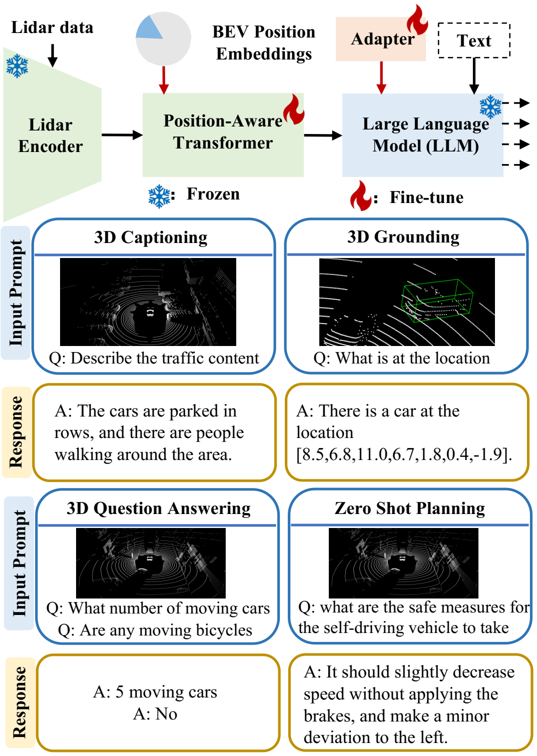
LiDAR-LLM: Exploring the Potential of Large Language Models for 3D LiDAR Understanding

Senqiao Yang[1*], Jiaming Liu[1,2*], Renrui Zhang [3*†], Mingjie Pan[1*], Ziyu Guo[3], Xiaoqi Li[1], Zehui Chen[1], Peng Gao[4], Hongsheng Li[3], Yandong Guo[2], Shanghang Zhang [1 ‡]

[1]State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University
[2]AI2Robotics
[3] The Chinese University of Hong Kong
[4] Shanghai Artificial Intelligence Laboratory

Abstract

Recently, Large Language Models (LLMs) and Multimodal Large Language Models (MLLMs) have shown promise in instruction following and image understanding. While these models are powerful, they have not yet been developed to comprehend the more challenging 3D geometric and physical scenes, especially when it comes to the sparse outdoor LiDAR data. In this paper, we introduce LiDAR-LLM, which takes raw LiDAR data as input and harnesses the remarkable reasoning capabilities of LLMs to gain a comprehensive understanding of outdoor 3D scenes. The central insight of our LiDAR-LLM is the reformulation of 3D outdoor scene cognition as a language modeling problem, encompassing tasks such as 3D captioning, 3D grounding, 3D question answering, etc. Specifically, due to the scarcity of 3D LiDAR-text pairing data, we introduce a three-stage training strategy and generate relevant datasets, progressively aligning the 3D modality with the language embedding of LLM. Furthermore, we design a Position-Aware Transformer (PAT) to connect the 3D encoder with the LLM, which effectively bridges the modality gap and enhances the LLM's spatial orientation comprehension of visual features. Our experiments demonstrate that LiDAR-LLM effectively comprehends a wide range of instructions related to 3D scenes, achieving a 40.9 BLEU-1 score on the 3D captioning dataset, a Grounded Captioning accuracy of 63.1%, and a BEV mIoU of 14.3%.



Q: Describe the traffic content

Q: What is at the location

A: The cars are parked in rows, and there are people walking around the area.

A: There is a car at the location [8.5,6.8,11.0,6.7,1.8,0.4,-1.9].

Q: What number of moving cars
Q: Are any moving bicycles

Q: what are the safe measures for the self-driving vehicle to take

A: 5 moving cars
A: No

A: It should slightly decrease speed without applying the brakes, and make a minor deviation to the left.

Figure 1: **Characteristics of LiDAR-LLM.** Our proposed LiDAR-LLM takes 3D LiDAR data as input and aligns the 3D modality with the language embedding space, leveraging the exceptional reasoning capabilities of LLMs to understand outdoor 3D scenes. The bottom part showcases examples derived from our generated or employed LiDAR-text data, covering a spectrum of 3D-related tasks.

Introduction

Recently, large language models (LLMs) (Touvron et al. 2023; OpenAI 2023; Brown et al. 2020) have demonstrated significant capabilities in complex reasoning and robust conversational abilities in the field of natural language processing. Building upon LLMs, Multimodal Large Language Models (MLLMs) (Liu et al. 2024; Li et al. 2023; et al 2022; Lin et al. 2023; Wang et al. 2023a), such as BLIP-2 and Flamingo, have been introduced. These models take in more modality (e.g., 2D images) as input, enabling LLMs to discuss and comprehend the visual scene. Despite MLLMs excelling at processing 2D image content, their comprehension of the more challenging 3D real-world scenes remains an open question. Understanding 3D scenes holds importance for various applications, including autonomous driving (Arnold et al. 2019; Chen et al. 2017) and robotics

[*]Equal contribution
[†]Project leader
[‡]Corresponding author

(Dhamo et al. 2021; Yao et al. 2018; Wang et al. 2024), due to the wealth of spatial information in 3D data.

Existing 3D understanding methods (Yang et al. 2021; Jiao et al. 2022; Parelli et al. 2023; Azuma et al. 2022; Ma et al. 2022) often lack sufficient generalization capabilities when faced with unseen scenarios. They are limited in expressing specific downstream tasks in a manner comprehensible to humans, such as generating scene captioning and question answering. Therefore, recent works (Wang et al. 2023b; Hong et al. 2023; Qi et al. 2024) take indoor 3D point clouds as input and leverage the powerful capabilities of LLMs to analyze them, aligning the 3D features with the textual features of LLMs. However, these approaches still encounter challenges when dealing with 3D outdoor LiDAR data. Specifically, due to the sparse property of LiDAR data, rendering the 3D data into multi-view images (Hong et al. 2023) results in poor rendering and feature extraction quality. Additionally, methods like (Guo et al. 2023; Qi et al. 2024) can only encode single object-level point clouds instead of understanding complex outdoor scenes.

In this paper, as shown in Figure 1, we introduce LiDAR-LLM, a novel approach that harnesses the reasoning capabilities of LLMs to comprehensively understand outdoor 3D scenes. The LiDAR-LLM architecture comprises a 3D LiDAR encoder, an alignment transformer, and an LLM, *e.g.*, LLaMA (Touvron et al. 2023). The key insight of LiDAR-LLM lies in redefining the problem of 3D scene cognition through interpretative language modeling. However, the introduction of LLMs for perceiving outdoor 3D scenes faces two challenges: **1)** In contrast to the abundant availability of image-text paired data (Sharma et al. 2018; Schuhmann et al. 2022; Changpinyo et al. 2021), 3D LiDAR-text paired data is exceedingly rare, and accessible multimodal encoders (e.g., CLIP (Radford et al. 2021)) are lacking. **2**)3D LiDAR is sparser than indoor point cloud and encompasses intricate geometric relationships with a variety of objects.

To tackle these challenges, we generate the required datasets and introduce a three-stage training strategy, including cross-modal alignment, perception, and high-level instruction. This strategy gradually transfers 3D representations into the language feature space, unleashing LLMs' reasoning capabilities for 3D scenes. In the first stage, we utilize MLLMs (Zhang et al. 2023; Li et al. 2023) and GPT-4 (OpenAI 2023) to facilitate communication between multi-view images and language within the nuScenes dataset (Caesar et al. 2020), which includes paired 3D LiDAR data for each scene. This process automatically generates a caption dataset of 420K LiDAR-text pairs, enabling the cross-modality alignment of 3D LiDAR features with LLM word embeddings. In the second stage, recognizing that perception is crucial for 3D scene understanding, we integrate 3D bounding boxes into the question-answer text, creating a 280K LiDAR grounding dataset. We then apply an object-centric learning strategy to equip the model with 3D perception capabilities. Finally, in the third stage, we efficiently fine-tune LiDAR-LLM on high-level instruction datasets (Qian et al. 2023; Contributors 2023), enhancing its capabilities for various 3D downstream tasks, such as autonomous driving prediction and planning. To more effectively bridge the modality gap between 3D LiDAR and text, we design a Position-Aware Transformer (PAT) that connects the 3D LiDAR encoder with the LLM, explicitly injecting BEV position embedding into the 3D features. Combined with the three-stage training strategy, PAT enhances the model's comprehension of the spatial orientation. In summary, our contributions are as follows:

- We propose LiDAR-LLM framework, which takes 3D LiDAR data and language as input, harnessing the reasoning capabilities of LLMs to understand outdoor 3D scenes. LiDAR-LLM is capable of performing tasks such as 3D captioning, 3D grounding, 3D question answering, high-level planning, and more.
- We introduce a three-stage training strategy for gradually transferring 3D representations into the text feature space, which involves cross-modal alignment, perception, and high-level instruction.
- To facilitate training, we collect a set of LiDAR-text paired datasets, comprising 420K 3D captioning data (nu-Caption) and 280K 3D grounding data (nu-Grounding). These datasets will be released for research.
- We specially design a Position-Aware Transformer (PAT) that connects the 3D LiDAR encoder with the LLM, bridging the modality gap, and enhancing the model's comprehension of the spatial orientation.

Method

Overview. The overall framework of LiDAR-LLM is presented in Fig. 2. The core concept involves transforming the sparse and intricate geometric LiDAR data into the representation space understandable by Large-Language Models (LLMs). However, the integration of LLMs to comprehend outdoor 3D scenes faces two challenges: (1) Unlike the abundance of available image-text paired data, 3D LiDAR-text paired data is exceptionally scarce; and (2) Sparse LiDAR data involves diverse objects and intricate geometric relationships among them. To overcome these challenges, we propose a three-stage training strategy and generate required LiDAR-text paired data to transfer 3D representations into the feature space of LLMs. Through this process, the alignment, perception, and common-sense reasoning capabilities of LLMs on 3D LiDAR data are gradually empowered. LiDAR-LLM can perform diverse tasks in the LiDAR modality and handle complex cross-modal scenarios at both the scene and instance levels. Furthermore, we introduce the Position-Aware Transformer (PAT) module, which connects the 3D encoder with the LLM. This module incorporates BEV position embeddings to enhance the spatial orientation understanding of the LLM. Hence, the specially designed training strategy, along with the PAT, jointly empowers LiDAR-LLM to attain a comprehensive understanding of intricate spatial information in outdoor 3D scenes. Finally, due to space limitations, we have included the detailed related works in Appendix D.

Model Architecture

Given a LiDAR input $L \in \mathbb{R}^{n\times3}$, where n is the number of points, VoxelNet (Zhou and Tuzel 2018) is employed to

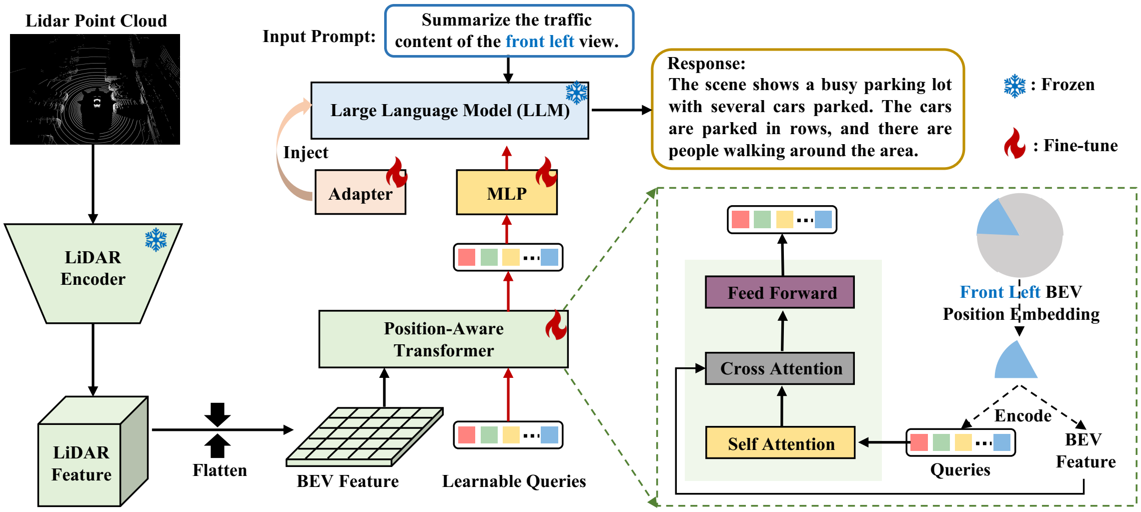

Lidar Point Cloud
Input Prompt:
Summarize the traffic content of the front left view.
Large Language Model (LLM)
Response:
The scene shows a busy parking lot with several cars parked. The cars are parked in rows, and there are people walking around the area.
❄ : Frozen
🔥 : Fine-tune
Inject
Adapter
MLP
LiDAR Encoder
Position-Aware Transformer
LiDAR Feature
Flatten
BEV Feature
Learnable Queries
Feed Forward
Cross Attention
Self Attention
Front Left BEV Position Embedding
Encode
Queries
BEV Feature

Figure 2: **Overview of our LiDAR-LLM framework.** The initial column showcases our 3D feature extractor, which processes the LiDAR point cloud input to derive a 3D voxel feature. Subsequently, the feature is flattened along the z-axis to produce the BEV feature. The Position-Aware Transformer (PAT) accepts BEV feature and learnable queries as input, with the output queries serving as soft prompt input to the frozen LLM. In the PAT, we introduce BEV position embeddings into the BEV feature along with corresponding queries to enhance the spatial orientation representation. This framework aligns the LiDAR modality with the language embedding space, empowering the LLM for a comprehensive understanding of outdoor 3D scenes.

extract its 3D voxel feature. Subsequently, considering the computational cost, we flatten the feature along the z-axis to generate the bird's-eye view (BEV) feature. Simultaneously, for the text input T with a maximum of m characters, LLaMA (Touvron et al. 2023) is utilized to extract text features. With the BEV feature $\mathcal{F}_v \in \mathbb{R}^{512\times180\times180}$ along with the text feature $\mathcal{F}_t \in \mathbb{R}^{576\times768}$ (where 768 is the dimension of the feature), our objective is to project these LiDAR BEV features into the word embedding space of a pre-trained LLaMA through our proposed Position-Aware Transformer (PAT). This alignment is crucial for conducting multi-modal understanding and generating accurate answers in 3D downstream tasks. During training, we only fine-tune the injected adapters (Hu et al. 2021) in the LLaMA and PAT module while freezing the major parameters. This aims to preserve the powerful feature extraction and reasoning ability of existing modules and further equip the model with capabilities in understanding 3D LiDAR scenes.

PAT design. In the right part of Fig. 2, the input to the PAT includes a set of K learnable query embeddings, with K set to 576 for convenient projection into the word embedding space of the LLM. These queries interact with the BEV feature through a cross-attention mechanism, where the learnable queries serve as the query and the BEV features serve as the key and value. The PAT produces an output comprising K encoded visual vectors, one for each query embedding. These vectors then undergo processing through a multi-layer perceptron (MLP) and are subsequently fed into the frozen LLM. However, outdoor LiDAR data, demands a comprehensive understanding of the orientation relationships between diverse objects and the ego car. Therefore, we introduce BEV position embeddings for the BEV feature, aiming to explicitly enhance spatial orientation comprehension. Specifically, we first construct the BEV position embedding $\mathcal{B}_p \in \mathbb{R}^{c\times v}$ with zero initial parameters, where c represents the embedding dimension and v represents the different orientations in the BEV scene. This division depends on the camera setup used during the original outdoor dataset collection. For example, nuScenes (Caesar et al. 2020) splits the BEV scene into six views (v=6), including front, front right, front left, back, back right, and back left views, while Waymo (Sun et al. 2020) contains five views (v=5). If a finer view division is required, our PAT design can efficiently adapt to accommodate it. Following this, we can clearly describe the spatial relationship between objects and the ego car, avoiding ambiguous descriptions, such as whether a car is positioned to the left back or the back of the ego car. When dealing with a question related to a specific view during the cross-modal alignment stage, we inject the corresponding BEV position embedding into both the BEV feature and queries. For instance, when training a caption sample related to the front left view, we inject only the front left part of the BEV position embedding $\mathcal{B}_p \in \mathbb{R}^{c\times 1}$ into the front left part of the BEV feature and the entire set of queries (as shown in Fig. 2). Moreover, when objects are situated at the junction of two views, they are depicted in both adjacent views, avoiding loss of information.

Note that our PAT design is intended to enhance our model's understanding of spatial orientation. As described in Sec. **Three-stage training strategy**, we incorporate separate BEV position embeddings only when training individual views during the initial cross-modal alignment stage. For training on panoramic scenes and subsequent tasks (i.e., perception and high-level instruction), we include the entire set of BEV position embeddings, enabling the model to perform end-to-end training with flexible language descriptions. Finally, we provide a quantitative analysis to validate the effectiveness of the PAT module in Sec. **Ablation Study**.

Three-Stage Training Strategy

In this section, we demonstrate how we empower LLMs with the capabilities to comprehend 3D LiDAR data and uniformly complete extensive 3D tasks. We introduce a three-stage training strategy and generate relevant datasets, gradually transferring 3D representations into the text feature space. Three stages contain cross-modal alignment, perception, and high-level instruction.

Cross-Modal Alignment (3D Captioning): To effectively address numerous 3D downstream tasks, the model needs to have a thorough understanding of the LiDAR scene. Scene captioning serves as a logical approach to enable the model to capture essential information and details from LiDAR data by integrating the entire 3D scene into LLMs.

However, there is a lack of open-source caption datasets pairing LiDAR data with text descriptions. Instead, datasets (Caesar et al. 2020; Sun et al. 2020) containing paired multi-view images and LiDAR data, are available. This motivates us to leverage multi-view images to automatically generate text descriptions, which can then be paired with corresponding LiDAR data. By employing powerful off-the-shelf 2D MLLMs (Li et al. 2023; Zhang et al. 2023), we input multi-view images to generate captions for each view, thereby obtaining single-view LiDAR-text paired data. Subsequently, we can obtain panoramic scene LiDAR-text paired data by merging multi-view captions. Nevertheless, the captions for LiDAR and 2D multi-view images are not perfectly aligned, as 2D MLLM may provide descriptions related to weather or colors for 2D images, which are not applicable to LiDAR data. To address this inconsistency, we further utilize GPT-4 (OpenAI 2023) to filter out irrelevant captions/words for LiDAR data. The filtering process can be divided into three steps. **1)** We first provide the prompt *"Please remove words and sentences describing the color and weather, ensuring that the meaning remains consistent."* to GPT-4 for filtering. **2)** After the first round, we input the filtered caption and the corresponding original caption back into GPT-4 using the prompt: *"Please check the following two sentences for similarity of meaning..."* This is to ensure that the original meaning of the descriptions is not changed. **3)** We automatically remove captions that alter the meaning and do not involve any manual verification in the process. More details of data generation can be found in Appendix C.

With the collected LiDAR-text caption pairs, our goal is to enable LLM to generate descriptive text conditioned on LiDAR input. We observe that textual captions for LiDAR data tend to be excessively detailed and lengthy due to their intricate geometric structures. Jointly learning overall captions could lead to entanglement in LLM reasoning. To mitigate this, we introduce individual-to-panoramic scene training mechanism. Specifically, we initially train the model to caption an individual view to reduce complexity. Following this, the subsequent step involves instructing the model to understand the entire panoramic scene and generate a global description. By doing so, we align the 3D feature representation with the text feature space of LLM, enabling the model to comprehend the context in the LiDAR data.

Perception: After equipping the model with a panoramic scene understanding, this stage focuses on endowing the model with instance-level 3D perception abilities, which form the foundation for high-level instructional tasks such as planning. Our aim is to align the feature representation of 3D objects with the corresponding text embedding of the LLM. To achieve this, we employ an object-centric learning strategy, ensuring the LiDAR-LLM learns various object details such as quantity, localization, and spatial relations.

Two tasks, visual grounding, and grounded captioning, are designed for this purpose. Objects are first represented as a sequence of discrete tokens, where each object's label and bounding box are extracted. Given a 3D object with its annotations, the category name and locations are encoded into a word embedding using the tokenizer of the pre-trained LLM. Unlike the previous indoor 3D MLLM (Wang et al. 2023b), there is no need to extract each object from the point cloud individually; instead, we achieve object perception across the entire 3D scene. For visual grounding, the model learns to generate location tokens specifying the region position $(x_1, y_1, z_1, x_2, y_2, z_2, \theta)$ based on the LiDAR input and instruction, where θ is the box angle. Grounded Captioning task is positioned as the inverse counterpart to visual grounding. The model is trained to generate descriptive text by leveraging the input LiDAR data and text with location information. The instructions are depicted in Fig. 3.

High-level Instruction: In this stage, after comprehensively understanding the LiDAR scene and equipping the model with basic 3D perception capabilities, we leverage high-level instruction datasets (i.e., nuScenes-QA (Qian et al. 2023) and DriveLM-nuScenes (Contributors 2023)) to further enhance the model's reasoning skills in 3D space. Through the fine-tuning of LiDAR-LLM using these dataset, we not only enhance its proficiency in comprehending a diverse array of instructions but also empower it to generate responses that are both creative and contextually appropriate. Meanwhile, rather than fine-tuning on planning QA data (DriveLM-nuScenes), we directly employ our trained model to infer planning-related questions. Through our proposed three-stage training strategy, LiDAR-LLM develops initial planning capabilities, as demonstrated in Appendix B. Conversely, incorporating DriveLM-nuScenes during High-level Instruction tuning can endow LiDAR-LLM with autonomous driving reasoning abilities in perception, driving behavior, and planning, yielding promising results as evidenced in Sec. **Experiment - High-level Instruction Task**.

Training and Task Inference

LiDAR-LLM undergoes joint fine-tuning with a variety of tasks and datasets, equipping it with a versatile skill set to adeptly handle diverse tasks within complex cross-modal scenarios. In the fine-tuning phase, we perform fine-tuning on a dataset consisting of 700K LiDAR-text pairs generated by us and 750K publicly available datasets (Qian et al. 2023; Contributors 2023). All the training steps are supervised through cross-entropy loss. During inference, our in-

Tasks	Models	BLEU-1	BLEU-4	Bert-F1	Bert-P	Bert-R
3D Captioning	Mini-GPT4	14.97	2.63	85.20	84.38	86.07
	LLaMA-AdapterV2	30.17	7.45	87.98	87.45	88.53
	Ours (Panoramic)	35.36	16.23	88.14	89.67	89.23
	Ours	**40.98**	**19.26**	**90.96**	**91.32**	**90.61**

Table 1: Experimental results on nu-Caption dataset. Our model outperforms all baseline models for all evaluation metrics. (Panoramic) means directly fine-tuning on overall captions, rather than training with individual-to-panoramic scene mechanisms. Bert-F1, -P, and -R represent the F1 score, Precision, and Recall of the Bert score.

Tasks	Models	ACC-19	ACC-5	mIoU-5	mIoU-car
3D Grounding	Mini-GPT4	5.1	21.2	-	-
	LLaMA-AdapterV2	7.1	23.4	-	-
	Ours	**34.4**	**63.1**	**9.9**	**14.3**

Table 2: Experimental results on the nu-Grounding dataset. ACC-19 and ACC-5 denote the mean Top-1 accuracy for scenarios with 19 categories and 5 categories, respectively. The mIoU-5 and mIoU-car are calculated for the 5 categories and the "Car" category, respectively. The fine-grained performances are provided in the Appendix A.

put still consists of LiDAR and question text. LiDAR-LLM is flexible enough to infer each question individually or to infer multiple questions consecutively.

Experiment

In this section, we conduct extensive experiments on our generated datasets, DriveLM-nuScene (Contributors 2023), and NuScenes-QA (Qian et al. 2023). In the next two sections, we first introduce the selected baselines and evaluation metrics, as well as the implementation details. Our main experiments evaluate the model on three tasks in the following sections, including Sec. **3D Captioning**, Sec. **3D Grounding**, and Sec. **High-level Instruction Task**. Finally, in Sec. **Ablation Study** and Sec. **Qualitative Analysis**, we provide a deeper analysis of our approach.

Baselines & Evaluation Metrics

Baselines. To the best of our knowledge, we are the first MLLM to utilize LiDAR data with textual instructions as input and implement a series of outdoor 3D tasks. Since there are no 3D/2D MLLM that can directly process LiDAR data, we project the depth information from LiDAR onto the 2D plane and employ current state-of-the-art (SOTA) 2D MLLM methods, MiniGPT-4 (Zhu et al. 2023), LLaVA1.5 (Liu et al. 2023), and LLaMA-Adapter V2 (Zhang et al. 2023), as our competitive counterparts. More baseline comparison experiments can be found in Appendix A, such as comparisons with 2D MLLM using image input.

Evaluation Metrics. To evaluate language generation in the 3D Captioning Task, we use BLEU (Papineni et al. 2002) and BERT Score (Zhang et al. 2019). For the 3D Grounding Task, we assess grounding ability using classification Top-1 accuracy and BEV mIoU. In NuScene-QA, model performance is measured with Top-1 accuracy, following VQA research practices (Azuma et al. 2022), with separate evaluations for different question types. For DriveLM-nuScene, we use BLEU and BERT Score, as in the 3D Captioning Task, to validate our method's effectiveness.

Implementation Details

Our LiDAR-LLM has three main components: a LiDAR feature extraction backbone, Position-Aware Transformers (PAT), and the LLM. For LiDAR feature extraction, we use the standard 3D detector CenterPoint-Voxel (Zhou and Tuzel 2018), pre-trained on the 3D detection task (Caesar et al. 2020). We input multi-adjacent frame point clouds (9 sweep frames per sample) to encode temporal information. The point cloud range is [-54.0m, -54.0m, -5.0m, 54.0m, 54.0m, 3.0m], and the BEV grid size is [0.6m, 0.6m]. The PAT module uses 576 learnable query tokens with a dimension of 768. For the LLM, we use LLaMA-7B (Touvron et al. 2023) for its efficiency and effectiveness. During three-stage training, we use the Adam optimizer $(\beta_1, \beta_2) = (0.9, 0.999)$ with an initial learning rate of 1e-4, halving it every 2 epochs. We fine-tune the PAT and LLaMA adapters for 6 epochs. All experiments are run on NVIDIA Tesla A100 GPUs.

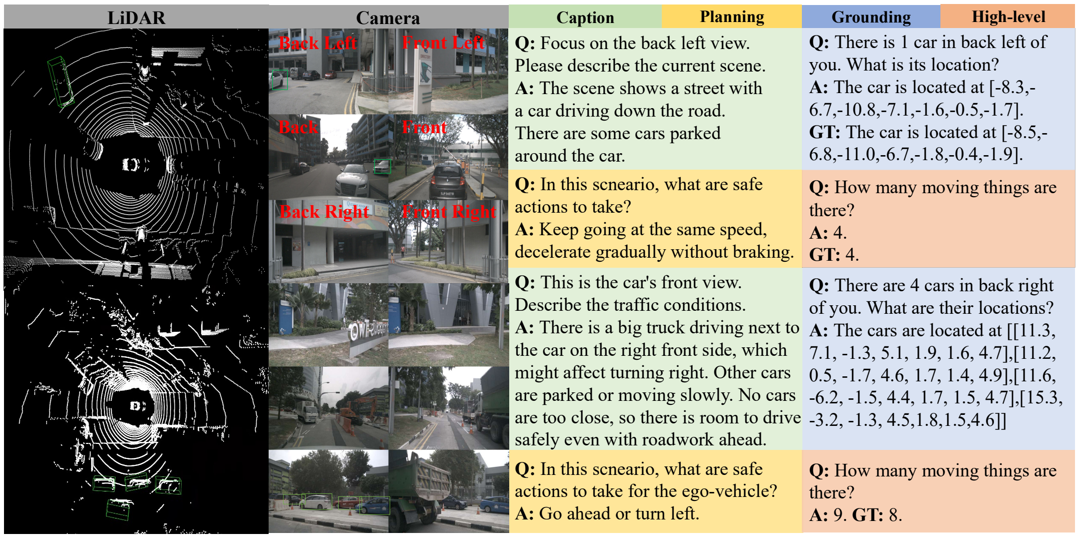
3D Captioning

Dataset Construction. Due to the absence of a caption dataset tailored for LiDAR data, we integrate GPT-4 (OpenAI 2023) and 2D MLLMs (Zhang et al. 2023) to construct a large-scale 3D captioning dataset on nuScenes (named nu-Caption), which consists of 420K high-quality LiDAR-text pairs. In nu-Caption, we employ 348K LiDAR-text pairs for the training set and 72K pairs for the validation set. The caption question covers three aspects, progressing from simple to difficult: 1) the general description of current scenes or traffic conditions, 2) the detailed description of objects and their relationships, 3) the recognition of potential risks on the road. Additional details can be found in Appendix C.

Results Analysis. We then evaluate the methods on our generated nu-Caption dataset and report the result in Table 1. LiDAR-LLM outperforms previous 2D MLLMs for all evaluation metrics. Specifically, our model achieves 19.26% BLEU-4 and 91.32% Bert score precision, surpassing Mini-GPT4, which achieved 2.63% BLEU-4 and 84.38% Bert-P; also get a 11.81% BLEU-4 and 3.87% Bert-P improvement compared to LLaMA-AdapterV2. These results indicate that directly employing 2D MLLMs to understand LiDAR data produces unsatisfactory outcomes, leading to the omission of crucial details in caption descriptions. Meanwhile, we compare our method (individual-to-panoramic scene) with the direct caption training on panoramic scenes. Ours (Panoramic) achieves 16.23% in BLEU-4 (a decrease of 3.03%), indicating that progressively learning from simple to complex scenes benefits the model's comprehension.

3D Grounding

Dataset Construction. Apart from captioning, 3D grounding requires perception capabilities. Using nuScenes anno-



Q: Focus on the back left view. Please describe the current scene.
A: The scene shows a street with a car driving down the road. There are some cars parked around the car.

Q: There is 1 car in back left of you. What is its location?
A: The car is located at [-8.3,-6.7,-10.8,-7.1,-1.6,-0.5,-1.7].
GT: The car is located at [-8.5,-6.8,-11.0,-6.7,-1.8,-0.4,-1.9].

Q: In this scneario, what are safe actions to take?
A: Keep going at the same speed, decelerate gradually without braking.

Q: How many moving things are there?
A: 4.
GT: 4.

Q: This is the car's front view. Describe the traffic conditions.
A: There is a big truck driving next to the car on the right front side, which might affect turning right. Other cars are parked or moving slowly. No cars are too close, so there is room to drive safely even with roadwork ahead.

Q: There are 4 cars in back right of you. What are their locations?
A: The cars are located at [[11.3, 7.1, -1.3, 5.1, 1.9, 1.6, 4.7],[11.2, 0.5, -1.7, 4.6, 1.7, 1.4, 4.9],[11.6, -6.2, -1.5, 4.4, 1.7, 1.5, 4.7],[15.3, -3.2, -1.3, 4.5,1.8,1.5,4.6]]

Q: In this scneario, what are safe actions to take for the ego-vehicle?
A: Go ahead or turn left.

Q: How many moving things are there?
A: 9. **GT:** 8.

Figure 3: Qualitative examples of prompt questions and LiDAR-LLM's prediction. Additional examples are shown in Appendix.

tations, we create a dataset called nu-Grounding with 280K question-answer pairs for visual grounding and grounded captioning tasks. Of these, 232K pairs are for training, and 48K are for validation. For Grounded Captioning, we measure accuracy for all 19 categories (Caesar et al. 2020) and the 5 primary ones: car, pedestrian, bus, truck, and construction vehicle. For Visual Grounding, we predict bounding boxes and calculate BEV mIoU for the "Car" category and the 5 primary categories.

Results Analysis. For Grounded Captioning, as shown in Table 2, our model achieves 63.1% accuracy in scenarios with 5 categories, surpassing LLaMA-Adapter and MiniGPT4 with accuracies of 23.4% and 21.2%, respectively. Meanwhile, when trained and tested on 19 categories, our approach still demonstrates a significant advantage over 2D MLLMs. This indicates that our LiDAR-LLM possesses an understanding of localization and classification information in 3D LiDAR data. For visual grounding, our LiDAR-LLM achieves 14.3% and 9.9% BEV mIoU for the single car category and the 5 primary categories, respectively. The results demonstrate that our approach exhibits basic perceptual capabilities and can generate fine-grained bounding boxes. In this task, our goal is not solely to obtain localization information for objects but also to enhance the model's understanding of the spatial relationships within LiDAR data.

High-Level Instruction Task

Dataset NuScenes-QA (Qian et al. 2023) is a multi-modal VQA benchmark for autonomous driving with five question types: existence, counting, query-object, query-status, and comparison. Questions are categorized by reasoning complexity into zero-hop and one-hop reasoning. DriveLM-nuScene (Contributors 2023) is another dataset for autonomous driving, with annotated QAs. However, some QAs in DriveLM-nuScene target 2D multi-view images with text prompts and 2D coordinates. We automatically filter these out during training. The resulting DriveLM-nuScene dataset, oriented toward LiDAR data, focuses on three reasoning tasks: perception, planning, and driving behavior. Since only the training dataset is public, we split 10% of it to create the validation set.

NuScenes-QA Results Analysis. As shown in Table 3, we first compare our model's performance on NuScenes-QA with different pre-training stages. In Ex1, training LiDAR-LLM from scratch achieves an accuracy of 41.2%, validating the effectiveness of our model design in high-level instruction tasks. Ex2, pre-training on the captioning task, results in a 6.2% accuracy improvement compared to Ex1. Ex3, pre-training on the grounding task, achieves a 5.3% accuracy improvement over Ex1. The results show that when LiDAR-LLM possesses basic 3D scene understanding or perception capabilities, it can more effectively accomplish high-level reasoning tasks. Compared to pre-training on a single task, pre-training on both captioning and grounding tasks (Ex4) shows significant improvement, achieving a total accuracy of 48.6%. This improvement is observed in both zero-hop and one-hop reasoning questions. In addition, as shown in the lower part of Table 3, we compare our model's performance with previous 2D MLLMs, which highlights a significant improvement. These results demonstrate our model's capability to handle various high-level reasoning tasks.

DriveLM-nuScene Results Analysis. As shown in Table 4, we fine-tune LiDAR-LLM on the DriveLM-nuScene dataset, showcasing its performance across three reasoning tasks. LiDAR-LLM achieved a BLEU-1 score of 57.44% in the perception reasoning task, demonstrating its capability to effectively describe high-level perception, which benefits from our proposed 3D perception training stage. Additionally, LiDAR-LLM achieved a BLEU-1 score of 37.47% in the planning task, indicating its strong reasoning abilities and potential in handling planning-related problems, particularly in the field of autonomous driving. Notably, in the driving behavior task, our method attained an unprecedented BLEU score. We attribute this to the numerous behavior

Method	Pretrain	Exist			Count			Object			Status			Comparison			Acc
		H0	H1	All	H0	H1	All	H0	H1	All	H0	H1	All	H0	H1	All	
Ex1	None	69.4	62.3	65.5	10.4	9.7	10.0	54.8	31.1	34.5	23.0	32.5	29.2	47.8	54.2	53.8	41.2
Ex2	C	80.0	71.7	75.5	13.8	12.6	13.2	59.6	32.0	36.0	45.8	40.8	42.5	73.9	54.8	56.2	47.4
Ex3	G	79.7	69.0	73.9	12.7	12.0	12.4	58.6	33.9	37.4	46.0	34.7	38.6	62.6	55.3	55.9	46.5
Ex4	C+G	79.1	70.6	74.5	15.3	14.7	15.0	59.6	34.1	37.8	53.4	42.0	45.9	67.0	57.0	57.8	48.6
LLaMA-AdapterV2	-	34.2	6.3	19.3	5.0	0.1	2.7	23.7	4.6	7.6	9.8	11.3	10.8	2.6	1.5	1.6	9.6
LLaVA1.5	-	38.9	51.9	45.8	7.7	7.6	7.7	10.5	7.4	7.8	7.0	9.9	9.0	64.5	50.8	52.1	26.2

Table 3: The high-level instruction results on NuScenes-QA. ”C” and ”G” denote loading the pre-trained parameters from the 3D captioning and 3D grounding task. H0 and H1 represent zero-hop and one-hop reasoning questions, respectively.

Dataset	Task	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Bert-F1	Bert-P	Bert-R
DriveLM nuScene	Perception	57.44	36.96	31.20	26.38	96.36	96.73	96.03
	Planning	37.47	26.94	16.99	13.58	92.18	93.19	91.25
	Behaviour	82.02	80.02	78.09	75.82	97.80	98.15	97.45
	Mean	60.54	40.16	33.84	29.26	95.45	96.02	94.91

Table 4: Experimental results on the DriveLM-nuScene dataset show that our model achieves promising performance in the three types of reasoning tasks.

	QTrans	BPE	BERT-P	BLEU-4
Ex_1			88.14	11.37
Ex_2	✓		90.60	15.41
Ex_3	✓	✓	91.32	19.26

Table 5: Ablation study of PAT, where BPE means BEV Position Embedding, QTrans means the Query Transformer.

QAs involving 2D coordinates, resulting in the filtering out of a significant amount of data and the remaining data with fixed prompt formats. Hence, focusing on Bert Score metric, LiDAR-LLM also achieved 98.15% precision, validating that our model can predict reasonable ego driving behavior based on traffic conditions. Detailed qualitative results are provided in Appendix B. To more comprehensively demonstrate the efficacy of our LiDAR-LLM, we also conduct comparisons with a 2D-MLLM utilizing multi-view input, as well as with the state-of-the-art indoor 3D LLM. Due to sapce limitation, the results are presented in Appendix A.

Ablation Study

Effectiveness of Position-Aware Transformer (PAT). To demonstrate the effectiveness of each component in PAT, we compare BLEU-4 and BERT scores precision in the 3D captioning task. As shown in Table. 5, Ex_1 without any transformer structure or position embedding. Before being fed into the LLM, the BEV Feature is directly processed by an MLP, resulting in Ex_1 achieving only 88.14% BERT-P score and 11.37% BLEU-4. In Ex_2, with the use of Query Transformer, we observe that the BERT Score achieves 90.60% and 15.41% BLEU-4. The results demonstrate that our proposed Query Transformer can better align LiDAR features with text embeddings. Compared with Ex_2 and Ex_3, the introduction of the BEV Position embedding achieves a 0.7% BERT score and 3.85% BLEU-4 improvement. This set of results confirms that incorporating the BEV Position embedding helps the model better understand 3D spatial relationships while not limiting the flexibility of language.

Qualitative Analysis

In the 3D captioning task, as indicated in the green part of Fig. 3, LiDAR-LLM demonstrates its proficiency in aligning language with LiDAR input. It showcases an understanding of contextual information within the LiDAR data and provides answers based on both textual questions and corresponding visual information. From the figure, our method excels in identifying crucial objects and evaluating their states, such as “parked” or “driving down,” showcasing its ability in comprehending 3D scenes. During the grounding stage, as shown in blue part of Fig. 3, the model showcases its ability to identify the referred object, exhibit spatial relation awareness, and demonstrate precise localization skills. In the yellow part of Fig. 3, we illustrate the interpretability of LiDAR-LLM in planning tasks. Leveraging the capabilities unlocked by our method, LiDAR-LLM generates coherent high-level actions. For example, in the second subfigure, we prompt the model to describe the safe action for the ego car. The model accurately identifies that there are no obstacles or potential dangers ahead, so the reasoning answer is ’Go ahead or turn left. Furthermore, in another high-level task (NuScenes-QA), as represented by the pink part of Fig. 3, LiDAR-LLM demonstrates relatively accurate panoramic scene perception and reasoning capabilities.

Conclusion

In conclusion, our paper represents a pioneering effort to unleash the reasoning capabilities of LLMs to comprehend outdoor LiDAR data. To train LiDAR-LLM, we generate a comprehensive set of LiDAR-text paired datasets, encompassing 420K 3D captioning and 280K 3D grounding data. We then introduce a three-stage training strategy, gradually aligning the LiDAR modality with the language embedding space of the LLM. Our architectural innovation introduces the Position-Aware Transformer to connect the 3D encoder with the LLM. Through extensive experimentation on both our generated datasets and open-source datasets, our LiDAR-LLM demonstrates promising performance across diverse tasks, including 3D captioning, 3D grounding, 3D question answering, autonomous driving planning, and a series of high-level instruction tasks.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62476011).

References

Arnold, E.; Al-Jarrah, O. Y.; Dianati, M.; Fallah, S.; Oxtoby, D.; and Mouzakitis, A. 2019. A survey on 3d object detection methods for autonomous driving applications. *IEEE Transactions on Intelligent Transportation Systems*, 20(10): 3782–3795.

Azuma, D.; Miyanishi, T.; Kurita, S.; and Kawanabe, M. 2022. ScanQA: 3D question answering for spatial scene understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19129–19139.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11621–11631.

Changpinyo, S.; Sharma, P.; Ding, N.; and Soricut, R. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3558–3568.

Chen, X.; Ma, H.; Wan, J.; Li, B.; and Xia, T. 2017. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 1907–1915.

Contributors, D. 2023. DriveLM: Drive on Language. https://github.com/OpenDriveLab/DriveLM.

Dhamo, H.; Manhardt, F.; Navab, N.; and Tombari, F. 2021. Graph-to-3d: End-to-end generation and manipulation of 3d scenes using scene graphs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16352–16361.

et al, J.-B. A. 2022. Flamingo: a Visual Language Model for Few-Shot Learning.

Guo, Z.; Zhang, R.; Zhu, X.; Tang, Y.; Ma, X.; Han, J.; Chen, K.; Gao, P.; Li, X.; Li, H.; et al. 2023. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. *arXiv preprint arXiv:2309.00615*.

Hong, Y.; Zhen, H.; Chen, P.; Zheng, S.; Du, Y.; Chen, Z.; and Gan, C. 2023. 3d-llm: Injecting the 3d world into large language models. *arXiv preprint arXiv:2307.12981*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Jiao, Y.; Chen, S.; Jie, Z.; Chen, J.; Ma, L.; and Jiang, Y.-G. 2022. More: Multi-order relation mining for dense captioning in 3d scenes. In *European Conference on Computer Vision*, 528–545. Springer.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. arXiv:2301.12597.

Lin, Z.; Liu, C.; Zhang, R.; Gao, P.; Qiu, L.; Xiao, H.; Qiu, H.; Lin, C.; Shao, W.; Chen, K.; et al. 2023. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2023. Improved Baselines with Visual Instruction Tuning.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Ma, X.; Yong, S.; Zheng, Z.; Li, Q.; Liang, Y.; Zhu, S.-C.; and Huang, S. 2022. Sqa3d: Situated question answering in 3d scenes. *arXiv preprint arXiv:2210.07474*.

OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Parelli, M.; Delitzas, A.; Hars, N.; Vlassis, G.; Anagnostidis, S.; Bachmann, G.; and Hofmann, T. 2023. 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*, 5606–5611.

Qi, Z.; Fang, Y.; Sun, Z.; Wu, X.; Wu, T.; Wang, J.; Lin, D.; and Zhao, H. 2024. Gpt4point: A unified framework for point-language understanding and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26417–26427.

Qian, T.; Chen, J.; Zhuo, L.; Jiao, Y.; and Jiang, Y.-G. 2023. NuScenes-QA: A Multi-modal Visual Question Answering Benchmark for Autonomous Driving Scenario. *arXiv preprint arXiv:2305.14836*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2556–2565.

Sun, P.; Kretzschmar, H.; Dotiwalla, X.; Chouard, A.; Patnaik, V.; Tsui, P.; Guo, J.; Zhou, Y.; Chai, Y.; Caine, B.; et al. 2020. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2446–2454.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wang, H.; Chen, A. G. H.; Li, X.; Wu, M.; and Dong, H. 2024. Find What You Want: Learning Demand-conditioned Object Attribute Space for Demand-driven Navigation. *Advances in Neural Information Processing Systems*, 36.

Wang, W.; Chen, Z.; Chen, X.; Wu, J.; Zhu, X.; Zeng, G.; Luo, P.; Lu, T.; Zhou, J.; Qiao, Y.; et al. 2023a. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *arXiv preprint arXiv:2305.11175*.

Wang, Z.; Huang, H.; Zhao, Y.; Zhang, Z.; and Zhao, Z. 2023b. Chat-3D: Data-efficiently Tuning Large Language Model for Universal Dialogue of 3D Scenes. *arXiv preprint arXiv:2308.08769*.

Yang, Z.; Zhang, S.; Wang, L.; and Luo, J. 2021. Sat: 2d semantics assisted training for 3d visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1856–1866.

Yao, S.; Hsu, T. M.; Zhu, J.-Y.; Wu, J.; Torralba, A.; Freeman, B.; and Tenenbaum, J. 2018. 3d-aware scene manipulation via inverse graphics. *Advances in neural information processing systems*, 31.

Zhang, R.; Han, J.; Zhou, A.; Hu, X.; Yan, S.; Lu, P.; Li, H.; Gao, P.; and Qiao, Y. 2023. Llama-adapter: Efficient finetuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhou, Y.; and Tuzel, O. 2018. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4490–4499.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.