NPM3D Project: Study Report
LiDAR-LLM: Exploring the Potential of Large Language Models for 3D LiDAR Understanding

Zhe HUANG from IASD

20/03/2024

Abstract

Recently, Large Language Models (LLMs) and Multimodal Large Language Models (MLLMs) have shown promise in instruction following and 2D image understanding. While these models are powerful, they have not yet been developed to comprehend the more challenging 3D physical scenes, especially when it comes to the sparse outdoor LiDAR data. In this report, I will introduce an insightful article titled **LiDAR-LLM: Exploring the Potential of Large Language Models for 3D LiDAR Understanding**[1]. The article introduces LiDAR-LLM, an innovative approach framing 3D scene understanding as a language modeling problem, encompassing 3D captioning, grounding, and Q&A tasks. Specifically, due to the scarcity of 3D LiDAR-text pairing data, the authors introduce a three-stage training strategy and generate relevant datasets, progressively aligning the 3D modality with the language embedding space of LLM. Furthermore, they design a **View-Aware Transformer (VAT)** 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. The article's experiments demonstrate that LiDAR-LLM not only excels in generating descriptive captions and accurately grounding objects within 3D scenes but also showcases superior performance in answering complex spatial questions.

1 introduction

In recent years, the rapid advancement in artificial intelligence has led to significant breakthroughs in natural language processing (NLP) and computer vision (CV), primarily fueled by the emergence of Large Language Models (LLMs) and Multimodal Large Language Models (MLLMs). These models have demonstrated remarkable abilities in generating coherent text, understanding complex instructions, and making sense of 2D visual content.

Despite these advancements, the comprehension of 3D physical scenes, particularly those captured through Light Detection and Ranging (LiDAR) technology, remains a challenging frontier. LiDAR sensors provide sparse, three-dimensional point clouds that capture the geometry of outdoor scenes with precision, making them indispensable for applications like autonomous driving and urban planning. However, the sparse and unstructured nature of LiDAR data poses unique challenges for interpretation: while we have models that excel in 2D image understanding and text generation, their capabilities extend only marginally into the realm of 3D scene understanding.

Addressing this challenge, the article titled ”LiDAR-LLM: Exploring the Potential of Large Language Models for 3D LiDAR Understanding” presents a pioneering approach that leverages the power of LLMs for the interpretation of outdoor 3D scenes captured by LiDAR sensors. As shown in Figure 1, the authors 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 Li-DAR encoder, an intermediate alignment transformer, and an LLM(e.g., LLaMA).

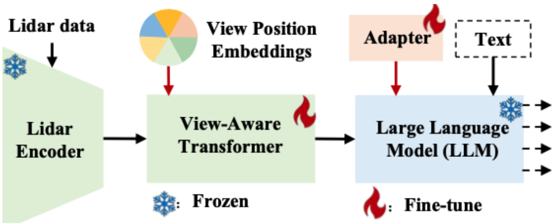


Figure 1: **Characteristics of LiDAR-LLM**. LiDAR-LLM takes 3D LiDAR data as input and aligns the 3D modality with the language embedding space, leveraging the exceptional reasoning capabilities of LLMs to understand outdoor 3D scenes.

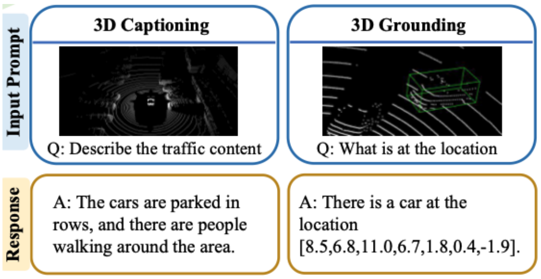

Input Prompt
3D Captioning
3D Grounding
Q: Describe the traffic content
Q: What is at the location
Response
A: The cars are parked in rows, and there are people walking around the area.
A: There is a car at the location [8.5,6.8,11.0,6.7,1.8,0.4,-1.9].

Figure 2: A little Example

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, 3D LiDAR-text paired data is exceedingly rare, and readily accessible multimodal models (e.g., CLIP) are lacking. (2) 3D LiDAR data encompasses a variety of objects and intricate geometric re- lationships among them. Take outdoor autonomous driving, for example, where the ego vehicle is surrounded by a di- verse array of moving and stationary objects, which both occlude and influence each other.

To tackle these challenges, for LiDAR-LLM, the authors introduce a three-stage training strategy and generate relevant datasets, gradually transferring 3D representations into the text feature space and unleashing LLMs' reasoning capabilities for 3D scenes. Specifically, in the first stage, they employ MLLMs and GPT4 for communication between multi-view images and language within the nuScenes dataset, where each scene is accompanied by paired 3D LiDAR data. During the second stage, as the perception forms the foundation of 3D scene understanding, they incorporate the 3D bounding boxes into the question-answer text. In the final stage, they perform efficient fine-tuning of their model on high-level instruction datasets, comprehensively expanding its capabilities for 3D downstream tasks. To more effectively bridge the modality gap between 3D Li-DAR and text, they design a View-Aware Transformer (VAT) that connects the 3D LiDAR encoder with the LLM, injecting six view position embeddings into the 3D feature. Combined with the three-stage training strategy, VAT enhances the LLM's comprehension of the spatial orientation of visual features.

2 Methodology

2.1 Overview

The fundamental framework of LiDAR-LLM is depicted in 3. At its core lies the concept of converting the intricately sparse geometric data from LiDAR into a representation that can be comprehended by Large-Language Models (LLMs). This transformation is achieved through the authors' proposed View-Aware Transformer (VAT), which integrates view position embeddings to enrich the LLM's understanding of spatial orientation. Consequently, it enables a thorough interpretation of the detailed elements within outdoor 3D scenes.

However, the incorporation of LLMs to grasp outdoor 3D scenes encounters two main challenges: (1) Unlike the abun- dance of available image-text paired data, 3D LiDAR-text paired data is exceptionally scarce; and (2) 3D LiDAR data involves diverse objects and intricate geometric relation- ships among them. Therefore, the authors adopt a three-stage training strategy and generate paired LiDAR-text data to align the 3D representations with the feature space of LLMs. This process enables LiDAR-LLM to tackle diverse tasks across different modalities and navigate complex cross-modal scenarios at both the scene and instance levels.

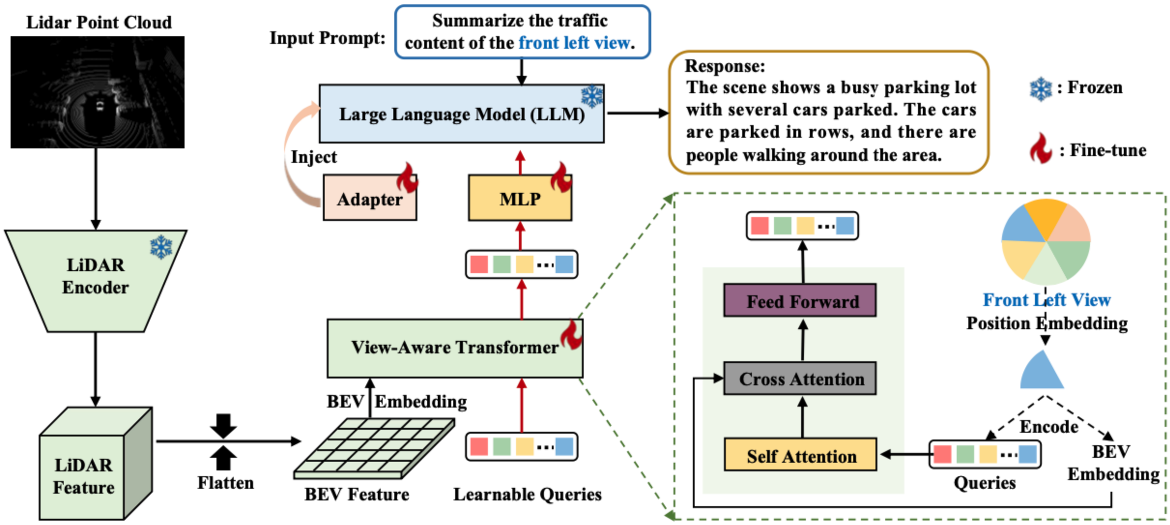


Figure 3: **Overview of LiDAR-LLM framework.**. Initially, the 3D feature extractor(in the first column) processes the LiDAR point cloud input to generate a 3D voxel feature. This feature is then flattened along the z-axis to produce the bird's-eye view (BEV) feature. The View-Aware Transformer (VAT) takes in the BEV embedding and learnable queries as input, producing output queries that serve as soft prompts for the frozen LLM. Within the VAT, they incorporate six view position embeddings into the BEV feature alongside corresponding queries to enhance spatial orientation representation. This framework effectively aligns the LiDAR modality with the language embedding space, enabling one to utilize the LLM for a comprehensive understanding of outdoor 3D scenes.

2.2 Model Architecture

2.2.1 LiDAR-LLM Architecture

- **LiDAR Input Feature Extraction**: For a given LiDAR input $L \in \mathbb{R}^{n\times 3}$, where n denotes the number of point measurements, a LiDAR Encoder(e.g. VoxelNet) is employed to extract its 3D voxel feature.

- **Feature Flattening**: Subsequently, considering the computational cost, we flatten the feature along the z-axis to generate the bird's-eye view (BEV) feature.

- **Text Feature Integration**: Concurrently, the architecture utilizes the pre-trained LLaMA model to extract text features of the text input T with a maximum of m characters.

- **View-Aware Transformer (VAT)**: With the BEV feature $F_v \in \mathbb{R}^{c\times h\times w}$ along with the text feature $F_t \in \mathbb{R}^{m\times d}$ (where d 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 the proposed VAT. During training, we only fine-tune the injected adapters [25] in the LLaMA and VAT 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.

2.2.2 VAT Architectures

- As shown in the right part of Figure 3, the input to the VAT includes a set of K learnable query embeddings, with K set to 576 for convenient projection into the word embed- ding space of the LLM. These queries interact with the BEV feature through a cross-attention mechanism. The VAT 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.

- Considering outdoor LiDAR datasets like nuScenes, there is a requisite for an in-depth perception of how different entities and the ego vehicle are oriented relative to each other. This also involves understanding the complex interrelations amongst the various objects. For this purpose, we incorporate a view position embedding into the BEV feature to enhance the system's ability to discern spatial orientations and geometric connections.

  Specifically, we first construct the view position embedding $V_p \in \mathbb{R}^{c\times 6}$ with zero initial parameters. Subsequently, the BEV feature is divided based on six distinct viewpoints: front, front right, front left, back, back right, and back left. During training, when dealing with a question related to a specific view, we inject the corresponding position embedding into both the BEV feature and queries. For instance, when training a caption sample related to the front left view, we only inject the front left position embedding $V_p \in \mathbb{R}^{c\times 1}$ into the front left view portion of the BEV feature and queries. If the training sample involves a question regarding the entire panoramic scene, we inject all six view position embeddings during training.

2.3 Three-stage Training Strategy

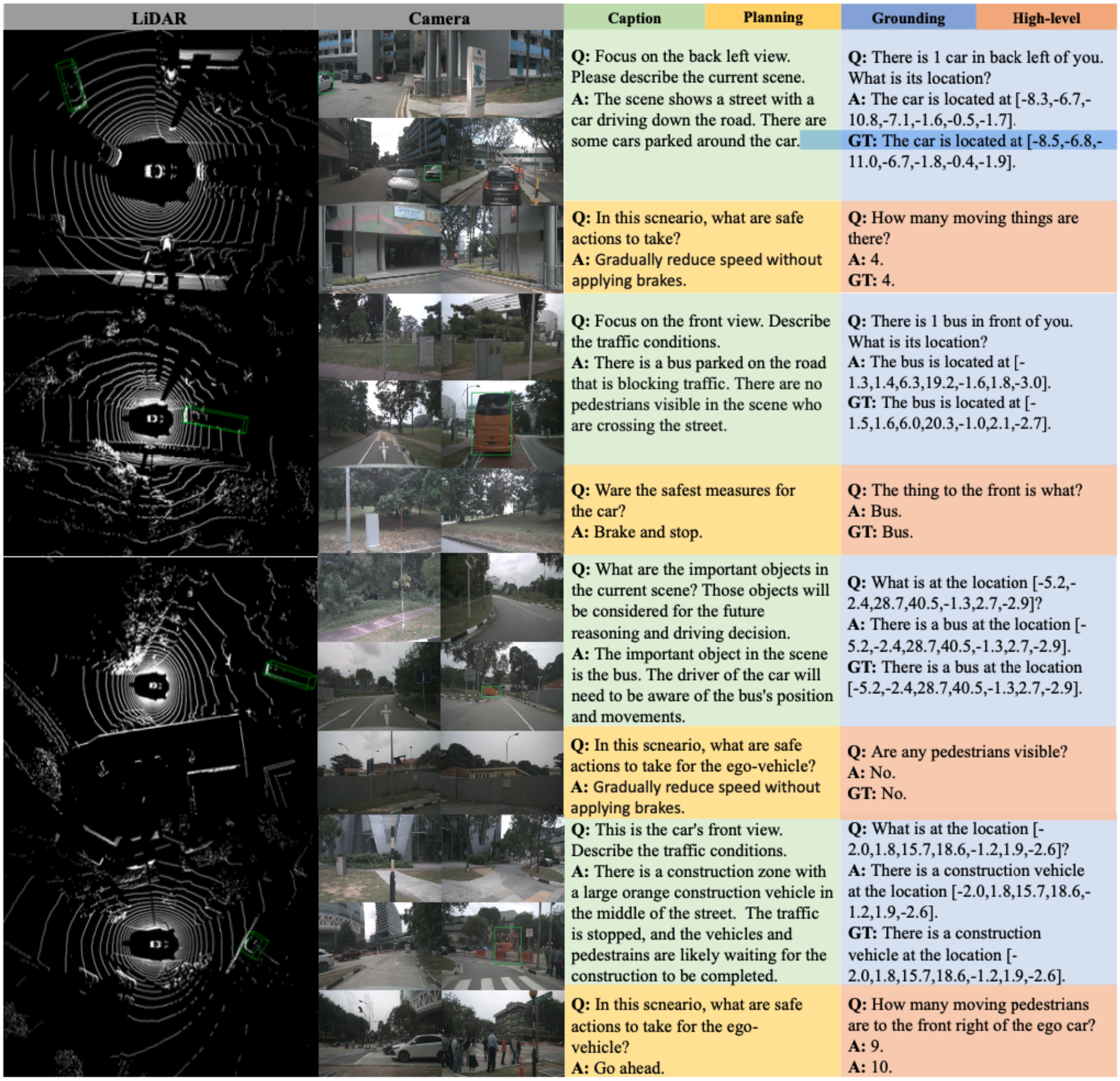
In this section, i will present the authors' three-stage training strategy to demonstrate how they empower LLMs with the capabilities to comprehend 3D LiDAR data and uniformly complete extensive 3D tasks. Three stages contain cross-modal alignment, perception, and high-level instruction, gradually transferring 3D representations into the text feature space.

2.3.1 Cross-Modal Alignment (3D Captioning)

To adeptly tackle a wide array of 3D downstream applications, the model necessitates a comprehensive grasp of LiDAR environments. Scene captioning is a logical approach to enable the model to capture essential information and details in the LiDAR data by integrating the entire 3D scene into LLMs.

The challenge, however, is the lack of paired LiDAR and textual data for captioning exercises, propelling the authors to utilize correlated multi-view imagery and LiDAR data available in nuScenes for text generation. Employing powerful off-the-shelf 2D Multi-Modal LLMs (MLLMs), they generate captions for each view, creating textual descriptions corresponding to the LiDAR scene. It must be noted, though, that the captions generated from 2D MLLM's insights might reference attributes like weather or hues in images, aspects that are incongruent with LiDAR information. To reconcile such disparities, they engage GPT-4 to sift through and select captions that align more coherently with the LiDAR context.

With the collected LiDAR-caption pairs, our goal is to enable LLaMA to generate descriptive text conditioned on LiDAR input. 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 circumvent this complication, they start by training the model to formulate captions for a solitary perspective, thereby simplifying the task. They direct the model's output to conform to the veracity of the specific view's actual annotation via cross-entropy loss. After enabling the model to acquire captioning skills for individual views, 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 to the text feature space of LLM, enabling the model to comprehend the context in the LiDAR data.



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: Gradually reduce speed without applying brakes.

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

Q: Focus on the front view. Describe the traffic conditions.
A: There is a bus parked on the road that is blocking traffic. There are no pedestrians visible in the scene who are crossing the street.

Q: There is 1 bus in front of you. What is its location?
A: The bus is located at [-1.3,1.4,6.3,19.2,-1.6,1.8,-3.0].
GT: The bus is located at [-1.5,1.6,6.0,20.3,-1.0,2.1,-2.7].

Q: Ware the safest measures for the car?
A: Brake and stop.

Q: The thing to the front is what?
A: Bus.
GT: Bus.

Q: What are the important objects in the current scene? Those objects will be considered for the future reasoning and driving decision.
A: The important object in the scene is the bus. The driver of the car will need to be aware of the bus's position and movements.

Q: What is at the location [-5.2,-2.4,28.7,40.5,-1.3,2.7,-2.9]?
A: There is a bus at the location [-5.2,-2.4,28.7,40.5,-1.3,2.7,-2.9].
GT: There is a bus at the location [-5.2,-2.4,28.7,40.5,-1.3,2.7,-2.9].

Q: In this scneario, what are safe actions to take for the ego-vehicle?
A: Gradually reduce speed without applying brakes.

Q: Are any pedestrians visible?
A: No.
GT: No.

Q: This is the car's front view. Describe the traffic conditions.
A: There is a construction zone with a large orange construction vehicle in the middle of the street. The traffic is stopped, and the vehicles and pedestrains are likely waiting for the construction to be completed.

Q: What is at the location [-2.0,1.8,15.7,18.6,-1.2,1.9,-2.6]?
A: There is a construction vehicle at the location [-2.0,1.8,15.7,18.6,-1.2,1.9,-2.6].
GT: There is a construction vehicle at the location [-2.0,1.8,15.7,18.6,-1.2,1.9,-2.6].

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

Q: How many moving pedestrians are to the front right of the ego car?
A: 9.
A: 10.

Figure 4: Qualitative examples of prompt questions and LiDAR-LLM's prediction(from the authors' paper)

2.3.2 Perception

At this juncture, having established a macroscopic comprehension of the scene, the model's development pivots towards attaining proficiency in instance-level discernment. This capability is pivotal for sophisticated command-based tasks, including navigational planning. The authors' methodology employs an object-focused approach to learning, thereby ensuring the model is cognizant of various object details such as quantity, localization, and spatial relations. It is through this method that the model becomes versed in the alignment of each object's 3D representation to its textual descriptor within the LLM.

Two tasks are designed for this purpose: visual grounding and grounded caption generation. Initially, objects within the scene are identified and encoded as a series of distinct tokens that encapsulate the object's classification and volumetric parameters. Given a 3D object with its annotations, its category name and locations are encoded into a word embedding using the tokenizer of the pre- trained LLM. Contrary to the previous method applied to indoor 3D MLLM that required individual object extraction from point clouds, the enhanced strategy comprehends object perception across the entire 3D landscape. In the visual grounding task, 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 denotes the box angle. The task of Grounded Captioning 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. Both tasks are regulated and refined using a cross-entropy loss function. The formulations of the instructions are depicted in Figure 4. This alignment process aims to align the 3D visual object embedding with the text embedding space, culminating in the LLM's 3D perception ability.

2.3.3 High-level Instruction

According to the paper, in this phase, the model, now versed in the LiDAR environment and equipped with foundational 3D perceptual skills, is further refined using an advanced instruction dataset, such as nuScenes-QA, to advance its inferential abilities in 3D space. Fine-tuning the LiDAR-LLM with this dataset does not only enhance its proficiency in compre- hending a diverse array of instructions but also empower it to generate responses that are both creative and con- textually appropriate. This stage of refinement further enables LiDAR-LLM to conduct complex spatial analysis and assimilate ancillary knowledge into its response construction. The adherence to cross-entropy loss during supervision ensures that the model's outputs remain in concord with the sophisticated instructions intended. Foregoing the creation of specific planning QA data, the authors instead rely on the model's inherent competencies to deduce queries pertinent to navigational strategy. Through the three-stage training strategy, LiDAR-LLM develops preliminary planning capabilities, as illustrated in Figure 4.

3 Implementation Details

The LiDAR-LLM mainly comprises three components: Li-DAR feature extraction backbone, View-Aware Transformers(VAT), and the Large Language Model(LLM).

- For the LiDAR feature extraction, I employ the standard pretrained 3D detector, CenterPoint-Voxel following its default settings. The point cloud range is $[-54.0m, 54.0m, -5.0m, 54.0m, 54.0m, 3.0m]$, and the BEV grid size is $[0.6m, 0.6m]$.
- For the VAT, i set the token number of learnable queries to 576(same as the original paper), and the dimension of the token is 768.
- In terms of the LLM, i employ LLaMA-7B considering both efficiency and efficacy.
- Throughout the three-stage training phase, i utilize the Adam optimizer $(\beta_1, \beta_2) = (0.9, 0.999)$ with an initial learning rate of $1e - 4$, halving it every 2 epochs. And we fine-tuning the VAT and adapters in LLaMA2 for 6 epochs. All code are conducted on colab.

Due to computational resource constraints, it was challenging to replicate the full scope of experiments detailed in the article. As a result, my implementation may not meet the desired level of comprehensiveness. But Efforts were made to approximate the methodologies described for LiDAR-LLM as closely as possible within the limitations.

4 Suggested Improvements

Despite the groundbreaking achievements of LiDAR-LLM, there are several potential areas where the framework could be enhanced. In subsequent research, i think it is worthwhile to explore these avenues:

- **Expanding Data Diversity**: Introduce a continuous learning paradigm where LiDAR-LLM is periodically updated with new data collected under diverse environmental conditions, including adverse weather and varying lighting. This approach can utilize a federated learning system to harness data from deployed autonomous systems across the globe, enriching the model's experience without compromising data privacy. Moreover, incorporating synthetic data generated through simulations can provide a breadth of scenarios that are not commonly encountered or are difficult to capture in real-world settings.

- **Real-Time Data Processing**: To facilitate real-time processing, LiDAR-LLM's VAT component could be optimized for speed without significantly sacrificing accuracy. This might include pruning the model to retain only the most critical parameters or leveraging quantization techniques to reduce computational demand. Additionally, the deployment of edge computing resources can minimize latency, as data processing can be conducted closer to where data is collected. During the fine-tuning stage of training, emphasis should be placed on optimizing the trade-off between performance and speed, ensuring the model remains agile and responsive in time-sensitive applications.

- **Seamless System Integration**: Developing a middleware framework that enables LiDAR-LLM to communicate and cooperate with other onboard systems such as GPS, IMU, and traffic prediction algorithms is essential. This middleware should act as an interpreter between LiDAR-LLM's outputs and the inputs required by the vehicle's control systems. During the second stage of training, the focus on perception can be leveraged to fine-tune this communication, ensuring that LiDAR-LLM's instance-level discernment feeds into the broader planning systems effectively. The final stage of training can then focus on simulating real-world operational scenarios to ensure that the model's output is in sync with the high-level instructions provided by autonomous driving planning systems.

5 conclusion

In conclusion, the authors' proposed LiDAR-LLM framework reframes the intricate challenge of understanding 3D outdoor scenes as a language modeling problem. Through a sophisticated three-stage training strategy, which encompasses cross-modal alignment, perception, and high-level instruction, the model aligns the LiDAR modality with the language embedding space of the LLM. This enables the model to comprehend the intricacies of outdoor scenes more effectively. A key architectural innovation introduced in the original paper is the View-Aware Transformer (VAT), which acts as a bridge between the 3D encoder and the LLM. This design enhancement effectively addresses the modality gap and enhances the LLM's spatial orientation comprehension, further enhancing the model's performance.

In my implementation, I followed the methodology outlined in the original paper for LiDAR-LLM. Additionally, I proposed several improvements aimed at ultimately making LiDAR-LLM feasible for deployment on edge devices. These improvements include exploring avenues such as expanding data diversity through continuous learning paradigms, optimizing real-time data processing by refining the VAT component for speed, and developing a middleware framework for seamless system integration with other onboard systems. By incorporating these enhancements, I aim to advance the practical utility and efficiency of LiDAR-LLM in real-world scenarios, particularly in applications such as autonomous driving and urban planning.

References

[1] S. Yang, J. Liu, R. Zhang, *et al.*, "Lidar-llm: Exploring the potential of large language models for 3d lidar understanding," National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University 2AI2Robotics 3 Shanghai Artificial Intelligence Laboratory, 2023.