


# M3DBench: Towards Omni 3D Assistant with Interleaved Multi-modal Instructions

Mingsheng Li<sup>1</sup>, Xin Chen<sup>2</sup>, Chi Zhang<sup>2</sup>, Sijin Chen<sup>1</sup>, Hongyuan Zhu<sup>3</sup>, Fukun Yin<sup>1</sup>, Zhuoyuan Li<sup>1</sup>, Gang Yu<sup>2</sup>, and Tao Chen<sup>1</sup>, 

<sup>1</sup> School of Information Science and Technology, Fudan University, Shanghai, China  
[limc22@m.fudan.edu.cn](mailto:limc22@m.fudan.edu.cn), [eetchen@fudan.edu.cn](mailto:eetchen@fudan.edu.cn)

<sup>2</sup> Tencent PCG, Shanghai, China

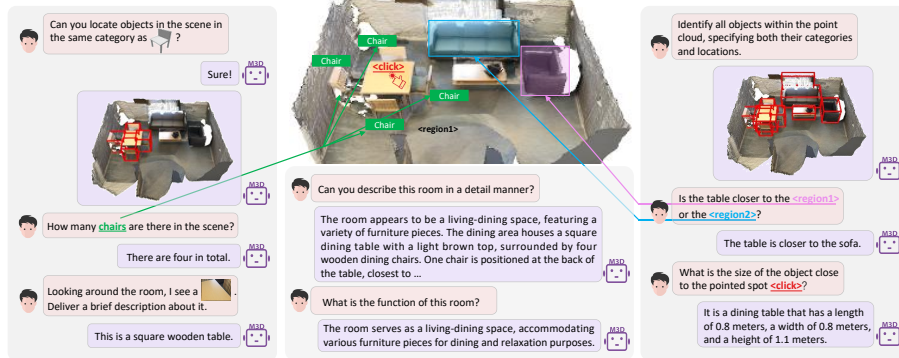
<sup>3</sup> Institute for Infocomm Research (I<sup>2</sup>R) & Centre for Frontier AI Research (CFAR), A\*STAR, Singapore

**Abstract.** Recently, the understanding of the 3D world has garnered increased attention, facilitating autonomous agents to perform further decision-making. However, the majority of existing 3D vision-language datasets and methods are often limited to specific tasks, limiting their applicability in diverse scenarios. The recent advance of **Large Language Models (LLMs)** and **Multi-modal Language Models (MLMs)** has shown mighty capability in solving various language and image tasks. Therefore, it is interesting to unlock MLM’s potential to be an omni 3D assistant for wider tasks. However, current MLMs’ research has been less focused on 3D due to the scarcity of large-scale visual-language datasets. In this work, we introduce M3DBench, a comprehensive multi-modal instruction dataset for complex 3D environments with **over 320k instruction-response pairs** that: 1) supports **general interleaved multi-modal instructions** with text, user clicks, images, and other visual prompts, 2) unifies **diverse region- and scene-level 3D tasks**, composing **various fundamental abilities** in real-world 3D environments. Furthermore, we establish a new benchmark for assessing the performance of large models in understanding interleaved multi-modal instructions. With extensive quantitative and qualitative experiments, we show the effectiveness of our dataset and baseline model in understanding complex human-environment interactions and accomplishing general 3D-centric tasks. We will release the data and code to accelerate future research on developing 3D MLMs.

**Keywords:** Multi-modal Learning

## 1 Introduction

The past year has witnessed remarkable success of **Large Language Models (LLMs)** families [55, 60, 20, 58] in addressing various general language processing tasks through instruction tuning [47]. **Multi-modal Language Models (MLMs)**, such as Flamingo [2], BLIP-2 [37], LLaVA [41] have progressed various visual comprehension and reasoning tasks on 2D domain, including image captioning [6, 56, 64], visual dialogue [13] and question-answering [26, 23]. To unlock the full



**Fig. 1:** Examples from M3DBench. Enabling effective interaction between humans and 3D Assistants poses numerous challenges, particularly in comprehending interleaved multi-modal instructions that may involve language, user clicks, captured images, and more, all while seamlessly executing diverse tasks. M3DBench introduces a range of tasks crafted to foster comprehensive interaction, addressing ambiguities through fine-grained multi-modal instructions.

potential of these MLMs, it is essential to curate a well-constructed instruction-following dataset [41, 35], which empowers models to handle diverse vision language (VL) tasks without extensive modifications to the architecture. However, current research on MLMs has predominantly overlooked 3D visual, and a comprehensive dataset for 3D instruction tuning is missing due to the daunting workload of collecting instructions in ambiguous and cluttered 3D environments.

Previous works have made efforts to construct datasets for specialized 3D task, such as object detection [21, 59], visual grounding [1, 11], dense captioning [1, 11], VQA [4, 67], and navigation [3]. Consequently, most of the models [52, 44, 8, 12, 19, 4] are specialists in only one or two of these tasks, potentially limiting their adaptability across various applications. Works such as LAMM [70], 3D-LLM [25], and Chat-3D [65] have made preliminary attempts in constructing 3D instruction-following datasets, achieving inspiring results. However, the range of visual tasks covered by these datasets is relatively *limited*, which constrains their effectiveness under diverse scenarios. These datasets primarily focus on language-only instructions, posing challenges in identifying a specific object within 3D environments, which are often cluttered and complex. For example, in a scene with multiple **wooden chairs**, distinguishing a particular one using language alone may require detailed instruction, as a simple reference like “a **wooden chair**” might result in *ambiguity*. Furthermore, the lack of a comprehensive evaluation *benchmark* poses challenges in assessing the capability of large models on 3D-centric tasks. Current works, such as LAMM [70], primarily evaluate the model’s performance on previous benchmarks that are not designed for assessing MLMs with open-form output [25].

**Table 1:** Comparison between M3DBench and other 3D VL datasets. M3DBench has the following characteristics: 1) a comprehensive instruction-following dataset tailored for 3D scenes. 2) Supporting multi-modal instructions that interleave text and diverse visual prompts. 3) Spanning fundamental abilities in real-world 3D environments, such as visual perception, scene understanding, spatial reasoning, and planning.

Dataset	Statistics		Prompt Modality				Perception				Understanding and Reasoning				Planning		
	#Instruction- response pairs	#Average length of instruction / response	Text	Click	Region	Image	3D Shape	Object Detection	Visual Grounding	Dense Caption	Visual Question Answering	Embodied Question Answering	Multi- region Reasoning	Scene Description	Multi- round Dialogue	Embodied Planning	Vision- Language Navigation
N3D [1]	-	-	✓	×	×	×	×	×	✓	✓	×	×	×	×	×	×	×
ScanRefer [11]	-	-	✓	×	×	×	×	×	✓	✓	×	×	×	×	×	×	×
ScanQA [4]	25K	8.77 / 2.42	✓	×	×	×	×	×	×	✓	×	×	×	×	×	×	×
SQA3D [43]	26K	10.49 / 1.10	✓	×	×	×	×	×	×	✓	✓	×	×	×	×	✓	×
ScanScribe [78]	-	-	✓	×	×	×	×	-	-	-	-	-	-	-	-	-	-
LAMM-3D [70]	10K	13.88 / 119.34	✓	×	×	×	×	✓	×	×	×	×	✓	✓	✓	×	×
3DLLM [25]	202K	43.80 / 8.11	✓	×	×	×	×	×	✓	✓	✓	✓	×	✓	✓	✓	✓
Chat-3D [65]	57K	9.11 / 48.75	✓	×	×	×	×	×	✓	×	×	×	×	×	✓	×	×
M3DBench	<b>327K</b>	24.79 / 18.48	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

In this paper, we introduce a comprehensive 3D instruction-following dataset called M3DBench, serving as the foundation for developing a versatile and practical assistant in the real-world 3D environment. Our dataset comprises general 3D-centric tasks at both object and scene levels and over 320K instruction-response pairs, covering fundamental capabilities such as visual perception, scene understanding, spatial reasoning, embodied planning, and VL navigation, as listed in Tab. 1. Furthermore, to tackle the challenge of ambiguity in language-only instructions, we interleave text instructions with visual prompts that provide rich clues about instances in the scene, including the user’s click, pointed region, photographed images, 3D shape (as shown in Fig. 1) in M3DBench, to enhance the granularity, diversity, and interactivity of generated instructions (such as “find the  $\langle$ image of a whiteboard captured by mobile phone $\rangle$  in the room”).

To evaluate the effectiveness of M3DBench, we develop a primary yet effective baseline model capable of processing interleaved multi-modal instructions, consisting of three components: scene perceiver, multi-modal instruction adapter, and LLM decoder. Furthermore, we develop a comprehensive benchmark for assessing the general capabilities of large models when handling multi-modal instructions. The evaluation benchmark comprises approximately 1.5K instances, encompassing both region-level and scene-level tasks, such as object description, multi-region reasoning, embodied planning, and multi-round dialogues. We believe that M3DBench will provide a solid foundation for future research in 3D MLMs with interleaved multi-modal instructions.

To summarize, our contributions are listed as follows:

- We introduce a large-scale 3D-centric instruction-response dataset that unifies both region-level and scene-level tasks, focusing on scene perception, understanding, reasoning, and planning.
- We propose an interleaved multi-modal instruction formula designed to enhance the granularity, diversity, and interactivity of instructions.
- We present an LLM-based model capable of understanding multi-modal instructions and executing multiple tasks.

- We establish a comprehensive benchmark for evaluating the capabilities of MLMs within 3D scenarios. Extensive experiments demonstrate the effectiveness of both the dataset and the baseline.

## 2 Related Work

### 2.1 Multi-modal Datasets and 3D Benchmarks

The progress of MLMs [54, 31, 38, 37] has been greatly accelerated by the availability of large-scale image-text data, such as MS COCO Caption [17], Visual Genome [33], LAION-5B [57]. In order to improve models’ comprehension of human instructions in visual tasks, several visual instruction-following datasets [41, 35, 70, 22] have been proposed. Additionally, while numerous studies in the field of 3D have presented benchmark datasets for visual grounding [1, 11], dense captioning [1, 11], and visual question answering [4, 43], these datasets are limited to specific tasks. In this paper, we propose a comprehensive dataset that supports interleaved multi-modal instructions and covers various 3D-centric tasks, including multi-region reasoning, scene description, multi-round dialogue, and so on. Refer to Tab. 1 for a detailed comparison between our dataset and other 3D VL datasets [1, 11, 4, 43] as well as exiting 3D visual instruction datasets [70, 25, 65]. Furthermore, rather than providing demonstrations only, we evaluate diverse tasks with quantitative results.

### 2.2 Multi-modal Foundation Models

With the triumph of LLMs [7, 75, 55, 60, 20], recent studies [2, 37, 41, 36, 29, 30, 69] start to explore Vision Language Models (VLMs), extending the capabilities of LLMs in solving diverse visual-related tasks. Early attempts include Flamingo [2], which incorporates visual features through gated cross-attention dense blocks, and BLIP-2 [37], which uses a Q-former as a bridge to reconcile the modality gap between the frozen image encoder and LLMs. In order to enhance the VLMs’ comprehension of human instructions, several visual instruction tuning methods [41, 36] have been proposed. Addressing the adaptation of LLMs to 3D-related tasks, LAMM [70] uses a simple projection layer to connect the 3d encoder and LLM. 3D-LLM [25] utilizes point clouds and text instructions as input, leveraging 2D VLMs as backbones. However, prior works that attempt to integrate the 3D world into MFMs have exhibited limitations in handling interleaved multi-modal instructions and accomplishing various tasks. In this work, we propose to improve the abilities of MFMs in addressing diverse 3D-centric tasks and handling interleaved multi-modal instructions with on a comprehensive 3D instruction dataset.

### 2.3 3D Vision-language Learning

Recently, there has been growing interest in 3D VL learning [24, 18, 28, 78, 27, 14]. While various 3D representations exist, including voxels, point clouds, and neural

fields, previous works have primarily focused on point cloud-text data. Among those, 3D dense captioning [18, 72, 16] aims to generate description of target object within a 3D scene, while 3D visual grounding [68, 77, 73] involves identifying object in a scene based on textual description. In 3D question answering [4, 67], models are required to answer questions based on the visual information. Although these works have achieved impressive results in connecting 3D vision and language, they heavily rely on task-specific model design. In contrast, we develop a unified model based on LLMs, capable of decoding multiple 3D-related tasks without the need for specific model designs. Furthermore, we establish a comprehensive benchmark to assess the model’s performance across various tasks.

### 3 Interleaved Multi-modal Instruction Dataset

We introduce the design recipe for interleaved multi-modal instructions in Sec. 3.1, along with the strategy for constructing a multi-modal 3D instruction dataset in Sec. 3.2. Then we detail the tasks at both the region-level and scene-level covered by the dataset in Sec. 3.3. The statistical and analytical examination of the dataset is presented in the supplementary materials.

#### 3.1 Interleaved Multi-modal Instruction

In contrast to prior 3D instruction datasets limited to language-only instructions, M3DBench introduces multi-modal instructions by interleaving text with diverse visual prompts. Specifically, we generate visual prompts for 3D scenes and design a formula that seamlessly integrates text with these interactive prompts, covering the user’s click, selected region, image, and 3D shape.

- **User’s Click** is generated by sampling a point from object-level annotations within the 3D scene, either inside the object or along its surface.
- **Region Prompts** are derived from the 3D bounding boxes annotations. Each box is further scaled with ratios in the interval  $[0.8, 1.2]$ .
- **Image Prompts** integrate objects from images corresponding to 3D scene, publicly available database(e.g. ImageNet [34]) and synthetic images (e.g. SDXL [51]).
- **3D shape prompts** are created by selecting instances from 3D scenes that come with object-level annotations. Furthermore, we enrich our collection with models sourced from the public 3D asset repository [10].

To generate 3D multi-modal instructions, we design a unified formula utilizing special tokens as placeholders to seamlessly integrate textual descriptions with different types of visual prompts. The special tokens are uniformly denoted as `<click>`, `<region>`, `<image>`, and `<shape>` respectively. For instance, a clicked point is represented as `<click>x, y, z</click>`, while a selected region is represented as `<region>cx, cy, cz, w, h, l</region>`. Visual prompts for images and 3D shapes are indicated by `<image>image_id</image>` and

`<shape>shape_id</shape>` respectively, where `image_id` and `shape_id` serve as unique identifiers for the images and 3D shapes within the dataset. Finally, an interleaved 3D multi-modal instruction  $I$  can be defined as an ordered sequence composed of text and visual prompts, represented as  $I = [x^1, x^2, \dots, x^M]$ , where each element  $x^i \in \{\text{text descriptions}, \text{visual prompts}\}$ . More examples and details can be found in the supplementary.

### 3.2 Dataset Construction Engine

Annotating large-scale visual instruction data requires expert-level detailed instructions and corresponding responses for various tasks, which is time-consuming and labor-intensive [40]. Drawing inspiration from advancements in image instruction generation [41, 35, 70], we utilize the GPT-API [58, 46] to construct M3Dbench, without the need for manual annotation. To achieve this, we develop a three-stage data generation pipeline.

**Stage I: Instruction & Response Generation.** To construct M3Dbench, we utilize existing datasets [21, 11, 1, 71, 9, 32, 34, 10] and generate instruction-response data through both template-based and LLM-prompting methods. Specifically, for 3D-only tasks such as object detection, we manually create instruction and response templates. The instruction templates consist of task descriptions and desired output format, while the response templates integrate ground-truth labels (e.g. coordinates) into a natural language context. For 3D-language tasks, like scene description and visual question answering, we prompt the GPT-API [58, 46] with processed object attributes, textual descriptions, carefully crafted system messages as well as few-shot examples to generate task-specific instruction data. Notably, within the system messages, we instruct GPT-API [58, 46] to generate instances in both declarative and interrogative forms to improve diversity. Furthermore, we impose explicit length constraints to ensure the generated content’s conciseness and relevance. More details regarding the templates and prompts can be found in the supplementary.

**Stage II: Data Quality Enhancement.** While most instruction-response pairs generated by GPT-APIs [58, 46] are of high quality, some may include unwanted content, such as “based on the given description”. To address this, we utilize pattern matching with specific keywords to exclude such pairs. Additionally, it has been observed in prior studies that LLMs often suffer from hallucinations while generating responses, such as non-existent objects. To address this challenge, we adopt a strategy where instructions are treated as queries, and scene information is provided as context when invoking GPT-APIs. We let human volunteers verify the consistency between the new responses and the original ones, and engage in multiple rounds of rewriting for inconsistent responses.

**Stage III: Visual Prompts Injection.** During Stage I of the instruction generation, we additionally preserve the IDs of instances within the instructions. Therefore, we can substitute instances’ IDs in the instructions with four types of visual prompts described in Sec. 3.1. By explicitly incorporating visual prompts into the instructions, the diversity and interactivity of instructions have been enhanced, with the experiment in Sec. 5.3 demonstrating the advantages of

multi-modal instructions over language-based ones. In total, we curate over 320K pairs of instruction-following data, of which more than 138K instructions include the visual prompts we proposed.

### 3.3 Task Coverage

M3DBench introduces a unified *instruction-response* format to cover diverse 3D-centric tasks, encompassing essential capabilities ranging from visual perception and understanding to reasoning and planning (detailed in Tab. 1). Examples of different tasks can be found in the supplementary.

**Object Detection(OD)** aims at identifying and locating all the objects of interest in a point cloud [48, 45]. Here, we transform the classic OD task into an instruction-following format by providing task descriptions and specifying the desired output format. Following LAMM [70], we manually design a set of instruction-response templates with placeholders, and each instruction includes the expected output format. The instruction and response templates can be found in the supplementary.

**Visual Grounding(VG)** involves identifying the target object in the scene based on a natural language referring expression [74, 66]. In M3DBench, we expand the task format of VG. Specifically, our description information for querying extends beyond textual input and includes various visual prompts, such as coordinate, clicked point, image, 3D object, and so on. Moreover, our output is not limited to locating a single target object but can also involve finding objects belonging to the same category.

**Dense Caption(DC)** requires a model to generate natural language descriptions for each object [18, 15]. However, existing DC datasets like ScanRefer [11] and Nr3D [1] provide only short captions. In M3DBench, we reconstruct the DC datasets and introduce terms like *brief* or *detailed* in instruction to generate either concise title or detailed description for the object, which allows for better control over the granularity of the generated caption. The instruction templates can be found in the supplementary.

**Visual Question Answering(VQA)** is a task that requires the model to correctly answer a given question based on the information present in a visual scene [4, 50]. In this work, we curate a collection of free-form, open-ended question-answer pairs using publicly available 3D-language datasets. These VQA pairs cover various aspects at both the object level and scene level, including instance locations and attributes, object counts, room functions, and more.

**Embodied Question Answering(EQA)**. Unlike traditional VQA tasks [4, 50] that primarily focus on answering questions related to global information, EQA requires the agent to first comprehend and analyze the surrounding environment to answer questions under that situation [43]. To collect instruction-following data for EQA, we start by randomly selecting a location within the scene and choosing to face a nearby object for reference direction, and then prompt GPT-4 to generate EQA pairs based on the given situation and text information.

**Multi-region Reasoning(MR)**. Datasets such as DC [11, 1] facilitate understanding and reasoning for individual objects. However, reasoning between

distinct regions is often overlooked. For instance, inquiries about the spatial relationship between ⟨region 1⟩ and ⟨region 2⟩. Here, we introduce MR, which is designed to enhance fine-grained comprehension of multiple regions of interest. Our methodology involves feeding object location, descriptions [71], few-shot examples, and language instructions to GPT-4 to obtain corresponding responses.

**Scene Description(SD).** Unlike DC [18, 15], which generates a caption for each object, SD focuses on producing descriptions of the entire scene, extending the descriptive ability of MLMs from the region level to scene level. To construct instruction-following data for SD, we extract 3D bounding box annotations from ScanNet [21] and dense captions from the 3D VL datasets [11, 1] as data sources. By prompting GPT-4, we can generate detailed descriptions for each scene.

**Multi-round Dialogue(MD).** To construct MDs, we make use of 3D VL datasets and follow a similar approach to that used in LLAVA [41]. During this process, we prompt GPT-4 to generate MDs in a self-questioning and self-answering format, taking advantage of coordinate information and language descriptions from [1, 11].

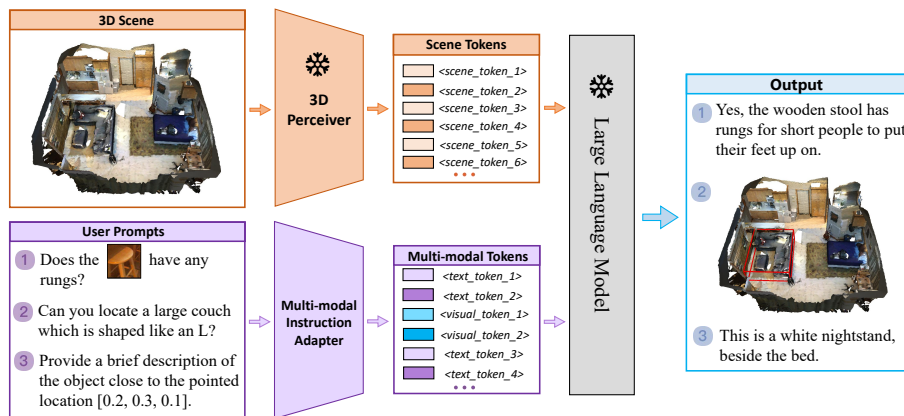
**Embodied Planning(EP).** Unlike EQA that primarily focuses on answering questions, EP requires agents to possess planning and decision-making capabilities. Specifically, the agent needs to perceive the environment, understand user’s intentions, and generate appropriate action instructions to achieve predefined goals [25].

**Vision Language Navigation(NLV)** require an agent to navigate and move in a real-world 3D environment based on human instructions. We leverage annotations from existing 3D-language navigation tasks [32] and transform them into an instruction-following format. Instructions are expressed in natural language, while the corresponding response is a trajectory formed by points in space.

## 4 Multi-modal Instruction Tuning

Existing LLM-based methods [70, 25, 65] are designed for text-based instructions and struggle with handling inputs that integrate multiple types of prompts. To address this, we introduce a baseline model that can perceive scenes and understand the interleaved multi-modal instructions, accomplishing a variety of tasks through a unified decoder. As shown in Fig. 2, the framework consists of three parts: scene perceiver, multi-modal instruction adapter, and LLM. First, the 3D scene is processed by the scene perceiver, and the features are then projected into the same feature space as the language embedding using a trainable projection layer (Sec. 4.1). Simultaneously, instructions are processed by the multi-modal instruction adapter, where information from various modalities is decoupled and encoded into the language space (Sec. 4.2). Then the visual and instruction tokens are concatenated and fed into the LLM (Sec. 4.3). Next, we will provide a detailed description of each module.





**Fig. 2:** Overview of our baseline model. We utilize scene perceiver to extract scene tokens from 3D visual input. Multi-modal instructions are transformed into corresponding instruction tokens via multi-modal instruction adapter. The scene tokens and multi-modal instruction tokens are then concatenated and fed into a frozen LLM, which generates the corresponding responses subsequently. The 3D encoder, image encoder and LLM are frozen during the training process.

#### 4.1 3D Scene Perceiver

Given the point cloud of a scene, denoted as  $P$ , we employ a pre-trained 3D encoder to extract scene feature:

$$f_s = \mathcal{E}^{3D}(P). \quad (1)$$

Similar to LLAVA [41], we also utilize a trainable visual feature projection matrix  $\mathcal{W}^{3D}$  to project the visual features into the language embedding space and obtain scene tokens:

$$X_s = \mathcal{W}^{3D} \cdot f_s. \quad (2)$$

The scene embeddings are denoted as  $X_s = \{x_s^n\}_{n=1}^N$ , where  $x_s^n \in \mathbb{R}^d$  and  $N$  represents the number of visual tokens.  $d$  represents the dimension of hidden states in LLM.

#### 4.2 Multi-modal Instruction Adapter

To enable LLMs to understand interleaved multi-modal instructions, we introduce **Multi-modal Instruction Adapter (MIA)**, capable of processing textual descriptions along with diverse visual prompts. Initially, the MIA discerns the types of visual prompts within an instruction, utilizing predefined special tokens as detailed in Sec. 3.1. Based on this identification, different types of prompts will be decoupled and fed into the appropriate prompt encoder to extract modal-specific features.

As for **textual descriptions** within the instructions, MIA utilizes the tokenizer and word embedding from LLMs to obtain corresponding tokens. To process **user’s click** and **pointed region** in the instructions, MIA encodes them using two separate and learnable, as demonstrated by the following formulation:

$$\begin{aligned} X_{click} &= \mathcal{W}^{click} \cdot p_{point} \\ X_{region} &= \mathcal{W}^{region} \cdot p_{region} \end{aligned} \quad (3)$$

Here,  $p_{click} \in \mathbb{R}^3$  and  $p_{region} \in \mathbb{R}^6$  represent the position of the user’s click and pointed region, respectively. For **image** prompts, MIA leverages the frozen CLIP [54] to extract image features, subsequently fed into a pre-trained projector from LLaVA [41] to generate image tokens. In the case of **3D shape**, MIA downsamples them to 1024 points and normalizes their coordinates into a unit sphere [78]. Then a pre-trained encoder is used to extract the object’s features, and a trainable projection matrix is inserted between the encoder and LLM to adjust these tokens.

### 4.3 LLM Decoder

We utilize the pre-trained LLM [75, 61, 20] as a unified decoder for various vision-centric tasks. To accomplish this, we employ a 3D scene perceiver (Sec. 4.1) to encode the input scene  $P$  into discrete scene tokens  $X_s = \{x_s^n\}_{n=1}^N$ . These tokens are then concatenated with the multi-modal instruction tokens  $X_i = \{x_i^n\}_{n=1}^M$ . LLM takes both the scene tokens and the multi-modal instruction tokens as input and predicts the probability distribution of the output token  $X_o = \{x_o^n\}_{n=1}^L$  in an auto-regressive manner:

$$P_\theta(X_o|X_s, X_i) = \prod_n P_\theta(x_o^n|x_o^{<n}; X_s, X_i). \quad (4)$$

Furthermore, for tasks that rely on coordinates for assessment, such as visual grounding, we decouple them from the output of LLMs (detailed in the supplements). This simple approach enables us to develop a unified framework for a wide range of 3D-only tasks without the need for modifications to the existing LLMs [75, 60, 7].

### 4.4 Training Strategy

The training objective is to maximize the likelihood of generating this target response sequence  $X_o = \{x_o^n\}_{n=1}^L$ , given the visual input  $X_s$  and multi-modal instruction  $X_i$ :

$$\mathcal{L}_\theta = - \sum_{n=1}^L \log P_\theta(x_o^n|x_o^{<n}; X_s, X_i). \quad (5)$$

Here,  $\theta$  represents the trainable parameters. Note that during training, we freeze 3D encoder, image encoder, as well as language decoder, and only train the

projection layers to enable rapid iterations. Exploring alternative architecture or refining training strategy may yield further improvements. We leave this as a direction for future work.

## 5 Experiments

We first introduce the baseline model, metrics, and implementation details in Sec. 5.1. Additionally, we provide the main results and analyses in Sec. 5.2. We showcase some visualization results in Sec. 5.4. In Sec. 5.3, we give insightful analyses of multi-modal instructions and zero-shot performance.

### 5.1 Baseline, Metrics, and Implementations

**Baseline.** Since there is no prior method that works out of the box with our interleaved multi-modal instruction setup, we develop several variants as baseline to accommodate M3DBench. Specifically, we incorporate two different types of 3D encoders, based on PointNet++ [53] and Transformer [62], into our baseline model. Furthermore, we consider two widely-used versions of LLMs as our language decoder: OPT-6.7B [75] and LLaMA-2-7B [61]. After end-to-end instruction tuning, we evaluate baseline models to assess their effectiveness.

**Evaluation Metrics.** The evaluation metrics include both traditional and GPT metrics. Traditional metrics, such as CiDEr [63], METEOR [5], Acc@0.25IoU [11], and so on, are used to measure the model’s performance on specific tasks. For a more comprehensive evaluation of the models’ instruction-following abilities, we employ GPT-4 to assess the quality of the different variants’ responses. Specifically, we provide GPT-4 with the answers generated by different variant models, the reference answers, and evaluation requirements. GPT-4 evaluates these responses and assigns a score ranging from 0 to 100. A higher average score indicates better performance of the model. To improve the robustness of our evaluation, we conduct three assessments with GPT-4 and use the average score as final result. Furthermore, we request GPT-4 to provide justifications for the scoring results, which helps us better judge the validity of the evaluation.

**Implementations.** Following previous works in 3D learning [15, 44], we down-sample each 3D scene to 40,000 points as scene input. For the PointNet++-based encoder, we initialize it with the checkpoint obtained from Depth Contrast [76]. As for the Transformer-based encoder, we employ the checkpoint from Vote2Cap-DETR [15]. Additionally, we use the pre-trained encoder ViT-L/14 [54] as the image encoder. We train all the baseline models using the Adam optimizer [42] with a cosine annealing scheduler where the learning rate decays from  $10^{-5}$  to  $10^{-6}$ . The model comprises roughly 52 million trainable parameters, accounting for less than 1% of the frozen LLM backbone’s (LLaMA-2-7B [61]) parameter.

### 5.2 Quantitative Evaluation

**Main Results.** As shown in Tab. 2, we comprehensively evaluate four variants and reported the quantitative results across five tasks: Dense Captioning, Visual

**Table 2:** Benchmark for Dense Caption, Visual Question Answering, Multi-region Reasoning, Embodied Question Answering, Embodied Planning. We present the performance of baseline methods on the evaluation dataset.  $\uparrow$  means the higher, the better.

3D Vision Encoder	LLM Decoder	BLEU-1 $\uparrow$	BLEU-2 $\uparrow$	BLEU-3 $\uparrow$	BLEU-4 $\uparrow$	ROUGE $\uparrow$	METEOR $\uparrow$	CIDEr $\uparrow$
<b>Dense Caption</b>								
Pointnet++ [53]	LLaMA-2-7B [61]	3.06	0.89	0.34	0.00	11.64	5.26	17.99
	OPT-6.7B [75]	6.47	2.78	1.29	0.51	20.73	9.55	41.80
Transformer [62]	LLaMA-2-7B [61]	8.61	3.54	1.07	0.00	15.83	13.01	24.81
	OPT-6.7B [75]	<b>17.49</b>	<b>7.67</b>	<b>3.43</b>	<b>1.80</b>	<b>24.01</b>	<b>13.76</b>	<b>50.62</b>
<b>Visual Question Answering</b>								
Pointnet++ [53]	LLaMA-2-7B [61]	31.92	26.56	22.65	19.49	43.27	20.71	183.06
	OPT-6.7B [75]	40.70	34.47	29.95	26.11	46.39	22.49	256.98
Transformer [62]	LLaMA-2-7B [61]	52.75	45.62	40.16	35.65	55.54	29.28	307.44
	OPT-6.7B [75]	<b>57.60</b>	<b>51.05</b>	<b>46.11</b>	<b>41.97</b>	<b>61.47</b>	<b>31.33</b>	<b>384.78</b>
<b>Multi-region Reasoning</b>								
Pointnet++ [53]	LLaMA-2-7B [61]	32.37	25.09	19.70	15.28	42.94	21.07	150.57
	OPT-6.7B [75]	33.60	28.68	24.26	20.27	39.58	20.73	215.23
Transformer [62]	LLaMA-2-7B [61]	41.42	35.92	31.34	27.43	45.28	25.19	251.00
	OPT-6.7B [75]	<b>45.75</b>	<b>39.32</b>	<b>33.90</b>	<b>29.61</b>	<b>53.93</b>	<b>29.77</b>	<b>311.35</b>
<b>Embodied Question Answering</b>								
Pointnet++ [53]	LLaMA-2-7B [61]	22.16	17.05	13.45	10.19	33.90	14.81	137.03
	OPT-6.7B [75]	37.72	28.93	23.20	17.80	43.87	22.84	184.93
Transformer [62]	LLaMA-2-7B [61]	38.77	30.11	24.43	18.73	40.58	20.18	154.21
	OPT-6.7B [75]	<b>48.17</b>	<b>39.19</b>	<b>33.18</b>	<b>27.22</b>	<b>52.95</b>	<b>26.81</b>	<b>240.51</b>
<b>Embodied Planning</b>								
Pointnet++ [53]	LLaMA-2-7B [61]	31.92	26.56	22.65	19.49	43.27	20.71	183.06
	OPT-6.7B [75]	39.63	33.52	28.22	23.58	42.87	21.31	39.35
Transformer [62]	LLaMA-2-7B [61]	35.38	29.10	24.53	21.04	40.52	20.38	85.85
	OPT-6.7B [75]	<b>56.72</b>	<b>51.40</b>	<b>46.89</b>	<b>42.88</b>	<b>60.93</b>	<b>32.28</b>	<b>256.59</b>

Question Answering, Multi-region Reasoning, Embodied Question Answering, and Embodied Planning. We employed BLEU 1-4 [49], ROUGE-L [39], METEOR [5], and CIDEr [63] as evaluation metrics.

Analyzing the results, one can see that when using the same language decoder, variants utilizing transformer-based encoder [47] outperform those based on Pointnet++ [53] in four out of five tasks. Upon examining the impact of different language decoders while keeping a constant 3D encoder, it is evident that the OPT-6.7B-based decoder [75] achieves superior performance across the majority of metrics, compared to LLaMA-2-7B-based decoder [61]. Overall, the variant with transformer-based encoder [47] and OPT-6.7B-based decoder [75] outperforms other variants across all evaluated tasks. It is noteworthy that the all of variants exhibit lower performance on DC compared to other tasks. This can be attributed to the unique challenge posed by the DC task in M3DBench, which involves both brief captions and detailed descriptions at the object level. Moreover, the suboptimal performance of current baseline models across various tasks offers potential direction for further development of 3D MLMs. For instance, improving the performance of MLMs on benchmark tasks is crucial for scene understanding, reasoning, and planning, and we leave them for future work to explore. In the supplementary, we provide further experimental results for other tasks.

### 5.3 Insightful Analyses

*Are Interleaved Multi-modal Instructions Superior to Language-Only Instructions?* As illustrated in Tab. 3, it is evident that compared to language-only

**Table 3:** Verifying the efficacy of interleaved multi-modal instructions. We use transformer-based [62] encoder and OPT-based [75] decoder, using ROUGE [39] metric for assessing.

Instruction Type	Dense Caption	Visual QA	Multi-region Reasoning	Embodied QA	Embodied Planning
Language-only	22.38	59.72	45.99	47.95	50.77
Multi-modal	<b>24.01</b>	<b>61.47</b>	<b>53.93</b>	<b>52.95</b>	<b>60.93</b>

**Table 4:** Zero-shot results on Embodied Question Answering (EQA) and Embodied Planning (EP). For held-out evaluation, we demonstrate the performance of baseline methods on two tasks. Notably, we find that leveraging LLaMA-2 [61] as the language decoder exhibits superior zero-shot generalization compared to OPT [75].

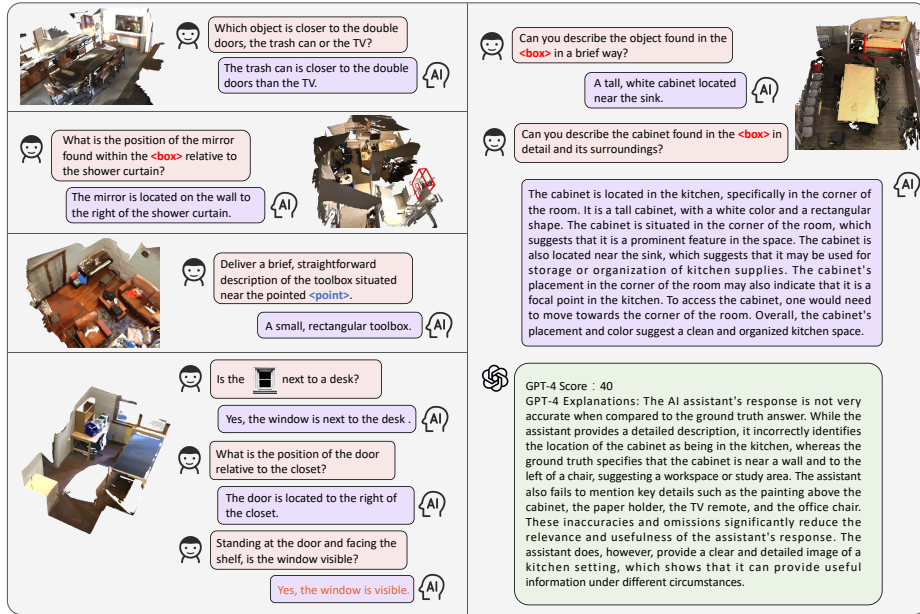
LLM Decoder	BLEU-1 $\uparrow$	BLEU-2 $\uparrow$	BLEU-3 $\uparrow$	BLEU-4 $\uparrow$	ROUGE $\uparrow$	METEOR $\uparrow$	CIDEr $\uparrow$
<i>Embodied Question Answering</i>							
OPT-6.7B [75]	12.04	8.94	7.12	5.39	16.87	11.09	57.60
LLaMA-2-7B [61]	<b>21.06</b>	<b>13.10</b>	<b>8.85</b>	<b>5.79</b>	<b>25.28</b>	<b>13.54</b>	<b>64.80</b>
<i>Embodied Planning</i>							
OPT-6.7B [75]	5.44	3.49	2.19	1.58	11.52	8.04	22.59
LLaMA-2-7B [61]	<b>20.80</b>	<b>11.75</b>	<b>6.32</b>	<b>3.69</b>	<b>17.83</b>	<b>12.08</b>	<b>28.73</b>

instructions, incorporating multi-modal prompts significantly boosts the model’s performance across a range of tasks. Notably, enhancements are observed in dense captioning (+1.63%), visual question answering (+1.75%), multi-region reasoning (+9.64%), embodied question answering (+5%), and embodied planning (+9.53%). This improved performance indicates the advantage of multi-modal instructions in enriching the model’s comprehension and execution of human queries, demonstrating their superiority over language-only instructions.

*How Does the Model Perform under Zero-Shot Scenarios?* To assess the zero-shot performance of baseline models, we partition M3DBench into held-in datasets for training and held-out datasets for evaluation. Analyzing the results presented in Tab. 4, we draw three key insights: 1), multi-modal instruction tuning enables the model to reason on new tasks; 2), the LLaMA-based model [61] surpasses the OPT-based model [75] in zero-shot generalization across both Embodied Question Answering and Embodied Planning tasks; 3), there still exist performance gaps compared to results obtained through full-supervised instruction fine-tuning (detailed in Tab. 2); These findings indicate that through multi-modal instruction tuning on M3DBench, model demonstrates reasoning abilities on tasks that it has not encountered before.

## 5.4 Qualitative Results

We showcase some qualitative examples of our baseline model on the evaluation dataset in Fig. 3. One can see that our proposed method, trained on M3DBench, is capable of performing corresponding tasks under a variety of interleaved multi-modal instructions.



**Fig. 3:** Qualitative Results. We provide visualization results on various 3D-centric tasks in diverse 3D environments. Orange highlights the wrong answer.

## 6 Discussion

**Limitation.** We provide a baseline method capable of handling multi-modal instructions, which has not been previously explored. Exploring the design of a better architecture to efficiently extract information from interleaved multi-modal instructions is meaningful. Additionally, exploring novel training strategies is worth further investigation.

**Conclusion.** In this paper, we present M3DBench, a comprehensive multi-modal 3D instruction-following dataset, designed to facilitate the development of MLMs in the 3D domain. M3DBench encompasses a wide range of 3D vision-centric tasks and over 320K pairs of 3D instruction-following pairs, covering fundamental functionalities such as visual perception, scene understanding, spatial reasoning, planning, and navigation. Additionally, M3DBench introduces a novel multi-modal prompting scheme, interweaving language instruction with user clicks, pointed regions, images, and other visual prompts. Comprehensive quantitative and qualitative results demonstrate that models trained with M3DBench can successfully follow human instructions and complete 3D visual-related tasks. We hope that our proposed multi-modal 3D instruction dataset, baseline model, and benchmarks will inspire and fuel future explorations in the field of 3D MLMs.

## Acknowledgements

This work is supported by National Key Research and Development Program of China (No. 2022ZD0160101), National Natural Science Foundation of China (No. 62071127 and 62101137), Shanghai Natural Science Foundation (No. 23ZR1402900), Shanghai Municipal Science and Technology Major Project (No.2021SHZDZX0103), the A\*STAR AME Programmatic Funding A18A2b0046, the RobotHTPO Seed Fund under Project C211518008, the 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. Achlioptas, P., Abdelreheem, A., Xia, F., Elhoseiny, M., Guibas, L.: Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I* 16. pp. 422–440. Springer (2020)
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems* **35**, 23716–23736 (2022)
3. Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., Van Den Hengel, A.: Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 3674–3683 (2018)
4. Azuma, D., Miyanishi, T., Kurita, S., Kawanabe, M.: Scanqa: 3d question answering for spatial scene understanding. In: *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 19129–19139 (2022)
5. Banerjee, S., Lavie, A.: 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*. pp. 65–72 (2005)
6. Bianco, S., Celona, L., Donzella, M., Napoletano, P.: Improving image captioning descriptiveness by ranking and llm-based fusion. *arXiv preprint arXiv:2306.11593* (2023)
7. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. *Advances in neural information processing systems* **33**, 1877–1901 (2020)
8. Cai, D., Zhao, L., Zhang, J., Sheng, L., Xu, D.: 3djcg: A unified framework for joint dense captioning and visual grounding on 3d point clouds. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 16464–16473 (2022)
9. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158* (2017)
10. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012* (2015)

11. Chen, D.Z., Chang, A.X., Nießner, M.: Scanrefer: 3d object localization in rgb-d scans using natural language. In: European conference on computer vision. pp. 202–221. Springer (2020)
12. Chen, D.Z., Wu, Q., Nießner, M., Chang, A.X.: D 3 net: A unified speaker-listener architecture for 3d dense captioning and visual grounding. In: European Conference on Computer Vision. pp. 487–505. Springer (2022)
13. Chen, F., Han, M., Zhao, H., Zhang, Q., Shi, J., Xu, S., Xu, B.: X-llm: Bootstrapping advanced large language models by treating multi-modalities as foreign languages. arXiv preprint arXiv:2305.04160 (2023)
14. Chen, S., Chen, X., Zhang, C., Li, M., Yu, G., Fei, H., Zhu, H., Fan, J., Chen, T.: Ll3da: Visual interactive instruction tuning for omni-3d understanding, reasoning, and planning (2023)
15. Chen, S., Zhu, H., Chen, X., Lei, Y., Yu, G., Chen, T.: End-to-end 3d dense captioning with vote2cap-detr. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11124–11133 (2023)
16. Chen, S., Zhu, H., Li, M., Chen, X., Guo, P., Lei, Y., Yu, G., Li, T., Chen, T.: Vote2cap-detr++: Decoupling localization and describing for end-to-end 3d dense captioning (2023)
17. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015)
18. Chen, Z., Gholami, A., Nießner, M., Chang, A.X.: Scan2cap: Context-aware dense captioning in rgb-d scans. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3193–3203 (2021)
19. Chen, Z., Hu, R., Chen, X., Nießner, M., Chang, A.X.: Unit3d: A unified transformer for 3d dense captioning and visual grounding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18109–18119 (2023)
20. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., et al.: Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023) (2023)
21. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5828–5839 (2017)
22. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning (2023)
23. Guo, J., Li, J., Li, D., Tiong, A.M.H., Li, B., Tao, D., Hoi, S.: From images to textual prompts: Zero-shot visual question answering with frozen large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10867–10877 (2023)
24. Guo, Z., Zhang, R., Zhu, X., Tang, Y., Ma, X., Han, J., Chen, K., Gao, P., Li, X., Li, H., 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)
25. Hong, Y., Zhen, H., Chen, P., Zheng, S., Du, Y., Chen, Z., Gan, C.: 3d-llm: Injecting the 3d world into large language models. arXiv preprint arXiv:2307.12981 (2023)
26. Hu, W., Xu, Y., Li, Y., Li, W., Chen, Z., Tu, Z.: Bliva: A simple multimodal llm for better handling of text-rich visual questions. arXiv preprint arXiv:2308.09936 (2023)



27. Huang, J., Yong, S., Ma, X., Linghu, X., Li, P., Wang, Y., Li, Q., Zhu, S.C., Jia, B., Huang, S.: An embodied generalist agent in 3d world. arXiv preprint arXiv:2311.12871 (2023)
28. Huang, S., Chen, Y., Jia, J., Wang, L.: Multi-view transformer for 3d visual grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15524–15533 (2022)
29. Jiang, B., Chen, X., Liu, W., Yu, J., Yu, G., Chen, T.: Motiongpt: Human motion as a foreign language. Advances in Neural Information Processing Systems **36** (2024)
30. Jiang, B., Chen, X., Zhang, C., Yin, F., Li, Z., Yu, G., Fan, J.: Motion-chain: Conversational motion controllers via multimodal prompts. arXiv preprint arXiv:2404.01700 (2024)
31. Kim, W., Son, B., Kim, I.: Vilt: Vision-and-language transformer without convolution or region supervision. In: International Conference on Machine Learning. pp. 5583–5594. PMLR (2021)
32. Krantz, J., Wijmans, E., Majumdar, A., Batra, D., Lee, S.: Beyond the navigraph: Vision-and-language navigation in continuous environments. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16. pp. 104–120. Springer (2020)
33. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision **123**, 32–73 (2017)
34. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25** (2012)
35. Li, B., Zhang, Y., Chen, L., Wang, J., Pu, F., Yang, J., Li, C., Liu, Z.: Mimic-it: Multi-modal in-context instruction tuning. arXiv preprint arXiv:2306.05425 (2023)
36. Li, B., Zhang, Y., Chen, L., Wang, J., Yang, J., Liu, Z.: Otter: A multi-modal model with in-context instruction tuning. arXiv preprint arXiv:2305.03726 (2023)
37. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
38. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022)
39. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
40. Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., Wang, L.: Mitigating hallucination in large multi-modal models via robust instruction tuning (2023)
41. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. arXiv preprint arXiv:2304.08485 (2023)
42. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
43. Ma, X., Yong, S., Zheng, Z., Li, Q., Liang, Y., Zhu, S.C., Huang, S.: Sqa3d: Situated question answering in 3d scenes. arXiv preprint arXiv:2210.07474 (2022)
44. Misra, I., Girdhar, R., Joulin, A.: An end-to-end transformer model for 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2906–2917 (2021)
45. Misra, I., Girdhar, R., Joulin, A.: An end-to-end transformer model for 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2906–2917 (2021)

46. OpenAI, R.: Gpt-4 technical report. arxiv 2303.08774. View in Article **2** (2023)
47. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* **35**, 27730–27744 (2022)
48. Pan, X., Xia, Z., Song, S., Li, L.E., Huang, G.: 3d object detection with pointformer. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 7463–7472 (2021)
49. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. pp. 311–318 (2002)
50. Parelli, M., Delitzas, A., Hars, N., Vlassis, G., Anagnostidis, S., Bachmann, G., Hofmann, T.: 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*. pp. 5606–5611 (2023)
51. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
52. Qi, C.R., Litany, O., He, K., Guibas, L.J.: Deep hough voting for 3d object detection in point clouds. In: *proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 9277–9286 (2019)
53. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems* **30** (2017)
54. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: *International conference on machine learning*. pp. 8748–8763. PMLR (2021)
55. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* **21**(1), 5485–5551 (2020)
56. Rotstein, N., Bensaïd, D., Brody, S., Ganz, R., Kimmel, R.: Fusecap: Leveraging large language models to fuse visual data into enriched image captions. arXiv preprint arXiv:2305.17718 (2023)
57. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems* **35**, 25278–25294 (2022)
58. Schulman, J., Zoph, B., Kim, C., Hilton, J., Menick, J., Weng, J., Uribe, J.F.C., Fedus, L., Metz, L., Pokorny, M., et al.: Chatgpt: Optimizing language models for dialogue. OpenAI blog (2022)
59. Song, S., Lichtenberg, S.P., Xiao, J.: Sun rgb-d: A rgb-d scene understanding benchmark suite. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 567–576 (2015)
60. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
61. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)

62. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. *Advances in neural information processing systems* **30** (2017)
63. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 4566–4575 (2015)
64. Wang, T., Zhang, J., Fei, J., Ge, Y., Zheng, H., Tang, Y., Li, Z., Gao, M., Zhao, S., Shan, Y., et al.: Caption anything: Interactive image description with diverse multimodal controls. *arXiv preprint arXiv:2305.02677* (2023)
65. Wang, Z., Huang, H., Zhao, Y., Zhang, Z., Zhao, Z.: Chat-3d: Data-efficiently tuning large language model for universal dialogue of 3d scenes. *arXiv preprint arXiv:2308.08769* (2023)
66. Wu, Y., Cheng, X., Zhang, R., Cheng, Z., Zhang, J.: Eda: Explicit text-decoupling and dense alignment for 3d visual grounding. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 19231–19242 (2023)
67. Yan, X., Yuan, Z., Du, Y., Liao, Y., Guo, Y., Li, Z., Cui, S.: Clevr3d: Compositional language and elementary visual reasoning for question answering in 3d real-world scenes. *arXiv preprint arXiv:2112.11691* (2021)
68. Yang, Z., Zhang, S., Wang, L., Luo, J.: Sat: 2d semantics assisted training for 3d visual grounding. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 1856–1866 (2021)
69. Yin, F., Chen, X., Zhang, C., Jiang, B., Zhao, Z., Fan, J., Yu, G., Li, T., Chen, T.: Shapegpt: 3d shape generation with a unified multi-modal language model. *arXiv preprint arXiv:2311.17618* (2023)
70. Yin, Z., Wang, J., Cao, J., Shi, Z., Liu, D., Li, M., Sheng, L., Bai, L., Huang, X., Wang, Z., et al.: Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *arXiv preprint arXiv:2306.06687* (2023)
71. Yuan, Z., Yan, X., Li, Z., Li, X., Guo, Y., Cui, S., Li, Z.: Toward explainable and fine-grained 3d grounding through referring textual phrases. *arXiv preprint arXiv:2207.01821* (2022)
72. Yuan, Z., Yan, X., Liao, Y., Guo, Y., Li, G., Cui, S., Li, Z.: X-trans2cap: Cross-modal knowledge transfer using transformer for 3d dense captioning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 8563–8573 (2022)
73. Yuan, Z., Yan, X., Liao, Y., Zhang, R., Wang, S., Li, Z., Cui, S.: Instancerefer: Cooperative holistic understanding for visual grounding on point clouds through instance multi-level contextual referring. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 1791–1800 (2021)
74. Yuan, Z., Yan, X., Liao, Y., Zhang, R., Wang, S., Li, Z., Cui, S.: Instancerefer: Cooperative holistic understanding for visual grounding on point clouds through instance multi-level contextual referring. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 1791–1800 (2021)
75. Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., et al.: Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068* (2022)
76. Zhang, Z., Girdhar, R., Joulin, A., Misra, I.: Self-supervised pretraining of 3d features on any point-cloud. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 10252–10263 (2021)
77. Zhao, L., Cai, D., Sheng, L., Xu, D.: 3dvg-transformer: Relation modeling for visual grounding on point clouds. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 2928–2937 (2021)

78. Zhu, Z., Ma, X., Chen, Y., Deng, Z., Huang, S., Li, Q.: 3d-vista: Pre-trained transformer for 3d vision and text alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2911–2921 (2023)