. 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 [53], BLEU-4 [47], METEOR [3], and Rouge-L [39] to evaluate the quality of the generated textual responses.

Implementation Details. Following previous works on 3D vision language tasks [10, 12], we randomly sample 40k points from each 3D scene as the 3D input. We adopt the pre-trained OPT-1.3B [63] as our causal LLM backbone, which is frozen and loaded in float16 to save memory cost. We adopt the AdamW [41] 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 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 ~30k iterations.

3D Dense Captioning demands a model to localize and describe any instance in a 3D scene. We benchmarks state-of-the-art methods on the widely-used ScanRefer [7] and Nr3D [1] dataset in Tab. 1 under the $m@k$IoU metric [12]. 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 [10, 12], 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 [13], 3DJCG [4], and 3D-VLP [34] are pre-trained on multiple 3D vision and language tasks annotated on ScanNet scenes. Additionally, UniT3D [13] adopts off-the-shelf image caption models [45] and multi-view images to generate additional instance-captions for pre-training. It is worth mentioning that we compare the results with the 3D-VisTA [70] model that is not trained on additional 3D scenes. To evaluate our model, we adopt the box predictions produced by Vote2Cap-DETR [10] as the visual prompt. Results show that our method consistently outperforms existing methods on 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 [29] tries to address 3D-QA via auto-regressive text generation (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.

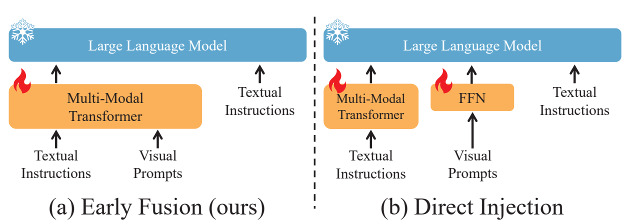

Figure 3. **Two Different Ways to Encode Visual Prompts.** Our proposed method (a) adopts a unified transformer to aggregate features from all kinds of interactions, while (b) directly concatenates the visual prompts to the scene embeddings. Tab. 3 shows that early fusion leads to a better performance.

Effectiveness of the Q-Former Design. We list two ways to process the visual prompts in Fig. 3. Here, Fig. 3 (a) is our proposed method that adopts a unified transformer to aggregate information from both text instructions and visual prompts, while Fig. 3 (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 feature 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[7].** 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 task 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 affect the performance when serving 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 they are both good auxiliary tasks for 3D-DC.

Table 4. **Effectiveness of Instructions as 3D Dense Captioning Auxiliary Tasks.** We train the models from scratch and evaluate on ScanRefer[7]. "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. We have made a similar study to analyze how adopting additional "answer and localize" instructions and visual prompts improves 3D-QA on ScanQA [2] as auxiliary tasks in Tab. 6. We do not use any visual interactions during inference. Results show that the additional textual instructions and visual prompts improve the task diversity and further improve the performance for 3D-QA.

Performance as a Generalist. To test whether LL3DA can distinguish different tasks given the textual instructions and visual prompts, we evaluate our model on different tasks

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[7] 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**

without task-specific fine-tuning in Tab. 5. The first three rows list the performance of LL3DA when trained from scratch on one specific task, while the following three rows represent the fine-tuned models. The last row indicates the direct evaluation of LL3DA. Results show that our model could distinguish 3D-DC and 3D-QA given the text instructions and visual prompts, and achieve strong performance (62.98% C@0.5 on ScanRefer, 75.67% CiDEr on ScanQA). However, the generalist model achieves poor performance on Nr3D [1], which is because we did not try to differentiate between Nr3D and ScanRefer during training in the first place, as ScanRefer and Nr3D are used for the same task. There is also an interesting observation that though we did not differentiate between these two datasets for 3D-DC, 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 can serve as a strong initialization 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 text instructions are necessary for 3D-DC in Tab. 7. The first row is our baseline method that directly generates the captions based on visual prompts without any text instructions, and the second row is our method that is trained with the text 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 when generating results in specific domains/tasks.

Table 7. **Effectiveness of Instructions on 3D Dense Captioning.** We perform experiments on ScanRefer[7]. 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.

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↑	BLEU-4↑	METEOR↑	Rouge-L↑
-	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 [22] and top-p [27] sampling strategy with $k = 50$ and $p = 0.95$.

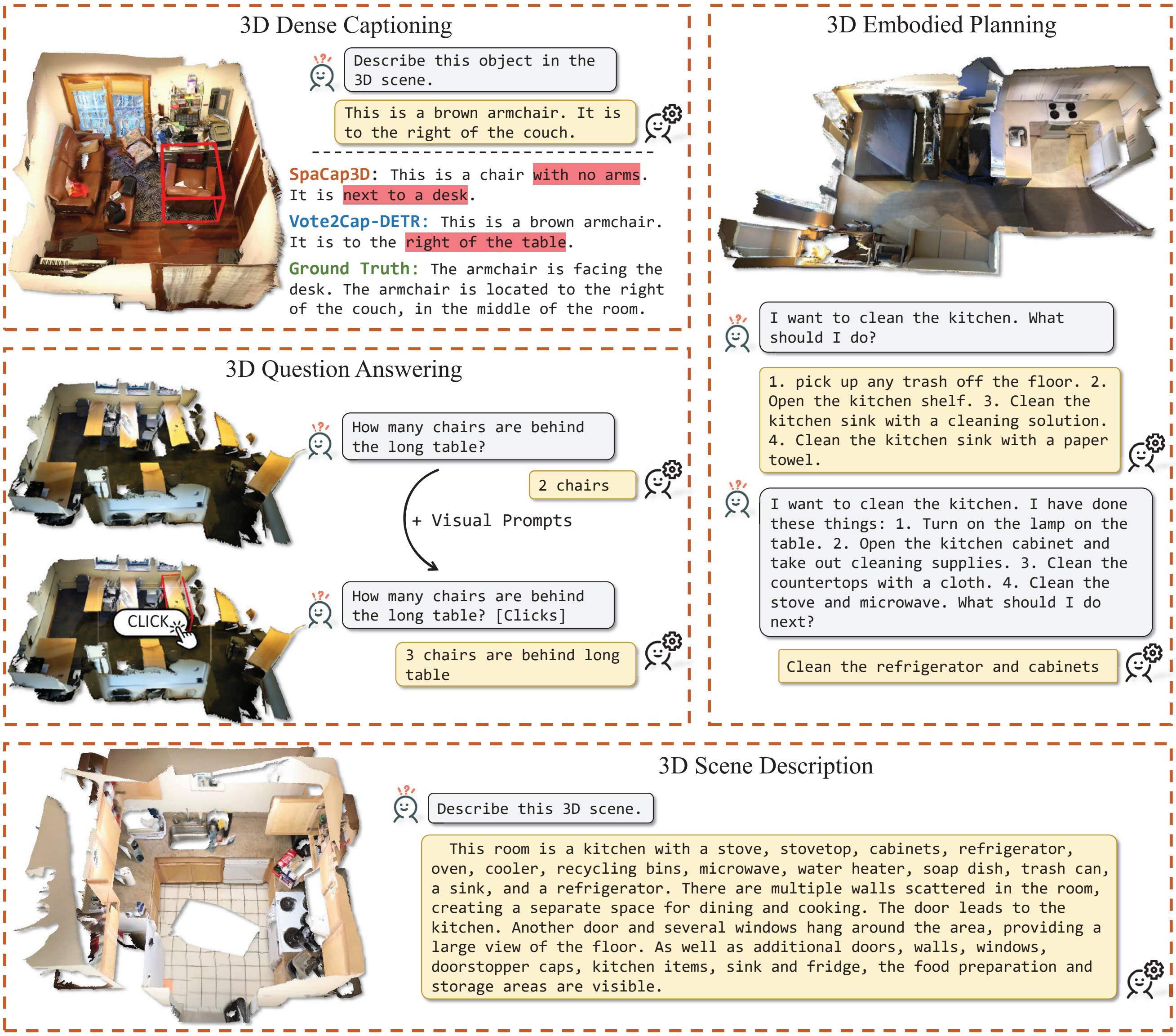

Figure 4. **Qualitative Results.** We provide several visualization results on various 3D vision and language tasks in diverse 3D environments (living room, classroom, kitchen, and bedroom). Red highlights the wrong answer.

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 the 3D point cloud and aggregates information from scenes and human interactions with the attention mechanism. We show that the visual interactions could remove the ambiguities in cluttered 3D environments, showing mighty instruction-following capacities. Experiments show that our method could achieve remarkable results on various 3D vision-language benchmarks. We hope our approach could inspire further designs and training strategies for large 3D language models. In future studies, we believe that the construction of high-quality and diverse annotations will further enhance the model's reasoning and planning capabilities.

7. Acknowledgements

This work is supported by National Natural Science Foundation of China (No. 62071127 and 62101137), National Key Research and Development Program of China (No. 2022ZD0160100), Shanghai Natural Science Foundation (No. 23ZR1402900), Shanghai Municipal Science and Technology Major Project (No. 2021SHZDZX0103), A*STAR AME Programmatic Funding A18A2b0046, RobotHTPO Seed Fund under Project C211518008, and EDB Space Technology Development Grant under Project S22-19016-STDP. The computations in this research were performed using the CFFF platform of Fudan University.

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, 7

[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

[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

[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] Yang Cao, Yihan Zeng, Hang Xu, and Dan Xu. Coda: Collaborative novel box discovery and cross-modal alignment for open-vocabulary 3d object detection. In *NeurIPS*, 2023. 2

[6] 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

[7] 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

[8] 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. 5

[9] 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. 1

[10] 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

[11] 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

[12] 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

[13] 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

[14] 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

[15] 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

[16] 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] 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, 2023. 2, 4

[18] 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, 5

[19] 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] 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

[21] 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

[22] Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*, 2018. 7

[23] Daocheng Fu, Xin Li, Licheng Wen, Min Dou, Pinlong Cai, Botian Shi, and Yu Qiao. Drive like a human: Rethinking autonomous driving with large language models. *arXiv preprint arXiv:2307.07162*, 2023. 1

[24] 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

[25] Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. Chartllama: A multimodal llm for chart understanding and generation. *arXiv preprint arXiv:2311.16483*, 2023. 2

[26] Dailan He, Yusheng Zhao, Junyu Luo, Tianrui Hui, Shaofei Huang, Aixi Zhang, and Si Liu. Transrefer3d: Entity-and-relation aware transformer for fine-grained 3d visual grounding. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2344–2352, 2021. 2

[27] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019. 7

[28] 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

[29] 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

[30] Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. *arXiv preprint arXiv:2311.12871*, 2023. 1

[31] 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

[32] 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

[33] 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. 5

[34] 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

[35] 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

[36] 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, 4

[37] Mingsheng Li, Xin Chen, Chi Zhang, Sijin Chen, Hongyuan Zhu, Fukun Yin, Gang Yu, and Tao Chen. M3dbench: Let's instruct large models with multi-modal 3d prompts. *arXiv preprint arXiv:2312.10763*, 2023. 1

[38] Yanda Li, Chi Zhang, Gang Yu, Zhibin Wang, Bin Fu, Guosheng Lin, Chunhua Shen, Ling Chen, and Yunchao Wei. Stablellava: Enhanced visual instruction tuning with synthesized image-dialogue data. *arXiv preprint arXiv:2308.10253*, 2023. 2

[39] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 5

[40] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 1, 2

[41] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

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

[43] 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

[44] 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. 5

[45] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 6

[46] OpenAI. Gpt-4 technical report, 2023. 1

[47] 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

[48] 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, 5

[49] 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

[50] Zhende Song, Chenchen Wang, Jiamu Sheng, Chi Zhang, Gang Yu, Jiayuan Fan, and Tao Chen. Moviellm: Enhancing long video understanding with ai-generated movies. *arXiv preprint arXiv:2403.01422*, 2024. 1

[51] 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

[52] 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

[53] 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

[54] 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, 5

[55] 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

[56] Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[57] 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

[58] 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

[59] Shuquan Ye, Dongdong Chen, Songfang Han, and Jing Liao. 3d question answering. *IEEE Transactions on Visualization and Computer Graphics*, 2022. 2

[60] Fukun Yin, Xin Chen, Chi Zhang, Biao Jiang, Zibo Zhao, Jiayuan Fan, Gang Yu, Taihao Li, and Tao Chen. Shapegpt: 3d shape generation with a unified multi-modal language model. *arXiv preprint arXiv:2311.17618*, 2023. 1

[61] 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

[62] Chi Zhang, Zhao Yang, Jiaxuan Liu, Yucheng Han, Xin Chen, Zebiao Huang, Bin Fu, and Gang Yu. Appagent: Multimodal agents as smartphone users. *arXiv preprint arXiv:2312.13771*, 2023. 1

[63] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui 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

[64] 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. 2

[65] 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

[66] Lichen Zhao, Daigang Cai, Jing Zhang, Lu Sheng, Dong Xu, Rui Zheng, Yinjie 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

[67] Yufeng Zhong, Long Xu, Jiebo Luo, and Lin Ma. Contextual modeling for 3d dense captioning on point clouds. *arXiv preprint arXiv:2210.03925*, 2022. 5

[68] Qiang Zhou, Chaohui Yu, Shaofeng Zhang, Sitong Wu, Zhibing Wang, and Fan Wang. Regionblip: A unified multi-modal pre-training framework for holistic and regional comprehension, 2023. 3

[69] 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

[70] 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