LLMI3D: Empowering LLM with 3D Perception from a Single 2D Image

Fan Yang[1,2] · **Sicheng Zhao**[2,*] · **Yanhao Zhang**[3] · **Haoxiang Chen**[3] · **Hui Chen**[2] · **Wenbo Tang**[4] · **Haonan Lu**[3] · **Pengfei Xu**[4] · **Zhenyu Yang**[3] · **Jungong Han**[5] · **Guiguang Ding**[1,2,*]

Abstract Recent advancements in autonomous driving, augmented reality, robotics, and embodied intelligence have necessitated 3D perception algorithms. However, current 3D perception methods, particularly small models, struggle with processing logical reasoning, question-answering, and handling open scenario categories. On the other hand, generative multimodal large language models (MLLMs) excel in general capacity but underperform in 3D tasks, due to weak spatial and local object perception, poor text-based geometric numerical output, and inability to handle camera focal variations. To address these challenges, we propose the following solutions: Spatial-Enhanced Local Feature Mining for better spatial feature extraction, 3D Query Token-Derived Info Decoding for precise geometric regression, and Geometry Projection-Based 3D Reasoning for handling camera focal length variations. We employ parameter-efficient fine-tuning for a pre-trained MLLM and develop LLMI3D, a powerful 3D perception MLLM. Additionally, we have constructed the IG3D dataset, which provides fine-grained descriptions and question-answer annotations. Extensive experiments demonstrate that our LLMI3D achieves state-of-the-art performance, significantly outperforming existing methods.

* Corresponding Authors: Sicheng Zhao and Guiguang Ding
Fan Yang: yfthu@outlook.com
Sicheng Zhao: schzhao@gmail.com
Yanhao Zhang: zhangyanhao@oppo.com
Haoxiang Chen: hxchen22@m.fudan.edu.cn
Hui Chen: jichenhui2012@gmail.com
Wenbo Tang: tangwenbo0104@gmail.com
Haonan Lu: luhaonan@oppo.com
Pengfei Xu: pfxu@outlook.com
Zhenyu Yang: yangzhenyu@oppo.com
Jungong Han: jungonghan77@gmail.com
Guiguang Ding: dinggg@tsinghua.edu.cn
1 School of Software, Tsinghua University, Beijing, China
2 BNRist, Tsinghua University, Beijing, China
3 OPPO AI Center, Shenzhen China
4 NavInfo, Beijing, China.
5 Computer Science Department, University of Sheffield, S1 4DP, UK.

1 Introduction

With the rapid development of deep learning, 2D perception tasks such as object detection (Chen et al., 2023d; Wang et al., 2024; Lyu et al., 2023), semantic segmentation (Zhao et al., 2021), and visual grounding (Liu et al., 2023b) have achieved remarkable progress (Li et al., 2021; Shen et al., 2023). However, the real world is three-dimensional, and many practical applications, such as autonomous driving, robotics, augmented reality, and embodied intelligence, demand enhanced spatial perception (Mao et al., 2023; Angelova et al., 2020; Addari and Guillemaut, 2023). Traditional 2D methods can no longer meet these demands. Therefore, researchers have introduced the concept of three-dimensional perception, which involves inferring the position, dimension, and pose of objects in three-dimensional space to achieve accurate predictions of their spatial locations (Mousavian et al., 2017; Qin et al., 2019).

Currently, 3D perception techniques primarily include methods using LiDAR point clouds (Shi et al., 2023; Stoiber et al., 2022; Xie et al., 2021; Lang et al., 2019; Zhang et al., 2022; Aumentado-Armstrong et al., 2023) and those based on camera images (Sundermeyer et al., 2020). LiDAR-based methods offer superior depth prediction capabilities; however, their high costs and complex components limit their applicability in many scenarios (Wang et al., 2023c; Zhang et al., 2024). In contrast, camera image-based methods are more cost-effective and are easier to integrate, making them widely used in autonomous driving, robotics, and augmented reality (Mao et al., 2023).

In recent years, numerous specialized perception models have been developed for image-based 3D perception.

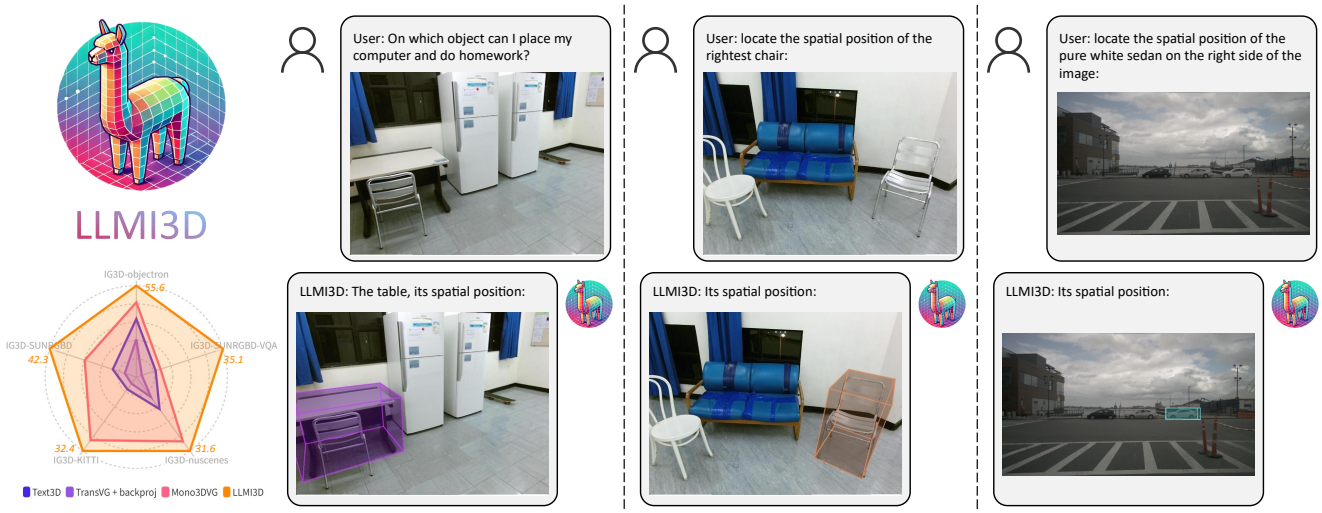

Fig. 1 Our LLMI3D endows multimodal large language models with 3D perception capabilities. When provided with a question or description about an object, our LLMI3D can return the object of interest and its 3D bounding box in 3D space. Across various datasets, our LLMI3D significantly outperforms existing methods on 3D perception.

Table 1 Existing specialized small models and multimodal large language models have various limitations. Only our approach, LLMI3D, exhibits comprehensive and robust 3D perception capabilities.

Methods	instruction processing	logical reasoning	question answering	open vocabulary	local spatial feature extracting	focal length variation handling	3D box outputting
MonoDETR	✗	✗	✗	✗	✓	✗	✓
Omni3D	✗	✗	✗	✗	✓	✓	✓
Mono3DVG	✓	✗	✗	✗	✓	✗	✓
VisionLLM	✓	✓	✓	✓	✗	✗	✗
Qwen-VL	✓	✓	✓	✓	✗	✗	✗
Ours	✓	✓	✓	✓	✓	✓	✓

However, these models face several issues: Firstly, single-modal 3D detection models lack the capability to process textual inputs, preventing them from accurately locating one specific object based on users' instructions. Current models generally detect all objects in an image but fail to pinpoint a specific object of interest. Secondly, the existing multimodal 3D perception model, Mono3DVG (Zhan et al., 2024), lacks logical reasoning and question-answering capabilities. While Mono3DVG employs BERT (Devlin et al., 2019) and CNN (He et al., 2016) to extract textual and visual features, BERT's model capacity restrict its performance. Such models cannot handle user questions and are deficient in logical reasoning and question-answering abilities. Finally, specialized small models are confined to predefined categories within the training dataset, rendering them incapable of performing open-vocabulary 3D perception for arbitrary categories. This limitation prevents them from addressing the vast number of categories and long-tail problems encountered in real-world scenarios (Zhao et al., 2024).

Recently, generative large language models like ChatGPT have demonstrated impressive capabilities. Multi-modal large language models (MLLMs) such as GPT-4V (OpenAI, 2023) and Gemini (Anil et al., 2023), exhibit strong logical reasoning and open-scene generalization abilities, performing well in multimodal tasks. The issues faced by specialized 3D perception small models can be addressed by MLLMs effectively and easily: MLLMs incorporate large-scale pre-trained language models capable of handling both text inputs and outputs, enabling them to follow user instructions easily. Generative MLLMs possess question-answering abilities, and their pre-training on extensive datasets and diverse tasks endows them with logical reasoning capabilities. Additionally, MLLMs are equipped with world knowledge and common sense, enabling the recognition of almost any categories, and they possess open-vocabulary object perception capabilities.

However, despite their general task capabilities, vanilla MLLMs face several issues when applied to specific 3D perception task:

1. Weak spatial and local object perception: Existing MLLMs, such as GPT-4V (OpenAI, 2023), Gemini (Anil et al., 2023), InternVL (Chen et al., 2023c), and Qwen-VL (Bai et al., 2023) are typically pre-trained on vast amounts of 2D images, videos, and text data, but lack exposure to 3D data during training. This results in these models focusing primarily on the semantic understanding of image content while overlooking the 3D spatial structures inherent in the images. Consequently, MLLMs generally lack the capability to extract 3D geometric spatial features, leading

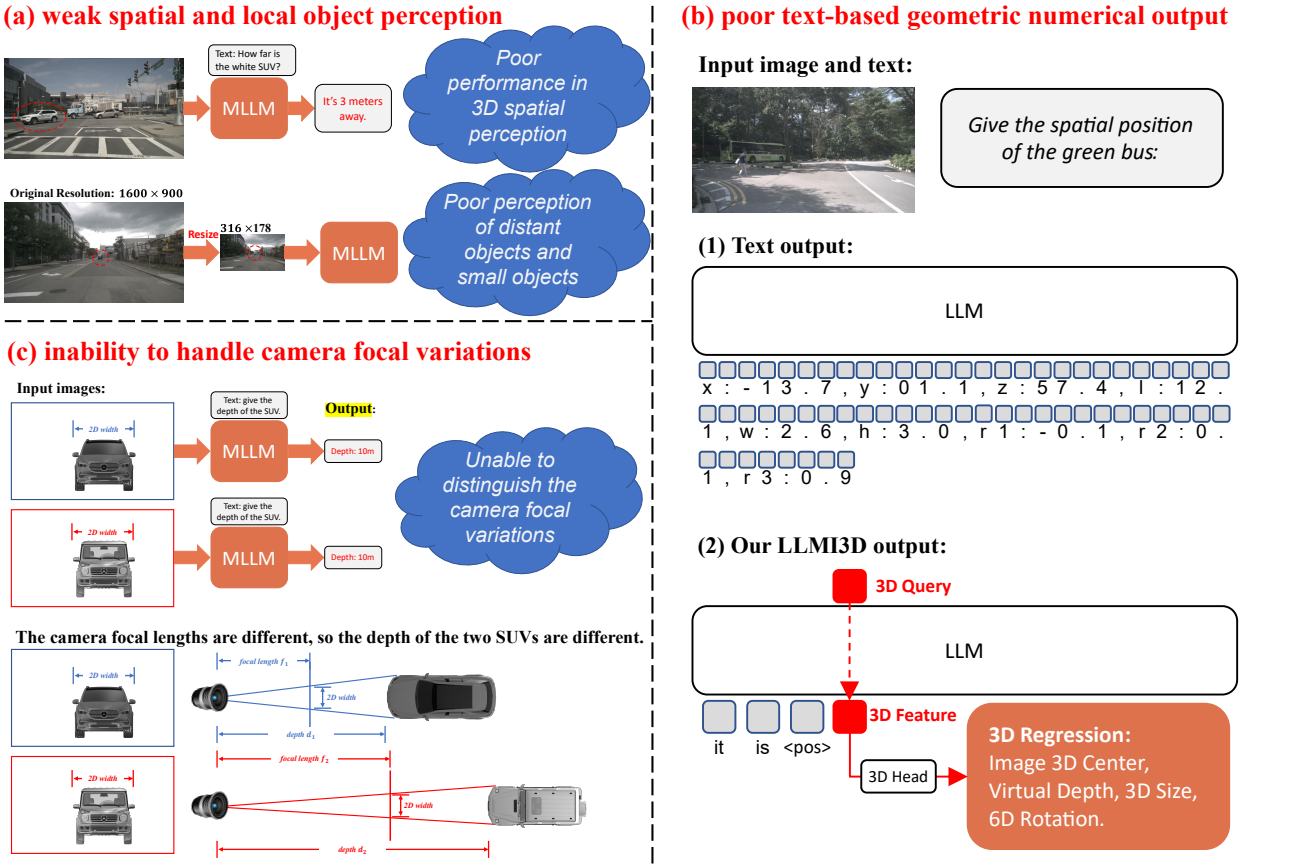

Fig. 2 In the task of 3D perception, vanilla MLLMs encounter significant challenges. (a) Weak spatial and local object perception: 3D spatial perception of existing MLLMs is poor, leading to notably erroneous answers regarding the 3D positions of objects. Additionally, typical MLLMs resize images to a fixed, lower resolution in the image encoder, which hinders their ability to perceive distant and small objects. (b) Poor text-based geometric numerical output: Existing perception models typically output 3D values in pure text format, which suffers from issues such as slow speed, low accuracy, and difficulty in parsing. In contrast, our method employs a learnable 3D query token, utilizing 3D heads to regress the geometric values of objects, significantly improving both accuracy and speed. (c) Inability to handle camera focal variations: It is challenging for models to distinguish changes in camera focal length solely from the input 2D images. As a result, objects that are the same size and have similar positions in 2D images are often predicted to have similar depth by neural networks. However, these images are taken by cameras with different focal lengths, and the actual depth of the two objects in 3D space differs significantly.

to poor performance in 3D perception tasks (Figure 2 (a)). Additionally, most MLLMs lack fine-grained local object feature extraction capabilities. Typical image encoders often resize images to low resolutions, which is inadequate for distant objects and small object detection in autonomous driving. Smaller objects may become nearly invisible after resizing, resulting in poor local feature extraction capabilities, as illustrated in Figure 2 (a).

2. Poor text-based geometric numerical output: Many 2D perception MLLM models output results in text form, which is unsuitable, especially for 3D numerical values (e.g., x, y, z, length, width, height, rotation) (Figure 2 (b)). This leads to issues such as poor precision, slow processing, and parsing complexity. LLMs struggle with numerical representation and mathematics, causing inaccuracies in spatial coordinates. Outputting all numerical values for one 3D box requires up to 50 tokens, significantly slowing down the decoding process. And outputting structured numerical text is prone to errors, with LLMs often misordering or misformatting the output.

3. Inability to handle camera focal variations: It is challenging for models to distinguish camera focal length variations solely from 2D images. 3D perception from a single 2D image is inherently under-constrained. Neural networks typically predict the depth based on the principle that nearby objects appear larger and distant ones appear smaller. However, as illustrated in Figure 2 (c), two objects can appear to have the same size and similar position in a 2D image. The network often predicts these objects to be at similar spatial positions. However, these two images were captured by cameras with different focal lengths, leading to substantial differences in the actual spatial positions of the objects. Consequently, neglecting the differences in camera focal lengths and directly using neural networks as a black

box to predict the 3D positions of objects can result in significant errors when dealing with focal length variations.

Both specialized small models for 3D perception and multimodal large language models face numerous challenges, with their strengths and weaknesses being highly complementary. Therefore, we leverage the strengths of specialized 3D perception small models and large language models. We propose **LLMI3D**, which empowers **LLM** with **I**mage-based **3D** perception capacity. We fine-tune a pre-trained MLLM using LoRA (Hu et al., 2022) and introduce a few extra structures into the image encoder and token decoder, thereby endowing the MLLM with powerful 3D perception capabilities.

To empower large language models with 3D perception capabilities, we introduce a 3D-friendly structure for multimodal large language models. Specifically, in the image encoder component, to address the weak spatial and local object perception in MLLMs, we propose Spatial-Enhanced Local Feature Mining. We comprehensively employ CNN (Liu et al., 2022) and depth predictors to extract spatial-enhanced local features from high-resolution images, and we use ViT (Radford et al., 2021) to obtain tokens from low-resolution images. Then we leverage spatial-enhanced cross-branch attention to effectively mine objects' spatial local features. In the LLM component, to overcome poor text-based geometric numerical output, we propose the 3D Query Token-Derived Info Decoding method. Instead of relying on text-based numerical outputs, we utilize a learnable 3D query token and 3D heads to accurately regress 3D geometric coordinates. This learnable 3D query token can adaptively extract 3D features from images and text within the LLM's self-attention mechanism. We then use 3D heads to precisely regress the output feature of the 3D query, obtaining 3D attribute outputs, including the image's 3D center, virtual depth, 3D size, and 6D rotation, as illustrated in Figure 2(b). For 3D box outputting, to address MLLMs' inability to handle variations in camera focal length, we do not solely rely on focal length-invisible black-box neural network. Instead, we combine black-box networks with white-box projection methods. We introduce Geometry Projection-Based 3D Reasoning, integrating camera parameters into geometric projection to mitigate the impact of varying camera focal lengths on 3D perception. A comparison of our method with existing specialized small models and multimodal large language models is shown in Table 1.

Furthermore, an appropriate dataset is crucial for fine-tuning MLLMs. Existing 3D perception datasets focus on object detection and lack fine-grained caption and question-answer labels. The recently released Mono3DRefer (Zhan et al., 2024) dataset also has significant issues. It includes 3D perception results in object caption inputs, which undermines the evaluation of real 3D perception capabilities.

Therefore, we developed **IG3D**: an **I**mage-based **3D** **G**rounding dataset. The IG3D dataset provides precise descriptions of objects within images, including detailed appearance and location information, thereby differentiating among objects of the same category within an image. This enables the 3D grounding task to be effectively performed. Furthermore, our IG3D dataset includes annotations for Visual Question Answering (VQA) instructions, allowing the assessment of a model's logical reasoning capabilities and accommodating users' personalized input requirements.

In summary, our contributions are as follows:

1. To the best of our knowledge, we are the first to apply parameter-efficient fine-tuning methods to adapt an MLLM for 3D perception from a single image. We identify three major issues in vanilla MLLMs for 3D perception and address these problems by incorporating a few extra structures into the image encoder and token decoder.

2. To address the issue of weak spatial and local object perception, we propose a spatial-enhanced local feature mining approach. This method integrates features extracted by ViT, CNN, and depth predictor while employing the spatial-enhanced cross-branch attention to effectively capture local spatial features of objects.

3. To overcome the problem of poor text-based geometric numerical output, we propose 3D query token-derived info decoding, which uses a single learnable 3D query to efficiently extract 3D features within the LLM and regress the 3D values accurately.

4. To mitigate the inability of MLLMs to handle camera focal variations, we propose geometry projection-based 3D reasoning, which combines black-box neural networks with white-box projection methods. By integrating camera parameters, we reduce the significant impact of varying camera focal lengths on 3D perception.

5. We construct the IG3D, an image-based 3D perception dataset designed to effectively assess a model's 3D grounding and question-answering capabilities. Extensive experiments demonstrate that our approach achieves state-of-the-art performance on various datasets, validating our motivation and the efficacy of our approach.

The rest of this paper is organized as follows: Section 2 reviews related work on multimodal small models, multimodal large language models, and image-based 3D perception. Section 3 elaborates on the proposed LLMI3D method in detail. Section 4 presents the extensive experimental results and analysis. We conclude the paper in Section 5.

2 Related Works

2.1 Multimodal Models

Researchers integrated multimodal data, such as images, video, and text, to construct multimodal models. Early advancements in this domain were pioneered by the "Show and Tell" model (Vinyals et al., 2015), which showcased the combination of convolutional neural networks and recurrent neural networks to generate textual descriptions from images. This foundational work underscored the potential of deep learning frameworks in bridging vision and language.

The introduction of attention mechanisms by Bahdanau et al. (Bahdanau et al., 2015) marked a significant shift in multimodal models. Xu et al.'s "Show, Attend and Tell" (Xu et al., 2015) applied these mechanisms to image captioning, allowing models to selectively focus on different parts of an image while generating descriptive text. This approach significantly enhanced the interpretability and quality of generated captions.

Building upon these foundations, transformer architectures have achieved remarkable success in multimodal integration. ViLBERT (Lu et al., 2019) introduced a model that processes images and text in parallel using two-stream transformers, aligning them through co-attentional mechanisms. This model set new benchmarks for various vision-and-language tasks, illustrating the power of joint vision-language pre-training.

Radford et al.'s CLIP (Radford et al., 2021) introduced a paradigm shift by leveraging a vast dataset of image-text pairs to learn transferable visual models through natural language supervision. CLIP demonstrated exceptional zero-shot performance across multiple datasets and tasks, showcasing the effectiveness of large-scale multimodal pre-training.

In the generative realm, DALL·E (Ramesh et al., 2021) exemplified the synergy between language models and image generation. DALL·E employs autoregressive transformers to create detailed images from textual descriptions, pushing the boundaries of text-to-image generation. FLAVA (Singh et al., 2022) proposed a unified multimodal model leveraging both unimodal and multimodal pre-training objectives. FLAVA combines the strengths of supervised learning and self-supervised learning to achieve comprehensive vision-language understanding.

2.2 Multimodal Large Language Models

In 2023, OpenAI released GPT-4V (OpenAI, 2023), which incorporated image input capabilities into the large language model framework, showcasing powerful vision-text multimodal capabilities. In May 2024, OpenAI introduced GPT-4o, a model capable of processing and generating any combination of text, audio, and image inputs. Notably, GPT-4o can respond to voice input in as little as 232 milliseconds, with an average response time of 320 milliseconds, approaching the reaction time of humans in daily conversations. However, powerful commercial multimodal models like GPT-4V and GPT-4o have not been open-sourced, limiting researchers' ability to build upon these closed models.

Recently, researchers have developed a series of open-source multimodal large language models (Li et al., 2023; Zhu et al., 2023; Bai et al., 2023; Chen et al., 2023c; Lu et al., 2024; Wang et al., 2023b). These open-source multimodal large language models have demonstrated impressive capabilities and have significantly contributed to the advancement of the community.

Liu et al. introduced LLaVA (Liu et al., 2023a), where the authors attempted to generate multimodal language-image instruction-following data using purely language-based GPT-4. By fine-tuning this generated data, they developed the large language and vision assistant. Utilizing CLIP (Radford et al., 2021) and LLaMA (Touvron et al., 2023) for instruction fine-tuning, they constructed the multimodal large model LLaVA, which achieved promising results. The Shanghai AI Lab proposed LLaMA-Adapter (Zhang et al., 2023b), an efficient fine-tuning method that adapts LLaMA into an instruction-following model. The method attaches a set of learnable adaptive prompts as prefixes to the input instruction tokens within the deep layers of the LLaMA transformer (Vaswani et al., 2017). For image input, it uses CLIP to extract multi-scale global features, subsequently concatenated and projected into a global information representation through a projection layer. Despite the promising demonstration of LLaMA-Adapter in handling vision inputs, it hasn't generalized well to open visual instructions and lags behind GPT-4. Peng Gao et al. proposed LLaMA-Adapter-V2 (Gao et al., 2023), a parameter-efficient (Xiong et al., 2024; Hao et al., 2024b) visual instruction model that enhances LLaMA-Adapter by unlocking more learnable parameters (e.g., norms, biases, and scales), extending its instruction-following capabilities (Xu et al., 2024) across the entire LLaMA model.

Wenhai Wang et al. proposed VisionLLM (Wang et al., 2023a), utilizing large language models to perform visual tasks such as detection and instance segmentation. They introduced a language-guided image tokenizer using BERT (Devlin et al., 2019) as the text encoder and deformable DETR (Carion et al., 2020) to capture high-level information. VisionLLM uses Alpaca-7B (Taori et al., 2023) as the backbone of the large language model. Jun Chen et al. proposed MiniGPT-v2 (Chen et al., 2023a), which uses a linear projection layer to map image features to the feature space of the large language model, reducing computational overhead by concatenating four adjacent

visual tokens. Weihai Wang et al. introduced CogVLM (Wang et al., 2023b), differing from shallow alignment methods like feature projection layers by employing deep fusion to better integrate visual features into the large language model. PerceptionGPT (Pi et al., 2023) proposed encoding and decoding perceptual information using a single token. However, it is limited to encoding and decoding 2D perceptual information and cannot handle 3D signals.

These multimodal large language models are trained on massive datasets consisting of 2D images, videos, and textual data, excelling in various 2D tasks. However, they have not been trained on 3D data, resulting in weak spatial perception capabilities.

In May 2024, Cho et al. introduced a pre-trained large model CubeLLM (Cho et al., 2024a). CubeLLM expended substantial resources on multimodal alignment and pre-training across 2D and 3D datasets, specifically involving a total of 9.6 million images and 40.9 million dialogues. A single experiment required 64 A100 GPUs with a batch size of 1024. CubeLLM's model, code, and training data have not been open-sourced. Unlike CubeLLM, which pre-trained a multimodal large language model, our LLMI3D requires much fewer resources because we apply parameter-efficient fine-tuning methods to adapt an MLLM for 3D perception capabilities. Our LLMI3D uses only two A100 GPUs with LoRA (Hu et al., 2022), significantly reducing the required training resources and costs while increasing flexibility.

2.3 3D Perception from a Single Image

Numerous studies have been conducted on 3D perception from a single image, with the vast majority focusing on monocular 3D object detection. Early work by Chen et al. in Mono3D (Chen et al., 2015) utilized geometric priors and a region proposal network to estimate 3D bounding boxes from single images. While effective, the reliance on hand-crafted features limited its applicability across diverse scenarios. Mousavian et al.'s Deep3DBox (Mousavian et al., 2017) introduced an approach combining 2D object detection with 3D pose estimation, marking a pivotal shift towards more accurate 3D localization. This method leveraged both appearance and geometric cues, setting the foundation for subsequent improvements. Further advancements were embodied in Qin et al.'s MonoGRNet (Qin et al., 2019) employing graph-based reasoning to capture spatial relationships within scenes, significantly enhancing the precision of 3D bounding box estimations. End-to-end learning frameworks have also shown great promise. Li et al. presented RTM3D (Li et al., 2020), a unified model integrating detection and localization stages, yielding robust performance through sophisticated feature extraction and attention mechanisms.

In recent years, image-based 3D perception has substantial progress (Wu et al., 2023; Zhang et al., 2023c; Yang et al., 2022; Yan et al., 2024). MonoFlex (Zhang et al., 2021) utilized key points to assist depth estimation and adopted an uncertainty-based ensemble method for improved accuracy. MonoRCNN (Shi et al., 2021) incorporated geometric information between 2D bounding box heights and 3D object heights to estimate depth. GUPNet (Lu et al., 2021) estimated object depth through the projection of 2D and 3D heights and employed an uncertainty loss for precise scoring. Gpro3D (Yang et al., 2023) significantly enhanced the accuracy of object depth and spatial predictions by leveraging ground plane priors.

In 2024, Zhan et al. (Zhan et al., 2024) proposed a multimodal specialized small model capable of performing 3D grounding tasks based on textual descriptions. However, it solely uses BERT (Devlin et al., 2019) and CNN (He et al., 2016) for simple feature extraction of text and images. Limited by model capacity and pre-training scale, BERT could only accept direct object descriptions. Mono3DVG lacks logical reasoning capabilities and cannot answer user questions. Furthermore, Mono3DVG is restricted to trained object classes and would perform poorly when encountering new classes during testing.

Zhan et al. also introduced the Mono3DRefer 3D grounding dataset (Zhan et al., 2024). However, Mono3DRefer has significant issues: it treats the 3D perception results, such as object depth and dimensions (length, width, height), as descriptive inputs for the objects. Consequently, the model can directly infer 3D results based on these descriptions, which prevents a fair evaluation of the model's image-based 3D perception capabilities.

3 Methodology

3.1 Overview

The architecture of our method is illustrated in Figure 3. We fine-tune a pre-trained multimodal large model to empower it with image-based 3D grounding capabilities. We propose a 3D-friendly structure by incorporating a few additional structures into the image encoder and token decoder of MLLMs. In the image encoder component, to address the weak spatial and local object perception issues of MLLMs, we introduce Spatial-Enhanced Local Feature Mining, enhancing the image encoder's ability to extract features from local objects and spatial structures, as detailed in Section 3.2.

In the LLM component, to overcome poor text-based geometric numerical output, we propose 3D Query Token-Derived Info Decoding. This method addresses the

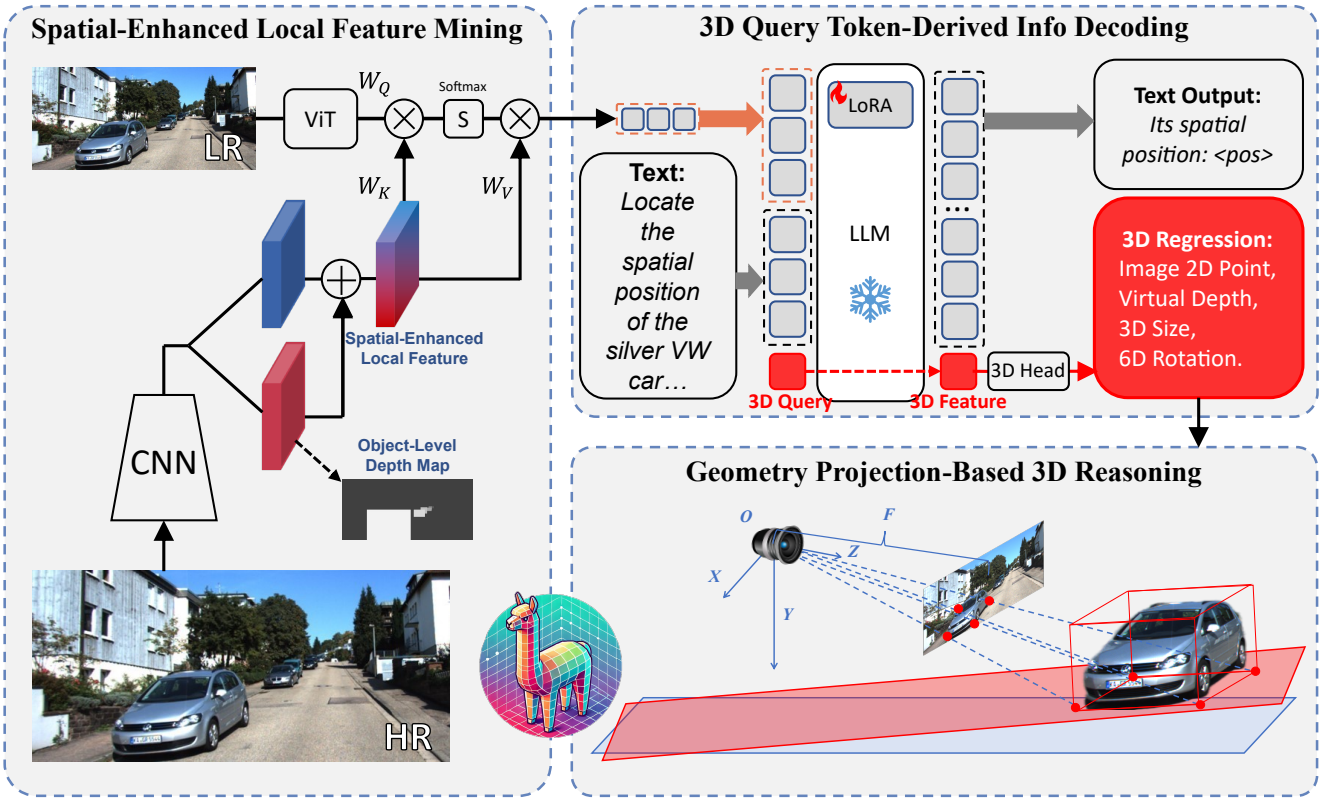

Fig. 3 Framework of our LLMI3D. We propose a 3D-friendly multimodal large language model architecture. (1) In the image encoder, we introduce Spatial-Enhanced Local Feature Mining. We employ a CNN and a depth predictor to extract spatial-enhanced local features from high-resolution (HR) images, and we use a ViT to obtain tokens from low-resolution (LR) images. Then we effectively mine the object's spatial local features with spatial-enhanced cross-branch attention. (2) In the LLM part, we propose 3D Query Token-Derived Info Decoding. We utilize a learnable 3D query token to extract 3D features from the LLM and employ 3D heads to accurately regress the object's geometric attributes. (3) To obtain the 3D bounding box of an object, we introduce geometry projection-based 3D Reasoning. Rather than using focal length-invisible neural network black-box methods for 3D reasoning, we adopt a combined approach of neural networks and geometric projection for 3D spatial reasoning. This effectively alleviates the significant errors introduced by varying camera focal lengths in 3D perception.

drawbacks of slow speed, low accuracy, and difficulty in parsing when 3D coordinates are outputted in text form. By using a single learnable 3D query token combined with 3D heads regression, we can accurately regress the 3D attributes of objects, as elaborated in Section 3.3.

To obtain the 3D bounding box and address MLLMs' inability to handle variations in camera focal lengths, we introduce geometry projection-based 3D Reasoning. Rather than relying solely on focal length-invisible black-box neural network 3D reasoning methods, we appropriately utilize camera intrinsic parameters. By combining neural networks with geometric projection, our method effectively mitigates the significant errors in 3D perception caused by different camera intrinsic parameters, as detailed in Section 3.4.

Additionally, in Section 3.5, we introduce the IG3D dataset. The IG3D dataset provides precise descriptions of objects in images, including detailed appearance and location, and distinguishes between different objects of the same category within the images, facilitating the 3D grounding task. Moreover, our IG3D dataset includes annotations for Visual Question Answering (VQA) instructions, assessing the model's logical reasoning capabilities, and catering to personalized user input requirements.

3.2 Spatial-Enhanced Local Feature Mining

The image encoder is a crucial component of multimodal large language models, tasked with extracting image features. However, existing image encoders encounter challenges with weak spatial and local object perception:

1. Current multimodal large language models exhibit inadequate capabilities in extracting 3D spatial features. Existing multimodal large language models are generally pre-trained and aligned on vast 2D image-text datasets. While the image encoder and projector components effectively capture semantic information from images, they often lose geometrical and spatial information. Unlike typical visual-language tasks, 3D grounding from images

is an inherently challenging problem that seeks to derive 3D positioning from 2D image inputs, which naturally lack depth information. Estimating depth is both the hardest and the most critical part of this task. Although pre-trained multimodal models like CLIP-ViT can efficiently extract semantic features, they struggle with spatial geometric features. Therefore, it is crucial to enhance geometric and spatial features using spatial-enhanced feature extractors.

2. Additionally, many image encoders in MLLMs exhibit insufficient capability in extracting features of local small objects. Typical image encoders reduce images to low resolutions in a simplistic manner (Zhang et al., 2023a). For example, LLaVA (Liu et al., 2023a), ShareGPT4V (Chen et al., 2023b), and InternVL (Chen et al., 2023c) resize images to 336x336, and Qwen-VL (Bai et al., 2023) resizes images to 448x448. These resolutions might be sufficient for global image understanding but fall short for detailed object-level perception. In autonomous driving, input images generally have very high resolutions, often up to millions of pixels. Downscaling these images to 336x336 or 448x448 can make small and distant objects, such as cones and pedestrians, unrecognizable. These objects are critical in autonomous driving and cannot be ignored. On the other hand, directly inputting high-resolution images into the image encoder is not feasible. This results in an excessive number of output tokens, surpassing the maximum token count of the LLM, and also significantly slows down the inference speed.

To address this issue, we propose the Spatial-Enhanced Local Feature Mining algorithm. Specifically, similar to typical multimodal large language models, we input low-resolution images into CLIP-ViT (Radford et al., 2021), obtaining a relatively small number of tokens. Next, we input high-resolution images into ConvNeXt (Liu et al., 2022), ensuring that even small and distant objects are clearly visible and a sufficient number of pixels can be processed. Furthermore, compared to the self-attention (Vaswani et al., 2017) mechanism in ViT (Dosovitskiy et al., 2021), the convolutional layers have a stronger ability to extract local features. This enhances the extraction of local object features, thereby improving the model's ability to identify small objects in the image.

In detail, the local-enhanced features F_{local} generated by ConvNeXt are then divided into two branches: the image RGB feature branch and the spatial depth feature branch. The two branches are both several convolution layers. And we get the spatial feature F_{spatial} and the local RGB feature F_{rgb}:

$$F_{\text{spatial}} = Conv_{\text{spatial}}(F_{\text{local}}) \quad F_{\text{rgb}} = Conv_{\text{rgb}}(F_{\text{local}}) \tag{1}$$

Then, we predict the object level depth map (Huang et al., 2022) from the spatial feature:

$$M_{\text{depth}} = Conv_{\text{depth}}(F_{\text{spatial}}) \tag{2}$$

The spatial depth feature branch is capable of extracting object-level depth features and enhancing the image feature branch for extracting local object features. We use the depth map ground truth and the L1 loss to supervise the object-level depth map.

Next, we add the image RGB features and spatial depth features to derive the Spatial-Enhanced Local Feature $F_{\text{spatial-local}}$:

$$F_{\text{spatial-local}} = F_{\text{spatial}} + F_{\text{rgb}} \tag{3}$$

Subsequently, we use the global feature tokens T_{vit} obtained from the ViT to mine the Spatial-Enhanced Local Feature $F_{\text{spatial-local}}$ from the CNN adaptively. The mining process enables the input to the LLM with fewer tokens while ensuring these tokens contain enough local and spatial features. Specifically, we employ a spatial-enhanced cross-branch attention mechanism. We utilize the global feature tokens T_{vit} generated by the ViT as the query, while the Spatial-Enhanced Local Feature $F_{\text{spatial-local}}$ serves as both the key and value. Concretely, We partition T_{vit} along the token length dimension into multiple queries. We partition the Spatial-Enhanced Local Features based on the height and width dimensions for use as multiple keys and values.

$$Q = T_{\text{vit}} \times W_Q \quad K = F_{\text{spatial-local}} \times W_K \quad V = F_{\text{spatial-local}} \times W_V \tag{4}$$

$$T = Softmax(Q\,K^T/\sqrt{d_k})\,V \tag{5}$$

Finally, we obtain the spatially localized enhanced token T, and we use T as input for the large language model. The proposed spatial enhanced cross-branch attention mechanism enables the tokens extracted by ViT to attend to their respective regions of interest within the image. This improves the alignment between ViT tokens and the localized features of the image.

3.3 3D Query Token-Derived Info Decoding

When multimodal large language models handle visual perception tasks, they typically output coordinates in the form of text tokens (Wang et al., 2023b,a) or discrete coordinate bin (Peng et al., 2023; Wang et al., 2022). To accomplish 3D grounding tasks, a straightforward approach would be to output the object's 3D spatial position in text format, including coordinates such as x, y, z, length, width, height, and rotation.

However, this text-based output approach has significant issues: 1. Low Speed: Within the LLM vocabulary table, like LLaMa (Touvron et al., 2023), each digit from 0-9 is typically a separate token. For example, outputting a value of 52.3 requires four tokens (three digits and a decimal point). For 3D detection results, if the object's x, y, z coordinates, dimensions, and Euler angles for rotation are

each output as text, this could require approximately 40-50 tokens. In a decoder-only LLM, tokens are generated one at a time, resulting in low output speed. 2. Poor Accuracy: Compared to LLMs' common sense and knowledge, they are less adept at handling numbers and mathematics. Existing LLMs usually have poor numerical reasoning capabilities, resulting in significant errors when decimal coordinates are output as text. For example, LLMs typically fail to accurately comprehend the mathematical meaning of pitch, roll, and yaw Euler angles for object rotation. Outputting three Euler angles as text poses a significant challenge for LLMs. 3. Parsing Complexity: 3D detection results are complex, involving at least nine degrees of freedom, including object x, y, z coordinates, length, width, height, and three Euler angles. It is challenging to force LLMs to output these nine values in a standard format. LLMs often output too many or too few numbers or do not follow the standard format, leading to frequent anomalies in parsing results from text.

To address the above poor text-based geometric numerical output issues, we propose a 3D query token-derived info decoding method.

Specifically, in the input tokens of the large language model, in addition to the image and text tokens, we introduce a 3D query token. The 3D Query is a set of learnable parameters that have the same dimension as the hidden layer features of the LLM. In the input part of LLM, the word embedding is replaced with the 3D Query. The purpose of the 3D Query is to extract spatial information from the hidden features of the LLM. Through adaptive learning, the 3D query, acting as the query (Q) in the attention mechanism, can effectively extract image and text 3D information from self-attention. Through adaptive learning, only one 3D query token is needed for this task. After processing the 3D query token through the LLM, the final hidden feature of the token is employed as the 3D feature F_{3D}.

Additionally, to determine when to use the 3D query token, we introduce a special `<pos>` token. The `<pos>` token is placed before the 3D query in the sequence. When the `<pos>` token is detected in the LLM's output, the subsequent next input token is replaced with the learnable 3D query token.

Unlike directly using LLM's text output to determine the object's spatial position, we employ an extra regression head to output the object's spatial position. Specifically, after obtaining the 3D Feature F_{3D}, we regress the object's 3D center projection on the image p_{img}, depth d_{v}, 3D size (length l, width w, height h), and rotation angles.

Specifically, for the object's center, we do not use the center of the 2D bounding box. Instead, we utilize the projection point of the object's 3D center onto the image, as the 3D center is more suitable for the subsequent geometric inverse projection process.

We normalize the width and height of the image to the range [0, 1], using an MLP to predict the relative position of the 3D center's projection point on the image, denoted as $p_{\text{norm}} = (u_{\text{norm}}, v_{\text{norm}})$:

$$u_{\text{norm}}, v_{\text{norm}} = \text{MLP}_{uv}(F_{3D}) \tag{6}$$

And the actual position of the 3D center projection point on the image: $p = (u, v)$:

$$u = u_{\text{norm}} \times w \quad v = v_{\text{norm}} \times h \tag{7}$$

where w, h are the image's width and height.

For the 3D size of objects in 3D space: length L, width W, and height H, we use an MLP to predict these dimensions:

$$L, W, H = \text{MLP}_{LWH}(F_{3D}) \tag{8}$$

For the depth d_v, we similarly use an MLP to predict this value:

$$d_v = \text{MLP}_{d}(F_{3D}) \tag{9}$$

For the rotation of objects in space, previous works (Cho et al., 2024b) directly predicted the Euler angles of objects. However, using neural networks to predict Euler angles directly poses several issues: 1. Discontinuity Issues: Euler angle representation is prone to singularities, also known as gimbal lock. This occurs when certain angles approach specific values, leading to indistinguishability between the angles, causing numerical instability and making it difficult for the model to learn. 2. Non-Uniqueness of Representation: Each rotation can have multiple Euler angle representations, leading to the multiple solutions problem. This non-uniqueness increases error and makes it challenging for the model to converge during training. 3. Complexity of Loss Function: Using Euler angles as outputs requires considering the periodicity and ambiguity of angles when calculating the loss, complicating the design of the loss function. 4. Asymmetry Issues: The range and symmetry of Euler angle components differ, making some angles more influential on the results than others. Neural networks may become more sensitive to certain angles while being less sensitive to others, affecting accuracy.

Previous works (Lu et al., 2021; Shi et al., 2021) focused on outdoor autonomous driving datasets, and they neglected the pitch and roll angles. They only predicted yaw angle, so the disadvantages of Euler angles were not apparent. However, in indoor datasets, such as SUNRGBD (Song et al., 2015) and Objectron (Ahmadyan et al., 2021) datasets, all three Euler angles (pitch, roll, yaw) are significant and impactful. The aforementioned issues of Euler angles become very apparent. Similarly, quaternions

are also discontinuous and difficult for neural networks to learn (Zhou et al., 2019).

To address these issues, we predict the 6D allocentric rotation (Zhou et al., 2019), which is continuous in 6D space and more suitable for learning. Specifically, we use an MLP to predict the object's 6D allocentric rotation representation:

$$\text{Rot}_{6D} = \text{MLP}_{6D}(F_{3D}) \tag{10}$$

We then convert the 6D rotation representation into a 3×3 rotation matrix:

$$\text{Rot} = \text{rotation\_6d\_to\_matrix}(F_{3D}) \tag{11}$$

In this section, we employ the learnable 3D query to extract 3D features from the large language model. Subsequently, we use 3D heads to regress the geometric attributes, including objects' image projection point, depth, dimensions (length, width, height), and rotation. We will elaborate on the reasoning of the object's 3D bounding box, particularly the inference of the object's 3D spatial position X, Y, and Z, in Section 3.4.

3.4 Geometry Projection-Based 3D Reasoning

The objective of 3D grounding is to output the object's 3D bounding box in space, where the origin is located at the camera center. A straightforward approach might involve directly predicting the object's X, Y, Z coordinates, dimensions, and rotation angles. However, image-based 3D grounding is inherently underdetermined. When provided with only a single image, the neural network lacks access to critical camera parameters, such as precise focal length and field of view, which significantly impact image-based 3D grounding.

The focal length and other intrinsic camera parameters significantly influence image-based 3D grounding. As illustrated in Figure. 2 (c), two objects may appear to have the same size and similar positions in a 2D image. Consequently, neural networks are prone to predict these two objects to be at the same 3D spatial location. However, these images are captured with different focal lengths, resulting in significant differences in the actual spatial positions of the objects. Therefore, neglecting the variations in camera parameters and relying solely on neural network predictions for object spatial positions can lead to substantial errors, particularly under varying focal lengths. Moreover, predicting 3D metrics (X, Y, Z) directly from a 2D image is highly challenging, as each value can introduce substantial errors in spatial positioning.

To resolve these issues, we propose geometry projection-based 3D Reasoning. Instead of directly predicting the 3D metrics X, Y, Z, we predict the 2D projection of the object's 3D center onto the image and its virtual depth, as mentioned in section 3.3.

Camera intrinsics greatly affect depth prediction. To mitigate this, we assume that all input images are captured by a virtual camera with unified focal length and resolution (Brazil et al., 2023). We do not regress the actual depth of the object but instead, regress the virtual depth under the virtual camera. Subsequently, we convert this virtual depth back to the actual depth of the object. Specifically, since that image width is typically greater than height and multimodal large models usually resize images to the same resolution along the width, we improved the calculation method for virtual depth by using width as a reference.

For a point P^i in space and its projection p^i on the image, we assume that in the virtual camera, P^v is projected to p^v, where P^v and P^i have the same X and Y values but differ in depth Z. The points p^v and p^i are at the same position on the image.

Assuming the intrinsic matrix of real camera C^i are $\mathbf{K^i}$ and the width of the real image is w^i, the projection of a point $P^i = (X^i, Y^i, Z^i)$ in 3D space onto the image, denoted as $p^i = (x^i, y^i)$, is given by the following equation:

$$Z^i[x^i, y^i, 1]^\top = \mathbf{K^i}[X^i, Y^i, Z^i]^\top \tag{12}$$

$$\mathbf{K^i} = \begin{bmatrix} f_x^i & 0 & c_x^i \\ 0 & f_y^i & c_y^i \\ 0 & 0 & 1 \end{bmatrix} \tag{13}$$

where $f_x^i, f_y^i, c_x^i, c_y^i$ are all the intrinsics of the real camera C^i.

From Equations 12 and 13, we derive:

$$x^i \cdot Z^i = f_x^i \cdot X^i + c_x^i \cdot Z^i \tag{14}$$

We assume the virtual camera intrinsic matrix is $\mathbf{K^v}$ and image width is w^v. The point P^v in the virtual camera differs from the point P^i in the real camera only in the Z-coordinate, with the X and Y coordinates being the same. Therefore, we can denote P^v as $P^v = (X^i, Y^i, Z^v)$. The projection of $P^v = (X^i, Y^i, Z^v)$ onto the image is denoted as $p^v = (x^v, y^v)$:

$$Z^v[x^v, y^v, 1]^\top = \mathbf{K^v}[X^i, Y^i, Z^v]^\top \tag{15}$$

$$\mathbf{K^v} = \begin{bmatrix} f_x^v & 0 & c_x^v \\ 0 & f_y^v & c_y^v \\ 0 & 0 & 1 \end{bmatrix} \tag{16}$$

where $f_x^v, f_y^v, c_x^v, c_y^v$ are all the intrinsics of the virtual camera.

Similarly, from Equations 15 and 16, we derive:

$$x^v \cdot Z^v = f_x^v \cdot X^i + c_x^v \cdot Z^v \tag{17}$$

The projection points $p^v = (x^v, y^v)$ under the virtual camera and $p = (x^i, y^i)$ under the real camera correspond to the same location. And the principal point c_x^v for the virtual camera corresponds to the principal point of the real camera:

$$x^v = \frac{x^i}{w^i} \cdot w^v, \qquad c_x^v = \frac{c_x^i}{w^i} \cdot w^v \tag{18}$$

Substituting Equations 18 into Equation 17:

$$\frac{x}{w^i} \cdot w^v \cdot Z^v = f_x^v \cdot X + \frac{c_x^i}{w^i} \cdot w^v \cdot Z^v \tag{19}$$

Thus:

$$x = f_x^v \cdot \frac{X}{Z^v} \cdot \frac{w^i}{w^v} + c_x \tag{20}$$

By substituting Equation 20 into Equation 14, we obtain:

$$\left(f_x^v \cdot \frac{X}{Z^v} \cdot \frac{w^i}{w^v} + c_x \right) \cdot Z = f_x^i \cdot X + c_x^i \cdot Z \tag{21}$$

Simplifying Equation 21, we derive the virtual depth:

$$Z^v = \frac{f_x^v}{f_x^i} \cdot \frac{w^i}{w^v} \cdot Z \tag{22}$$

where f_x^i, f_x^v are the focal length of the real camera and virtual camera, respectively. w^i, w^v are image widths of the real camera and virtual camera. Z is the object's real depth.

Virtual depth $Z^v = \frac{f_x^v}{f_x} \cdot \frac{w^i}{w^v} \cdot Z$ accounts for the effects of different focal lengths and image sizes, enabling the neural network to make predictions invariant to camera parameter discrepancies.

Thus, in Section 3.3 3D Query Token-Derived Info Decoding, we do not regress the object's actual depth. Instead, we regress the object's virtual depth and subsequently convert it back to the actual depth. Therefore, in Section 3.3, the depth d_v output by the MLP represents the virtual depth Z^v.

Next, we use the inverse of Equation 22 to convert the virtual depth predicted by the neural network back to the actual depth:

$$Z_1 = d_v \cdot \frac{f_x}{f_x^v} \cdot \frac{w^v}{w} \tag{23}$$

where f_x and w are the focal length and width of the real camera image, and f_x^v and w^v are the focal length and width of the virtual camera image.

Moreover, in outdoor autonomous driving scenes, the range of object depths can vary significantly, making prediction highly challenging. Therefore, in these datasets, we also utilize geometric projection constraints in the Y direction to predict a second independent depth value. We apply the projection equations (Equations 12 and 13) to two points with the same depth at the top and bottom of the 3D bounding box:

$$y_1 \cdot Z = f_y \cdot Y_1 + c_y \cdot Z \quad y_2 \cdot Z = f_y \cdot Y_2 + c_y \cdot Z \tag{24}$$

From Equation 24, we get:

$$(y_1 - y_2) \cdot Z = f_y \cdot (Y_1 - Y_2) \tag{25}$$

In outdoor autonomous driving datasets, the pitch and roll angles of objects are typically very small. Therefore, $(y_1 - y_2)$ can be approximated as the 2D height h of the object, and $(Y_1 - Y_2)$ can be approximated as the 3D height H of the object. Thus, we derive the object's second depth value:

$$Z_2 = \frac{H}{h} \cdot f_y \tag{26}$$

Finally, we combine the virtual depth regression value Z_1 with the projected depth Z_2, taking the average of Z_1 and Z_2 as the object's final predicted depth Z to enhance the accuracy of depth prediction.

$$Z = \frac{Z_1 + Z_2}{2} = \left(d_v \cdot \frac{f_x}{f_x^v} \cdot \frac{w^v}{w} + \frac{H}{h} \cdot f_y \right) / 2 \tag{27}$$

Currently, we have obtained the Z-coordinate of the 3D bounding box center P. Next, we need to determine the X and Y coordinates of P. Using the 3D center image projection point $p = (u, v)$ predicted in Section 3.3 3D Query Token-Derived Info Decoding, along with the depth Z, we can calculate the X and Y coordinates of P. According to the projection formula:

$$Z[u, v, 1]^\top = \mathbf{K}[X, Y, Z]^\top \tag{28}$$

$$\mathbf{K} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{29}$$

Here, $\mathbf{K}$ represents the camera intrinsic matrix, with f_x, f_y, c_x, and c_y being the camera intrinsic parameters.

Then, we can get:

$$X = \frac{Z}{f_x} \cdot (u - c_x), \quad Y = \frac{Z}{f_y} \cdot (v - c_y) \tag{30}$$

According to Equation. 27, Substituting into the formulas above, we can obtain:

$$X = \left(\frac{d_v}{f_x^v} \cdot \frac{w^v}{w} + \frac{H}{h} \cdot \frac{f_y}{f_x} \right) \cdot (u - c_x)/2 \tag{31}$$

$$Y = \left(\frac{f_x}{f_y} \cdot \frac{d_v}{f_x^v} \cdot \frac{w^v}{w} + \frac{H}{h} \right) \cdot (v - c_y)/2 \tag{32}$$

$$Z = \left(d_v \cdot \frac{f_x}{f_x^v} \cdot \frac{w^v}{w} + \frac{H}{h} \cdot f_y \right) / 2 \tag{33}$$

Consequently, we derive the expression for the coordinates of the 3D bounding box center $P = (X, Y, Z)$, where all terms are either the outputs from the neural network or known quantities.

Additionally, in Section 3.3 3D Query Token-Derived Info Decoding, we obtained the dimensions of the 3D bounding box: length L, width W, and height H, as well as the rotation matrix Rot. Therefore, we can obtain the final 3D bounding box of the object.

3.5 IG3D: Image-Based 3D Grounding Dataset

To accomplish the task of 3D grounding, we require 3D grounding training data that includes 3D bounding boxes of objects as well as descriptions of those objects. However, existing 3D datasets such as KITTI (Geiger et al., 2012), nuScenes (Caesar et al., 2020), Waymo (Ettinger et al., 2021), SUNRGBD (Song et al., 2015), and Objectron (Ahmadyan et al., 2021), only provide 3D bounding boxes and object categories without any descriptive texts about the objects. When an image contains multiple objects of the same category, it is impossible to accurately identify which object is being referred to based solely on the object category.

Moreover, in applications of embodied intelligence and robotics, tasks such as Visual Question Answering (VQA) require the agent to possess common sense and perform logical reasoning based on user inquiries to identify and respond to specific objects. Current 3D detection datasets only provide object categories and lack data or annotations related to logical reasoning and question answering. The recently released Mono3DRefer (Zhan et al., 2024) dataset in 2024 also has significant issues. Mono3DRefer uses the results of 3D perception, such as object depth and dimensions (length, width, height), as input descriptions. This allows the model to directly obtain 3D perceptual answers based on the caption texts. As a result, it is difficult for Mono3DRefer to assess the model's real ability of 3D perception from images. Additionally, Mono3DRefer lacks question-answer data, making it unsuitable for evaluating the model's world knowledge and logical reasoning capabilities.

To address these issues, we constructed the IG3D dataset. The IG3D dataset provides detailed descriptions of objects in images, including detailed information on appearance and location, allowing us to distinguish between different objects of the same category within an image. This enables the execution of 3D grounding tasks. Additionally, the IG3D dataset includes visual question answering instructions. The input is a personalized question, and the output should be the object of interest in text form and its 3D box in space.

Specifically, we use an auto-annotation technique based on multimodal large models to label our dataset efficiently. Initially, we obtain images and the corresponding 2D bounding boxes box_{2D} and 3D bounding boxes box_{3D} from a 3D detection dataset. We draw 2D bounding boxes around objects of interest in the images and input these images into a pre-trained multimodal large language model (we used a frozen Mini-Gemini-34B (Li et al., 2024)).

As shown in Figure 4, we expect the multimodal large language model to provide an accurate description of the object of interest, including color, shape, position in the image, etc. Particularly, if there are multiple objects of the same category in the image, we hope the model describes the appearance and position of the object to distinguish it from others of the same category. For instance, the model might indicate that the object is the second one from the left in the image.

We use the 3D bounding box box_{3D} from the 3D object detection dataset as the object's 3D bounding box label. Finally, using the precise object descriptions and box_{3D}, we construct the IG3D dataset, as illustrated in Figure 4.

Moreover, to test the model's Visual Question Answering capabilities, we constructed an additional VQA section of the dataset. Specifically, we use GPT-4o to automatically generate a large number of questions for objects. We instructed GPT-4o to generate questions whose answers should be the specific object. For example, when asked, "I want to read a book, but the light is too dark. Which object should I use?", the model not only needs to provide the answer "the lamp" in text form but also needs to identify the lamp's 3D bounding box in space. This evaluates the model's common sense and logical reasoning ability, which is highly beneficial for applications in embodied intelligence and robotics. Examples of categories and questions from the IG3D-VQA dataset are provided in Table 2. The answers to these questions are the specific objects in the image, in the form of textual responses and their 3D bounding boxes in space.

In summary, we constructed the 3D Grounding datasets based on the following 3D detection datasets: SUNRGBD (Song et al., 2015), nuScenes (Caesar et al., 2020), KITTI (Geiger et al., 2012), and Objectron (Ahmadyan et al., 2021), and we name them as IG3D-SUNRGBD, IG3D-nuScenes, IG3D-KITTI, and IG3D-Objectron respectively. Additionally, we created a 3D Visual Question Answering (VQA) dataset from the SUNRGBD dataset, referred to as IG3D-SUNRGBD-VQA. For the training, validation, and test splits of all these datasets, we followed the divisions established by Omni3D (Brazil et al., 2023). Furthermore, due to the inevitable errors in object descriptions provided by existing multimodal models, we filtered out many problematic annotations to enhance the quality of our datasets.

4 Experiments

In this section, we first introduce the experimental settings, including datasets, metrics, baselines, and the implementation details of our method. We then conducted a comprehensive comparison of our method with existing state-of-the-art approaches, followed by an extensive ablation study and visualization results.

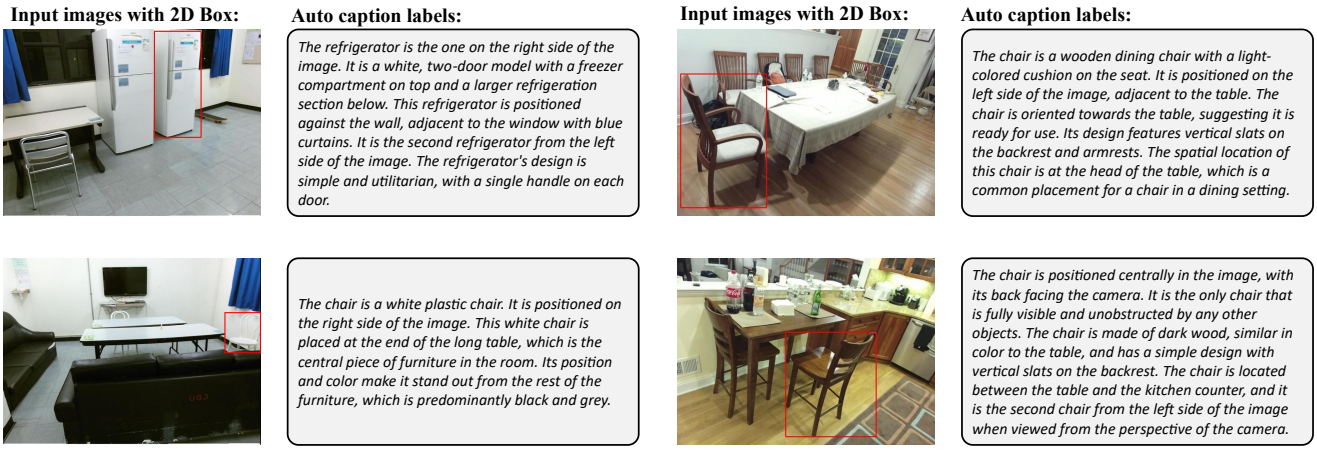

Fig. 4 Some examples of our IG3D 3D grounding datasets. Given an input image and a 2D box around the object of interest, we employ a pre-trained frozen MLLM to automatically generate a caption describing the object. This automatically generated caption serves as the object description part of our IG3D grounding dataset. Additionally, we utilize the 3D bounding box labels from the original 3D detection dataset as the 3D box ground truth for our IG3D dataset.

Table 2 Some category and question examples in our IG3D-SUNRGBD-VQA dataset. The answers to these questions are the specific objects in the image, in the form of textual responses and their 3D bounding boxes in space.

Categories	Question Examples
camera	I want to take a picture, which object should I use? Which object is a device for taking photographs or making films? What piece of equipment is essential for recording videos? Which device typically has a lens for capturing images?
chair	What is often used to provide seating in offices and homes? What do people usually sit on while working at a desk? What piece of furniture has a seat, back, and sometimes armrests? What piece of furniture is commonly found at a dining table for seating?
window	I want to look outside, which object should I use? I want to breathe fresh air, which object should I use? Which object is on the wall, with glass to allow light or air to enter and allow people to see outside? What allows natural light to enter a room without opening a door?
mirror	What is used to reflect an image clearly on a flat surface? What is often used in dance studios for students to watch themselves? What do people often look into to see a reflection of themselves? What is employed in physics experiments to reflect light?
bicycle	What is a two-wheeled vehicle powered by human pedaling? Which object is a two-wheeled, pedal-driven, environmentally friendly transportation vehicle? Which object is a two-wheeled small land vehicle, powered by pedaling? What vehicle requires pedals and chains to operate?

4.1 Experimental Settings

In this section, the datasets, metrics, baselines, and implementation details are described.

4.1.1 Datasets, Metrics, and Baselines

We conducted experiments on multiple datasets, including 3D grounding datasets: IG3D-SUNRGBD, IG3D-nuScenes, IG3D-KITTI, IG3D-Objectron, Mono3DRefer (Zhan et al., 2024), as well as the 3D VQA dataset IG3D-SUNRGBD-VQA. These datasets are based on 3D detection datasets with added descriptions or questions about the objects,

creating grounding or VQA scenarios. Specifically, SUN-RGBD and Objectron are indoor scene datasets with a rich variety of categories. In contrast, nuScenes, KITTI, and Mono3DRefer are outdoor autonomous driving datasets with fewer categories but more expansive and open scenes. In the experiments, a 2D image is provided as input, along with descriptions or questions about the objects within the image. The model is required to output the 3D bounding box of the object(s) of interest in the image.

In real-world scenarios, object categories are vast, with many rare categories and long-tail distribution issues (Yin et al., 2024; Yang et al., 2024b). Therefore, we conducted open-vocabulary 3D grounding experiments to evaluate the model's ability to perform 3D localization of novel categories. Additionally, we performed domain generalization experiments to assess the model's robustness to domain shifts in real-world settings (Chen et al., 2024).

Regarding the evaluation metrics, we follow the standards set in Mono3DVG (Zhan et al., 2024). Specifically, "Acc@0.25" considers a prediction correct if the Intersection over Union (IoU) between the predicted 3D bounding box and the ground truth exceeds 25%, while "Acc@0.5" requires an IoU greater than 50%. It is important to note that predicting 3D bounding boxes with high accuracy is very challenging. Even slight errors in the position or size of the 3D box can cause a significant drop in IoU, leading to generally low accuracy values across various methods.

"DepthError" measures the average error in depth between the predicted and ground truth 3D bounding boxes in real-world coordinates, expressed in meters. "LengthError," "WidthError," and "HeightError" assess the discrepancies in length, width, and height between the predicted and ground truth 3D bounding boxes, respectively, also in meters. It is worth noting that large "DepthError" values typically originate from the model misidentifying the object, resulting in the predicted 3D bounding box not corresponding to the object of interest, thereby causing large depth errors.

"Text3D" refers to the results obtained from a vanilla multimodal large language model, which outputs 3D bounding boxes in the form of text tokens. "TransVG+backproj" follows the baseline from Mono3DVG (Zhan et al., 2024), using 2D vision grounding combined with back-projection to adapt the results to 3D. We extended TransVG by incorporating a simple depth predictor to estimate object depth. Mono3DVG is currently the state-of-the-art image-based 3D grounding method. LLMI3D refers to our proposed method.

4.1.2 More Implement Details

Since we utilize pre-trained multimodal large language models, we perform parameter-efficient fine-tuning using LoRA (Hu et al., 2022) instead of full parameter training. Specifically, we use LoRA to fine-tune Mini-Gemini (Li et al., 2024). The rank of LoRA is set to 64, and alpha is set to 16. For the LLM component, we employ the Vicuna-7B version (Chiang et al., 2023). We use the L1 loss for the 3D regression heads. We follow the fine-tuning hyper-parameters set by Mini-Gemini, using the AdamW optimizer and the learning rate 2e-5. The batch size is 4, and gradient accumulation steps are set to 4. The fine-tuning process is conducted with two NVIDIA A100 GPUs.

4.2 Comparison with State-of-the-art

To evaluate our method against others, we conducted experiments on multiple datasets and under various experimental settings. Tables 3, 4, 5, 6, 7, and 8 compare our method with other state-of-the-art methods.

4.2.1 3D Grounding and 3D VQA Results

As shown in Table 3, we present a comparison of our method with other methods on the 3D grounding and 3D VQA task across several datasets: IG3D-SUNRGBD, IG3D-SUNRGBD-VQA, IG3D-nuScenes, IG3D-KITTI, and IG3D-Objectron. As observed in Table 3, our method outperforms existing state-of-the-art approaches by a considerable margin.

"TransVG+backproj" performs poorly across all datasets, highlighting the significant differences between 2D and 3D grounding tasks and the inherent difficulties of 3D grounding. Unlike 2D grounding, which is relatively straightforward as it only requires locating objects within an image, image-based 3D grounding involves predicting 3D bounding boxes in real-world coordinates from 2D images, requiring estimates of object depth and physical dimensions. This makes image-based 3D grounding a highly under-constrained problem when dealing with a single image. Without specialized optimization for 3D grounding, combining 2D grounding with simple back-projection is insufficient for accurately locating objects in 3D space.

"Text3D" also performs poorly across various datasets. It directly uses a large language model to output the object's X, Y, Z position coordinates, as well as its dimensions and Euler angles, which is unsuitable for multimodal large language models. These models excel in general knowledge and common sense but struggle with numerical and mathematical concepts. Consequently, the large language model can only rely on training data to guess about objects' spatial positions. "Text3D" performs particularly poorly on outdoor datasets with large scene spaces, such as nuScenes and KITTI.

Moreover, current approaches for monocular 3D object detection and multi-view BEV object detection predominantly target outdoor autonomous driving datasets. The state-of-the-art method Mono3DVG (Zhan et al., 2024), is also primarily designed for outdoor scenes, explaining its reasonable performance on the KITTI and nuScenes datasets. However, Mono3DVG's performance diminishes on the indoor SUNRGBD dataset. From Table 3, it is evident that our method significantly outperforms Mono3DVG in indoor environments. This is because KITTI and nuScenes contain fewer scene categories, allowing smaller models like Mono3DVG to achieve reasonable accuracy through overfitting. At the same time, SUNRGBD encompasses a broader variety of scenes and object categories, presenting a more challenging grounding task. Additionally, outdoor datasets often default to zero pitch and roll angles for objects, whereas indoor datasets like SUNRGBD encompass significant pitch and roll variations. Mono3DVG predicts only the yaw angle using a bin classification approach, neglecting pitch and roll angles. Our method, utilizing a 6D allocentric rotation approach, is better suited for indoor datasets.

Specifically, the IG3D-SUNRGBD-VQA dataset is a question-answer dataset. It aims to test the model's problem-solving capabilities in real-world scenarios involving embodied intelligence and robotics, where the model must handle various complex questions from users. To effectively answer the questions in the IG3D-SUNRGBD-VQA dataset, the model requires a certain degree of logical reasoning ability. Originally, IG3D-SUNRGBD-VQA dataset requires the model to provide answers about objects in the image in the form of textual responses, as well as output their corresponding 3D bounding boxes. However, only large language models are capable of generating textual answers, whereas smaller models lack this ability. To enable a fair comparison of different methods, we did not evaluate the textual answers produced by the models. Instead, we only assessed the 3D bounding boxes of the objects of interest that each model outputs. On the SUNRGBD-VQA dataset, both TransVG+backproj and Mono3DVG show a marked decline in performance. This indicates that smaller models like TransVG or Mono3DVG are almost incapable of handling VQA tasks that require logical reasoning and basic common sense. They fail not only in generating textual answers but also in producing accurate 3D bounding boxes. Only our method, built upon a large language model, maintains strong performance and shows a minimal decrease from IG3D-SUNRGBD to IG3D-SUNRGBD-VQA, demonstrating robustness against complex questions.

4.2.2 Open Vocabulary 3D Grounding

The real world comprises an infinite number of categories, including numerous rare sample classes and corner cases (Hao et al., 2024a; Weng et al., 2024; Feng et al., 2024). Existing works and methods predominantly evaluate the accuracy of models on known categories encountered during training without testing their performance on open-world categories (Weng et al., 2024).

To evaluate the models' capability in open-world scenarios, we tested their zero-shot open-vocabulary 3D grounding performance. Specifically, we divided the categories in datasets such as SUNRGBD, nuScenes, KITTI, and Objectron into 80% base classes and 20% novel classes. We trained all models on the base classes and assessed their 3D grounding performance on the novel classes, as illustrated in Table 4.

In Table 4, our model demonstrates significantly higher accuracy than existing works. Models like Mono3DVG, when trained on base classes, perform well on these classes but exhibit poor performance on novel classes. This indicates the inability of specialized small models like Mono3DVG to handle open-world scenarios and address the vast array of object categories in real-world settings.

MLLMs inherently possess significant advantages over specialized small models in open vocabulary tasks. When faced with new categories, small models struggle to ascertain the 3D spatial sizes of new classes, failing to accurately determine their dimensions. In contrast, large models are endowed with world knowledge and common sense, enabling them to understand the actual sizes of novel categories in the physical world and infer their physical positions from the environment. Thus, large models outperform small models significantly in open vocabulary tasks.

4.2.3 Domain Generalization for the 3D Grounding

In real-world scenarios, we not only encounter novel categories but also face the challenge of domain shift (Lehner et al., 2024; Rodriguez and Mikolajczyk, 2023; Rodriguez et al., 2023). The domain gap in 3D localization is significantly more complex than in 2D scenarios. In 2D scenarios, the domain gap typically includes variations in image appearance, such as different image styles. However, the domain gap in 3D scenarios encompasses not only differences in 2D image appearance, such as style or scene variations across different datasets, but also discrepancies at the camera level. Different datasets use different cameras with substantial differences in focal lengths, resolutions, and fields of view. For instance, early datasets like KITTI have lower resolutions than newer datasets like nuScenes. Variations in camera parameters such as focal length and

Table 3 Comparison of our LLMI3D with other methods on the IG3D-SUNRGBD, IG3D-SUNRGBD-VQA, IG3D-nuScenes, IG3D-KITTI, and IG3D-Objectron datasets.

dataset	Method	Acc@0.25↑	Acc@0.5↑	DepthError↓	LengthError↓	WidthError↓	HeightError↓
IG3D-SUNRGBD	TransVG + backproj	5.6	0.4	0.88	0.57	0.84	0.59
	Text3D	11.5	1.7	0.45	0.20	0.34	0.21
	Mono3DVG	25.2	6.8	0.53	0.14	0.26	0.16
	LLMI3D	**42.3**	**11.8**	**0.32**	**0.12**	**0.21**	**0.12**
IG3D-SUNRGBD-VQA	TransVG + backproj	2.2	0.2	0.97	0.71	0.93	0.77
	Text3D	7.8	1.0	0.56	0.24	0.45	0.29
	Mono3DVG	9.8	1.4	0.63	0.21	0.41	0.27
	LLMI3D	**35.1**	**8.6**	**0.36**	**0.16**	**0.28**	**0.17**
IG3D-nuScenes	TransVG + backproj	8.6	3.5	7.51	2.28	0.77	0.82
	Text3D	13.7	5.2	4.25	1.75	0.28	0.27
	Mono3DVG	27.5	9.8	2.80	0.55	0.19	0.21
	LLMI3D	**31.6**	**13.2**	**2.19**	**0.50**	**0.16**	**0.17**
IG3D-KITTI	TransVG + backproj	2.9	0.3	8.42	1.39	0.31	0.35
	Text3D	5.4	0.7	4.15	0.70	0.16	0.17
	Mono3DVG	27.7	7.74	2.08	0.44	0.13	0.14
	LLMI3D	**32.4**	**10.3**	**1.56**	**0.34**	**0.11**	**0.11**
IG3D-Objectron	TransVG + backproj	23.0	6.7	0.14	0.05	0.03	0.05
	Text3D	35.4	10.7	0.08	**0.03**	**0.02**	0.04
	Mono3DVG	45.5	12.4	0.09	**0.03**	**0.02**	**0.03**
	LLMI3D	**55.6**	**18.7**	**0.05**	**0.03**	**0.02**	**0.03**

Table 4 In the open vocabulary 3D grounding task, results comparison of our LLMI3D with other methods on the IG3D-SUNRGBD, IG3D-nuScenes, IG3D-KITTI, and IG3D-Objectron datasets.

dataset	Method	Acc@0.25↑	Acc@0.5↑	DepthError↓	LengthError↓	WidthError↓	HeightError↓
IG3D-SUNRGBD	Text3D	8.6	1.5	0.52	0.31	0.47	0.24
	Mono3DVG	19.4	2.4	0.54	0.21	0.40	0.18
	LLMI3D	**40.1**	**4.4**	**0.38**	**0.19**	**0.29**	**0.11**
IG3D-nuScenes	Text3D	8.2	1.9	5.50	4.81	0.47	0.75
	Mono3DVG	10.5	2.2	5.42	3.45	0.58	0.93
	LLMI3D	**30.5**	**7.5**	**3.11**	**2.56**	**0.36**	**0.57**
IG3D-KITTI	Text3D	3.1	0.3	5.30	3.88	0.37	0.62
	Mono3DVG	2.2	0.2	5.97	1.88	0.46	0.51
	LLMI3D	**30.1**	**8.1**	**1.84**	**0.84**	**0.17**	**0.26**
IG3D-Objectron	Text3D	20.5	6.0	0.08	0.04	**0.03**	**0.06**
	Mono3DVG	6.4	0.1	0.35	0.07	0.15	0.16
	LLMI3D	**32.8**	**7.9**	**0.07**	**0.03**	**0.03**	**0.06**

Table 5 In the domain generalization for the 3D Grounding task, results comparison of our LLMI3D with other methods. nuScenes→KITTI refers to models trained on the nuScenes dataset and tested on the KITTI dataset, while KITTI→nuScenes represents models trained on the KITTI dataset and tested on the nuScenes dataset.

dataset	Method	Acc@0.25↑	DepthError↓	LengthError↓	WidthError↓	HeightError↓
nuScenes → KITTI	Text3D	5.0	4.98	0.67	0.37	0.20
	Mono3DVG	16.5	2.24	0.84	0.29	0.16
	LLMI3D	**23.1**	**2.11**	**0.59**	**0.20**	**0.13**
KITTI → nuScenes	Text3D	0.9	8.47	0.72	0.65	0.62
	Mono3DVG	4.3	5.30	0.62	0.35	0.24
	LLMI3D	**20.5**	**3.32**	**0.54**	**0.29**	**0.22**

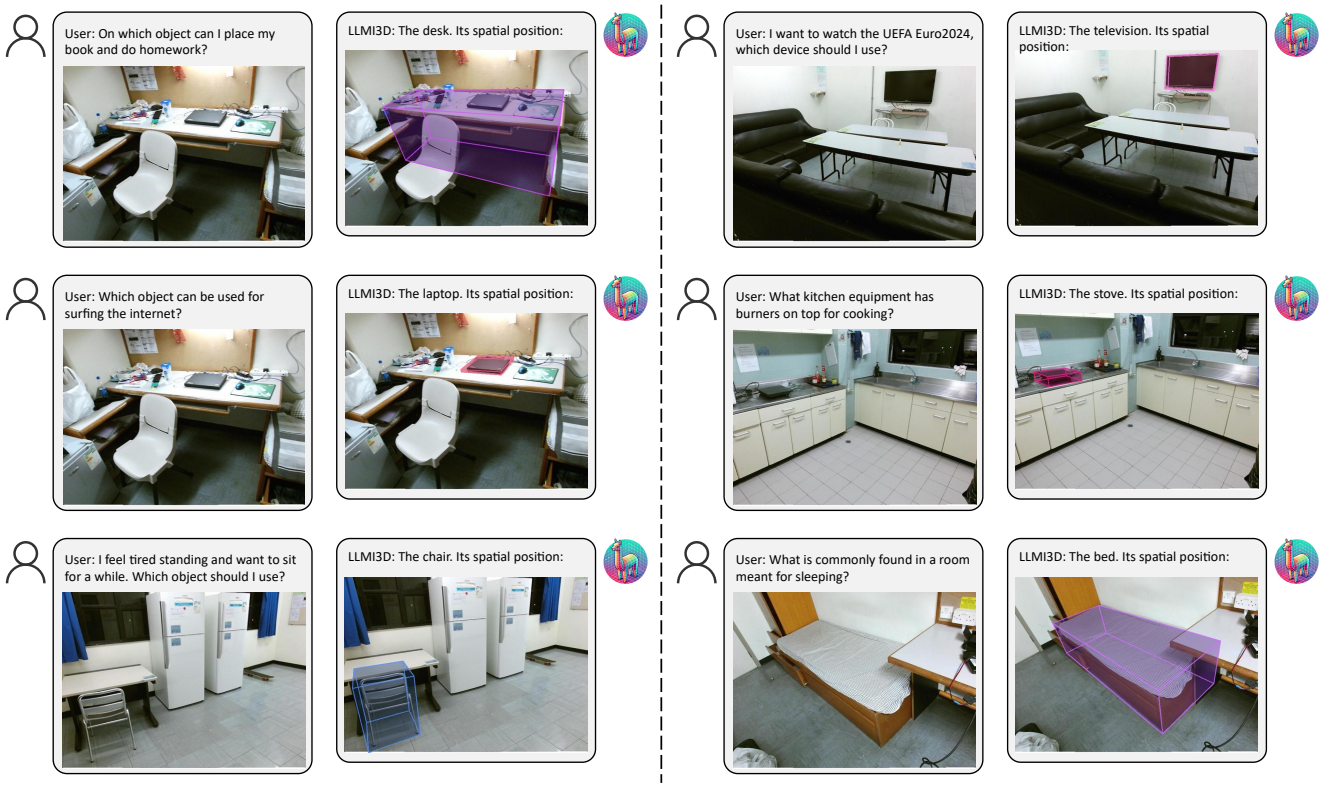

Fig. 5 Some examples of the 3D VQA visualizations of our LLMI3D in the IG3D-SUNRGBD-VQA dataset. Our LLMI3D is capable of understanding user-input personalized questions, leveraging common knowledge and logical reasoning to identify objects of interest, and returning the corresponding 3D bounding boxes.

field of view have minimal impact on 2D localization but significantly affect 3D localization (Yang et al., 2024a).

In Table 5, we present experiments conducted under different domain generalization settings. The nuScenes → KITTI setting involves training on the nuScenes dataset and testing on the KITTI dataset, while the KITTI → nuScenes setting involves training on the KITTI dataset and testing on the nuScenes dataset.

The results in Table 5 show that both the Text3D experiment, which directly outputs text tokens, and the specialized 3D grounding small model Mono3DVG experience significant accuracy drops under domain generalization settings. When the camera changes, these models struggle to discern focal length variations solely based on images, continuing to infer object spatial positions using the focal lengths from the training set. They fail to adapt to focal length variations for accurate spatial position inference. Under domain generalization settings, our method, benefiting from the geometry projection-based 3D Reasoning, treats images as results of virtual camera imaging and predicts virtual depth under a unified camera. Consequently, our approach significantly outperforms the baselines.

4.2.4 Results on Mono3DRefer Dataset

Finally, to demonstrate the strong robustness of our model, we conducted experiments on the existing dataset Mono3DRefer (Zhan et al., 2024). Mono3DRefer is the latest dataset proposed by Zhan et al. in 2024. We followed various metrics proposed in Mono3DVG for evaluation. Specifically, "unique" refers to images with only one object of the given category, and "multiple" refers to images containing multiple objects of the given category. We adhered to the baselines set by Mono3DVG, comparing our results against methods such as ZSGNet (Sadhu et al., 2019), FAOA (Yang et al., 2019), ReSC (Yang et al., 2020), and TransVG (Deng et al., 2021), which are all 2D grounding methods. These 2D grounding methods, when combined with back projection, performed poorly on the Mono3DRefer dataset, highlighting the difficulty of the 3D task. The 3D grounding task requires models to understand geometrical spatial structures, a capability that 2D grounding models do not possess. While the state-of-the-art method Mono3DVG demonstrated better performance, it still fell short compared to our LLMI3D. Our LLMI3D surpassed all existing methods, achieving the best accuracy across various metrics, thus proving the generalization capability of our model on diverse datasets.

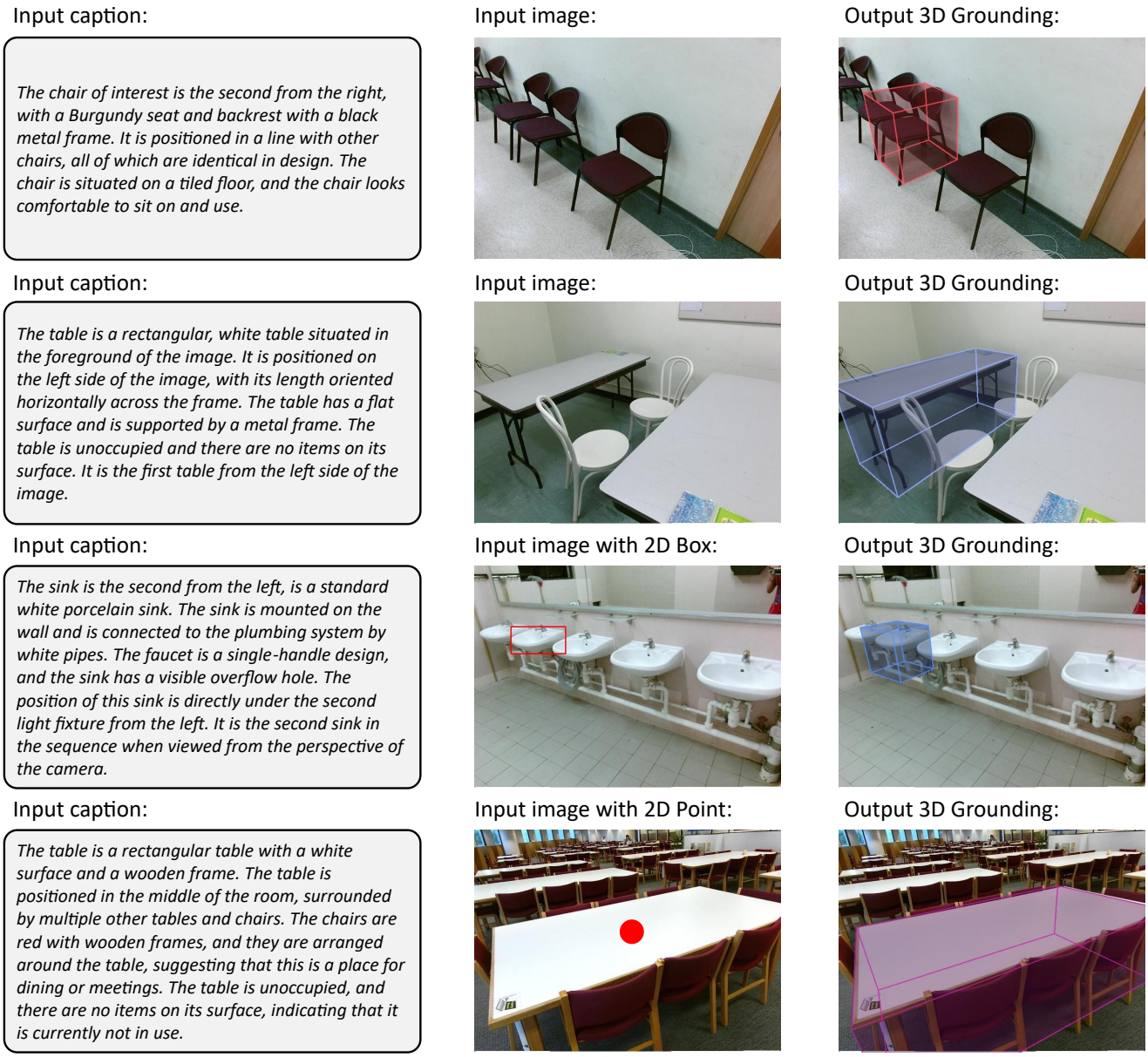

Fig. 6 The 3D grounding visualizations of our LLMI3D in the IG3D-SUNRGBD dataset. When users input an image and a caption describing the object of interest, our LLMI3D can return the corresponding 3D bounding box. Moreover, our LLMI3D can accept various input prompts, such as a caption with a 2D bounding box or a caption with a 2D point, and output the corresponding 3D bounding box, thereby accommodating various user input forms.

4.2.5 Various Types of Input Prompts

Table 7 presents the experimental results when the input prompt includes both the caption and the 2D box. Incorporating the 2D box into the prompt improves the accuracy across all experiments compared to using only the caption as the prompt. This demonstrates the difficulty of 3D grounding when only the text caption is used as a prompt. When an image contains multiple objects of the same category, these objects may look identical and could be spatially overlapping. Distinguishing such objects using only language prompts is very challenging. When using the object caption and box combination as input prompts, the localization accuracy greatly increases.

Table 8 displays the experimental results when the input prompt includes both the caption and a 2D point. In real-world scenarios involving robots, embodied intelligence, and augmented reality (AR), users interact through various forms of input. These diverse prompts correspond to the multiple input modalities required in such environments. Our method demonstrates high performance across these varied prompts, surpassing all state-of-the-art methods,

Table 6 The results on Mono3DRefer Datasets.

Method	Type	Unique		Multiple		Overall	
		Acc@0.25↑	Acc@0.5↑	Acc@0.25↑	Acc@0.5↑	Acc@0.25↑	Acc@0.5↑
ZSGNet + backproj	One-Stage	9.02	0.29	16.56	2.23	15.14	1.87
FAOA + backproj	One-Stage	11.96	2.06	13.79	2.12	13.44	2.11
ReSC + backproj	One-Stage	11.96	0.49	23.69	3.94	21.48	3.29
TransVG + backproj	Tran.-based	15.78	4.02	21.84	4.16	20.70	4.14
Mono3DVG-TR	Tran.-based	57.65	33.04	65.92	46.85	64.36	44.25
LLMI3D	Tran.-based	**60.14**	**35.91**	**69.19**	**49.20**	**67.48**	**46.70**

Table 7 When the input prompt is changed to caption+2D Box, comparison of our LLMI3D with other methods on the IG3D-SUNRGBD, IG3D-nuScenes, IG3D-KITTI, and IG3D-Objectron datasets.

dataset	Method	Acc@0.25↑	Acc@0.5↑	DepthError↓	LengthError↓	WidthError↓	HeightError↓
IG3D-SUNRGBD	Text3D	21.3	4.7	0.39	0.17	0.28	0.16
	Mono3DVG	47.8	18.1	0.35	0.13	0.20	0.12
	LLMI3D	**56.6**	**18.6**	**0.25**	**0.11**	**0.18**	**0.11**
IG3D-nuScenes	Text3D	22.4	8.9	2.16	1.19	0.26	0.19
	Mono3DVG	34.1	11.5	2.31	0.46	0.17	0.15
	LLMI3D	**42.4**	**17.6**	**1.45**	**0.43**	**0.15**	**0.14**
IG3D-KITTI	Text3D	10.2	2.8	2.75	0.77	0.17	0.17
	Mono3DVG	43.7	13.4	0.88	0.38	0.12	**0.10**
	LLMI3D	**50.1**	**17.5**	**0.77**	**0.34**	**0.11**	**0.10**
IG3D-Objectron	Text3D	36.7	11.5	0.08	**0.03**	**0.02**	0.04
	Mono3DVG	53.4	16.4	0.05	**0.03**	**0.02**	**0.03**
	LLMI3D	**64.4**	**21.9**	**0.04**	**0.03**	**0.02**	**0.03**

showcasing the robust and general capabilities of our approach.

4.3 Ablation Study

4.3.1 Ablation Study on the Spatial-Enhanced Local Feature Mining

Table 9 presents a detailed ablation study on the Spatial-Enhanced Local Feature Mining method in the image encoder. The "CNN HR branch" denotes the use of an additional ConvNeXt (Liu et al., 2022) CNN branch to extract high-resolution image features. If the CNN branch is not selected, only the ViT branch is used for image feature extraction. The "depth branch" refers to the addition of an extra spatial depth branch within the ConvNeXt in the image encoder. This branch is utilized for predicting the object-level depth map. "SECBA", an abbreviation for Spatial-Enhanced Cross-Branch Attention, refers to leveraging spatial-enhanced local features extracted from the CNN and the global feature tokens generated by the ViT to perform spatial-enhanced cross-branch attention.

In Table 9, Experiment (a) solely utilizes image tokens extracted by CLIP ViT, which are then input into a large language model. Due to the low resolution of images processed by CLIP ViT, it primarily extracts semantic information and lacks capabilities in local spatial feature extraction, leading to weak spatial and local feature extraction issues. Consequently, Experiment (a) exhibits a relatively high DepthError and low accuracy.

In Experiment (b), an additional CNN branch is incorporated. High-resolution images are fed into the CNN HR branch, allowing distant and small objects to have sufficient pixels processed by the neural network. Compared to Experiment (a), the accuracy in Experiment (b) improves significantly due to better perception of distant and small objects. Notably, in Experiments (b) and (c), max pooling is applied to the feature maps extracted by the CNN, which are then summed with the tokens extracted by the ViT before being input into the large language model.

Experiment (c), compared to Experiment (b), introduces the depth branch within the CNN to predict object-level depth maps and uses object depth for supervision, enhancing the CNN branch's ability to extract spatial features. Experiment (d) represents the final, complete version of our proposed method. It utilizes the CNN, enhanced with the depth branch, to extract Spatial-Enhanced Local Features and applies spatial-enhanced cross-branch

Table 8 When the input prompt is changed to caption+2D Point, comparison of our LLMI3D with other methods on the IG3D-SUNRGBD, IG3D-nuScenes, IG3D-KITTI, and IG3D-Objectron datasets.

dataset	Method	Acc@0.25↑	Acc@0.5↑	DepthError↓	LengthError↓	WidthError↓	HeightError↓
IG3D-SUNRGBD	Text3D	15.7	2.5	0.40	0.19	0.33	0.18
	Mono3DVG	42.6	11.4	0.37	0.13	0.23	0.13
	LLMI3D	**53.7**	**15.8**	**0.25**	**0.12**	**0.20**	**0.12**
IG3D-nuScenes	Text3D	19.7	7.6	2.20	1.23	0.29	0.25
	Mono3DVG	30.1	10.4	2.41	0.53	0.18	0.18
	LLMI3D	**37.4**	**17.4**	**1.74**	**0.46**	**0.17**	**0.16**
IG3D-KITTI	Text3D	8.3	2.1	3.19	0.78	0.16	0.17
	Mono3DVG	31.4	9.5	1.72	0.38	0.13	0.12
	LLMI3D	**39.0**	**14.3**	**1.05**	**0.34**	**0.11**	**0.10**
IG3D-Objectron	Text3D	37.9	12.6	0.07	**0.03**	**0.02**	0.04
	Mono3DVG	49.6	15.1	0.06	**0.03**	**0.02**	**0.03**
	LLMI3D	**62.8**	**22.9**	**0.05**	**0.03**	**0.02**	**0.03**

Fig. 7 The 3D grounding visualizations of our LLMI3D in the IG3D-KITTI dataset. When users input an image and a caption describing the object of interest, our LLMI3D can generate the corresponding 3D bounding box of the object.

attention to integrate these with the tokens from the ViT. Experiments (b) and (c) naively use max pooling to sum the feature maps extracted by the CNN with the ViT tokens before inputting them into the large language model. Although this naive approach shows some effectiveness, it tends to blend the entire image feature from CNN with the ViT tokens, resulting in poor local feature matching. And it fails to focus on the critical local object regions within the images. In contrast, our proposed spatial-enhanced cross-branch attention enables the ViT tokens to attend to the areas of interest in the image more effectively, ensuring a better match between the ViT tokens and the local features of the image.

4.3.2 Ablation Study on the 3D Query Token-Derived Info Decoding

As shown in Table 10, we conduct an ablation study on the 3D Query Token-Derived Info Decoding method using the IG3D-SUNRGBD dataset. In Experiment (a), we directly use text token outputs for 3D grounding results. As previously discussed, directly using text tokens results in poor accuracy, slow speed, and difficulty in parsing. Consequently, the accuracy of text token outputs is quite low.

In Experiment (b), we extract the last hidden layer of the text token and use these text features as inputs to MLP to regress the 3D geometric values. Compared to Experiment (a), the accuracy in Experiment (b) improves significantly. This demonstrates that using regression methods to compute

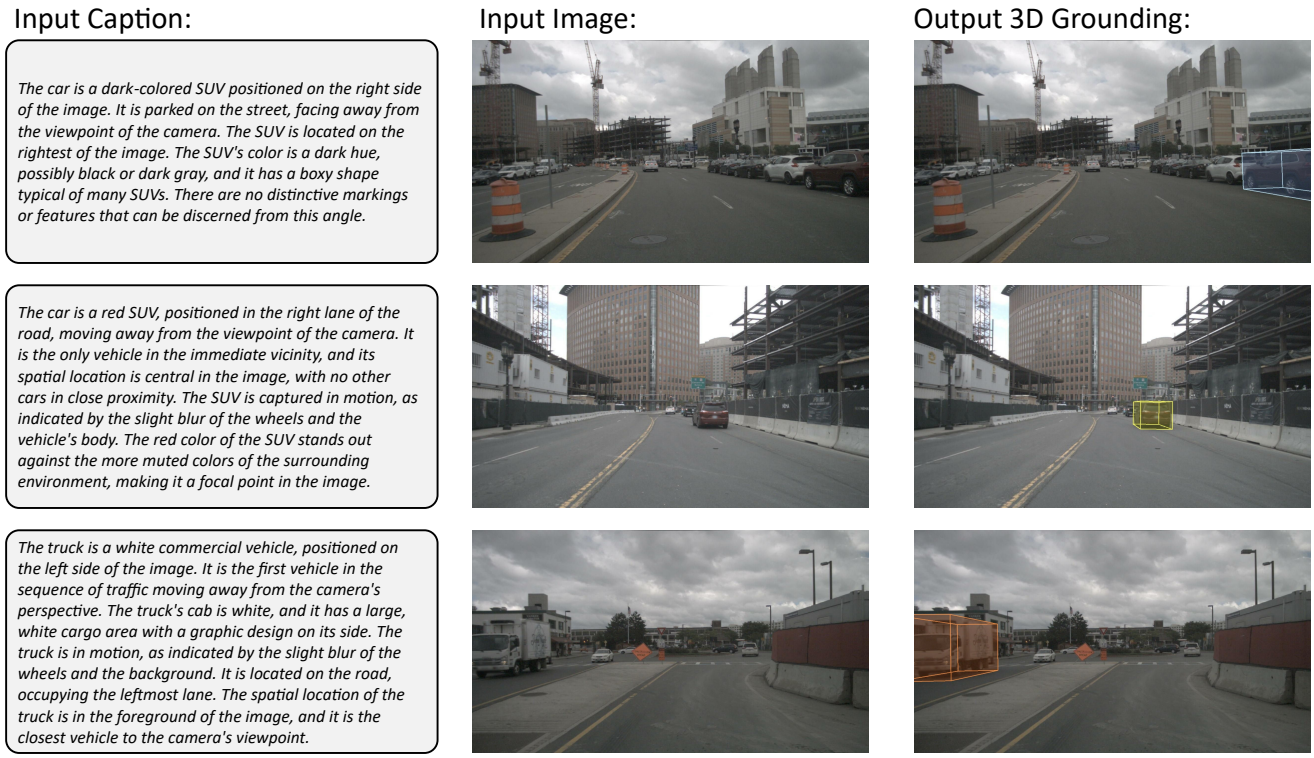

Fig. 8 The 3D grounding visualizations of our LLMI3D in the IG3D-nuScenes dataset. When users input an image and a caption describing the object of interest, our LLMI3D can generate the corresponding 3D bounding box of the object.

Table 9 Ablation study on the Spatial-Enhanced Local Feature Mining method using the IG3D-SUNRGBD dataset. "SECBA" stands for Spatial-Enhanced Cross-Branch Attention.

Exp	CNN HR branch	depth branch	SECBA	Acc@0.25↑	Acc@0.5↑	DepthError↓
(a)				30.7	6.1	0.73
(b)	✓			35.4	8.6	0.58
(c)	✓	✓		37.4	9.3	0.42
(d)	✓	✓	✓	**42.3**	**11.8**	**0.32**

Table 10 Ablation Study on the IG3D-SUNRGBD dataset to evaluate the decoding methods in the LLM part.

Exp	setting	Acc@0.25↑	Acc@0.5↑	DepthError↓
(a)	text output	11.5	1.7	0.45
(b)	text feature	21.6	4.7	0.42
(c)	position token	40.5	10.4	0.37
(d)	3D decoder token with 3D query	**42.3**	**11.8**	**0.32**

geometric values is considerably more effective than text-based outputs, as large models are not particularly adept at handling numbers. Utilizing MLP regression helps compensate for the large model's deficiencies in numerical output.

In Experiment (c), instead of using the text token, we use the hidden layer features of a special <pos> token to MLP for regression. While using the <pos> token's features is reasonably effective, its accuracy is not as high as our 3D query method.

Finally, in Experiment (d), we present our complete 3D query token-derived info decoding method. In comparison to Experiment (c), we introduce a learnable token, the 3D query, which serves as the Query in self-attention in the LLM. This allows effective and adaptive extraction of image and text features within the self-attention mechanism.

Hence, we can achieve precise 3D feature extraction using just a single token.

4.3.3 Ablation Study on the Geometry Projection-Based 3D Reasoning

As shown in Table 11, we present the results of our ablation study on Geometry Projection-Based 3D Reasoning using the IG3D-nuScenes dataset. The nuScenes dataset consists of images captured by six different cameras, with varying focal lengths. The "back projection" indicates whether the back projection is performed using the predicted image projection of the 3D center and depth. The "2D-3D height depth" refers to the depth value obtained using the height of the object's 2D and 3D bounding box, through Equation 26: $Z_2 = \frac{H}{h} \cdot f_y$. "Virtual depth" is the depth derived using Equation 23: $Z_1 = d_v \cdot \frac{f_x}{f_x^v} \cdot \frac{w^v}{w}$.

In Table 11, Experiment (a) does not perform back projection. Consequently, it does not predict the image projection or the depth value of the object's center. Instead, Experiment (a) directly regresses the spatial coordinates x, y, and z of the object. This approach has significant issues: x, y, and z are absolute coordinates in space, making them difficult to regress from the image. If any of the x, y, or z values have a prediction error, it results in substantial errors in the 3D box's position. Therefore, the accuracy of Experiment (a) is considerably lower compared to other experiments.

Experiment (b) uses the 3D center and depth, along with back projection, to obtain the spatial position of the 3D bounding box. Comparatively, Experiment (b) significantly improves accuracy over Experiment (a). Experiment (c) builds upon Experiment (b) by merging 2D-3D height depth. Using the equation $Z_2 = \frac{H}{h} \cdot f_y$ to obtain depth values in outdoor scenes provides a certain reference value, resulting in improved accuracy over Experiment (b).

Experiment (d) further adds virtual depth. Virtual depth is assumed to be generated under a uniform virtual camera. When dealing with images of different focal lengths, regressing a unified virtual depth through the neural network shows clear advantages. The nuScenes dataset comprises images captured by six cameras with different focal lengths. Thus, utilizing virtual depth overcomes the focal length differences, enhancing 3D grounding performance on the nuScenes dataset.

Experiment (e) represents our complete version of LLMI3D, which integrates 2D-3D height and virtual depth to obtain the final depth, as shown in Equation 27. Additionally, it leverages the predicted image projection of the object's 3D center to derive the 3D bounding box. This approach overcomes the issue of MLLM's inability to handle camera focal variations, resulting in the best experimental accuracy observed.

4.3.4 Experiments on the Rotation Angle Prediction

Table 12 presents our experiments on different methods for predicting rotation angles. Experiments (a) and (b), respectively, use Euler Angles and 6D allocentric rotation for predicting the rotations. As discussed in Section 3.3, Euler Angles suffer from several issues including discontinuity, non-uniqueness of representation, complexity of the loss function, and asymmetry, making them unfriendly for neural networks. Consequently, Experiment (a), which directly predicts Euler Angles, exhibits significant errors. In contrast, Experiment (b) utilizes 6D allocentric rotation, which is continuous in 6D space and better suited for learning by neural networks.

4.4 Visualization Results

We present the visualizations of our LLMI3D model's performance on the 3D Grounding and 3D VQA tasks.

Figure 5 showcases the question-answering results of our model on the IG3D-SUNRGBD-VQA dataset. There is a significant demand for embodied intelligence and robotic scenarios to provide answers to users' questions and determine the spatial locations of relevant objects. In real-world scenarios, users' questions are highly diverse, and answering these questions requires the model to possess common sense knowledge and logical reasoning capabilities. These are challenging capacities that specialized small models lack completely. Currently, only large-scale pre-trained language models possess the necessary common sense and logical reasoning abilities. As demonstrated in Figure 5, our model can comprehend user questions, perform reasoning, provide accurate responses, and indicate the object's location in 3D space, showcasing its powerful capabilities and strong application potential in robotics and other scenarios.

Figure 6 demonstrates our model's 3D Grounding performance on the IG3D-SUNRGBD dataset. The IG3D-SUNRGBD dataset comprises indoor scenes with a variety of object categories. When the user provides an image and a caption of the object of interest, our LLMI3D can return the corresponding 3D Bounding Box. Moreover, our LLMI3D is capable of handling various input prompts, such as a caption plus a 2D Box or a caption plus a 2D point, and subsequently outputs the object's 3D Bounding Box. This flexibility meets diverse user input requirements. As observed, our model exhibits strong spatial localization capabilities and precise rotation predictions for objects.

Figures 7 and 8 illustrate our model's 3D Grounding performance on the IG3D-KITTI and IG3D-nuScenes datasets, respectively. Both IG3D-KITTI and IG3D-nuScenes are outdoor datasets aimed at autonomous driving scenarios, encompassing common object categories found in street

Table 11 Ablation study on the IG3D-nuScenes dataset to evaluate the Geometry Projection-Based 3D Reasoning method.

Exp	back projection	2D-3D height depth	virtual depth	Acc@0.25↑	Acc@0.5↑	DepthError↓
(a)				19.8	8.6	3.46
(b)	✓			26.9	10.4	2.94
(c)	✓	✓		28.8	12.1	2.73
(d)	✓		✓	29.4	12.5	2.57
(e)	✓	✓	✓	**31.6**	**13.2**	**2.19**

Table 12 Experiments on the IG3D-SUNRGBD dataset to evaluate the rotation angle prediction methods.

Exp	rotation prediction	Acc@0.25↑	Acc@0.5↑
(a)	Euler Angle	37.1	8.8
(b)	6D allocentric rotaion	**42.3**	**11.8**

scenes. Outdoor scenes cover larger spatial ranges, demanding higher spatial perception capabilities. In Figures 7 and 8, when the user provides an image and a caption of the object of interest, our LLMI3D returns the corresponding object's 3D Bounding Box, and our model accurately locates the target object's spatial position.

5 Conclusion

The demand for 3D perception in real-world applications such as autonomous driving, augmented reality, robotics, and embodied intelligence is growing rapidly. In this paper, we identified three major challenges faced by multimodal large language models in 3D perception: (1) Weak spatial and local object perception, (2) Poor text-based geometric numerical output, and (3) Inability to handle camera focal variations. These issues significantly limit the application of multimodal large language models in 3D perception scenarios.

To address these challenges, we proposed LLMI3D, a 3D-friendly multimodal large language model architecture. For the image encoder, we introduced Spatial-Enhanced Local Feature Mining. In the LLM part, we proposed 3D Query Token-Derived Info Decoding. To obtain the 3D bounding boxes of objects, we proposed Geometry Projection-Based 3D Reasoning. Additionally, we introduced the IG3D dataset, which enables the evaluation of fine-grained grounding, logical reasoning, and question-answering capabilities of 3D perception models.

We conducted extensive experiments on various 3D perception datasets, including 3D grounding and 3D VQA experiments, open-vocabulary 3D grounding experiments, and domain generalization experiments. The experimental results demonstrate that our approach achieves state-of-the-art performance, significantly outperforming existing methods. Detailed ablation studies validate the rationality of our motivations and the effectiveness of each module in our method.

References

Addari G, Guillemaut J (2023) A family of approaches for full 3d reconstruction of objects with complex surface reflectance. International Journal of Computer Vision (IJCV) 131(9):2243–2266, DOI 10.1007/S11263-023-01795-W, URL https://doi.org/10.1007/s11263-023-01795-w

Ahmadyan A, Zhang L, Ablavatski A, Wei J, Grundmann M (2021) Objectron: A large scale dataset of object-centric videos in the wild with pose annotations. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021, Computer Vision Foundation / IEEE, pp 7822–7831, DOI 10.1109/CVPR46437.2021.00773, URL https://openaccess.thecvf.com/content/CVPR2021/html/Ahmadyan_Objectron_A_Large_Scale_Dataset_of_Object-Centric_Videos_in_the_CVPR_2021_paper.html

Angelova A, Carneiro G, Sünderhauf N, Leitner J (2020) Special issue on deep learning for robotic vision. International Journal of Computer Vision (IJCV) 128(5):1160–1161, DOI 10.1007/S11263-020-01324-Z, URL https://doi.org/10.1007/s11263-020-01324-z

Anil R, Borgeaud S, Wu Y, Alayrac J, Yu J, Soricut R, Schalkwyk J, Dai AM, Hauth A, Millican K, Silver D, Petrov S, Johnson M, Antonoglou I, Schrittwieser J, Glaese A, Chen J, Pitler E, Lillicrap TP, Lazaridou A, Firat O, Molloy J, Isard M, Barham PR, Hennigan T, Lee B, Viola F, Reynolds M, Xu Y, Doherty R, Collins E, et al CM (2023) Gemini: A family of highly capable multimodal models. CoRR abs/2312.11805, DOI 10.48550/ARXIV.2312.11805, URL https://doi.org/10.48550/arXiv.2312.11805, 2312.11805

Aumentado-Armstrong T, Tsogkas S, Dickinson SJ, Jepson AD (2023) Disentangling geometric deformation spaces in generative latent shape models. International Journal of Computer Vision (IJCV) 131(7):1611–1641, DOI

10.1007/S11263-023-01750-9, URL https://doi.org/10.1007/s11263-023-01750-9

Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In: Bengio Y, LeCun Y (eds) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, URL http://arxiv.org/abs/1409.0473

Bai J, Bai S, Yang S, Wang S, Tan S, Wang P, Lin J, Zhou C, Zhou J (2023) Qwen-vl: A frontier large vision-language model with versatile abilities. CoRR abs/2308.12966, DOI 10.48550/ARXIV.2308.12966, URL https://doi.org/10.48550/arXiv.2308.12966, 2308.12966

Brazil G, Kumar A, Straub J, Ravi N, Johnson J, Gkioxari G (2023) Omni3d: A large benchmark and model for 3d object detection in the wild. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023, IEEE, pp 13154–13164, DOI 10.1109/CVPR52729.2023.01264, URL https://doi.org/10.1109/CVPR52729.2023.01264

Caesar H, Bankiti V, Lang AH, Vora S, Liong VE, Xu Q, Krishnan A, Pan Y, Baldan G, Beijbom O (2020) nuscenes: A multimodal dataset for autonomous driving. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, Computer Vision Foundation / IEEE, pp 11618–11628, DOI 10.1109/CVPR42600.2020.01164, URL https://openaccess.thecvf.com/content_CVPR_2020/html/Caesar_nuScenes_A_Multimodal_Dataset_for_Autonomous_Driving_CVPR_2020_paper.html

Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S (2020) End-to-end object detection with transformers. In: Vedaldi A, Bischof H, Brox T, Frahm J (eds) Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I, Springer, Lecture Notes in Computer Science, vol 12346, pp 213–229, DOI 10.1007/978-3-030-58452-8_13, URL https://doi.org/10.1007/978-3-030-58452-8_13

Chen J, Zhu D, Shen X, Li X, Liu Z, Zhang P, Krishnamoorthi R, Chandra V, Xiong Y, Elhoseiny M (2023a) Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. CoRR abs/2310.09478, DOI 10.48550/ARXIV.2310.09478, URL https://doi.org/10.48550/arXiv.2310.09478, 2310.09478

Chen K, Gal E, Yan H, Li H (2024) Domain generalization with small data. International Journal of Computer Vision (IJCV) 132(8):3172–3190, DOI 10.1007/S11263-024-02028-4, URL https://doi.org/10.1007/s11263-024-02028-4

Chen L, Li J, Dong X, Zhang P, He C, Wang J, Zhao F, Lin D (2023b) Sharegpt4v: Improving large multi-modal models with better captions. CoRR abs/2311.12793, DOI 10.48550/ARXIV.2311.12793, URL https://doi.org/10.48550/arXiv.2311.12793, 2311.12793

Chen X, Kundu K, Zhu Y, Berneshawi AG, Ma H, Fidler S, Urtasun R (2015) 3d object proposals for accurate object class detection. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R (eds) Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada, pp 424–432, URL https://proceedings.neurips.cc/paper/2015/hash/6da37dd3139aa4d9aa55b8d237ec5d4a.html

Chen Z, Wu J, Wang W, Su W, Chen G, Xing S, Zhong M, Zhang Q, Zhu X, Lu L, Li B, Luo P, Lu T, Qiao Y, Dai J (2023c) Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. CoRR abs/2312.14238, DOI 10.48550/ARXIV.2312.14238, URL https://doi.org/10.48550/arXiv.2312.14238, 2312.14238

Chen Z, Zhang J, Xu Y, Tao D (2023d) Transformer-based context condensation for boosting feature pyramids in object detection. International Journal of Computer Vision (IJCV) 131(10):2738–2756, DOI 10.1007/S11263-023-01830-W, URL https://doi.org/10.1007/s11263-023-01830-w

Chiang WL, Li Z, Lin Z, Sheng Y, Wu Z, Zhang H, Zheng L, Zhuang S, Zhuang Y, Gonzalez JE, et al. (2023) Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See https://vicuna lmsys org (accessed 14 April 2023) 2(3):6

Cho JH, Ivanovic B, Cao Y, Schmerling E, Wang Y, Weng X, Li B, You Y, Krähenbühl P, Wang Y, Pavone M (2024a) Language-image models with 3d understanding. CoRR abs/2405.03685, DOI 10.48550/ARXIV.2405.03685, URL https://doi.org/10.48550/arXiv.2405.03685, 2405.03685

Cho JH, Ivanovic B, Cao Y, Schmerling E, Wang Y, Weng X, Li B, You Y, Krähenbühl P, Wang Y, Pavone M (2024b) Language-image models with 3d understanding. CoRR abs/2405.03685, DOI 10.48550/ARXIV.2405.03685, URL https://doi.org/10.48550/arXiv.2405.03685, 2405.03685

Deng J, Yang Z, Chen T, Zhou W, Li H (2021) Transvg: End-to-end visual grounding with transformers. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October

10-17, 2021, IEEE, pp 1749–1759, DOI 10.1109/ICCV48922.2021.00179, URL https://doi.org/10.1109/ICCV48922.2021.00179

Devlin J, Chang M, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein J, Doran C, Solorio T (eds) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, pp 4171–4186, DOI 10.18653/V1/N19-1423, URL https://doi.org/10.18653/v1/n19-1423

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N (2021) An image is worth 16x16 words: Transformers for image recognition at scale. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, OpenReview.net, URL https://openreview.net/forum?id=YicbFdNTTy

Ettinger S, Cheng S, Caine B, Liu C, Zhao H, Pradhan S, Chai Y, Sapp B, Qi CR, Zhou Y, Yang Z, Chouard A, Sun P, Ngiam J, Vasudevan V, McCauley A, Shlens J, Anguelov D (2021) Large scale interactive motion forecasting for autonomous driving : The waymo open motion dataset. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, IEEE, pp 9690–9699, DOI 10.1109/ICCV48922.2021.00957, URL https://doi.org/10.1109/ICCV48922.2021.00957

Feng J, Yang Y, Xie Y, Li Y, Guo Y, Guo Y, He Y, Xiang L, Ding G (2024) Debiased novel category discovering and localization. In: Wooldridge MJ, Dy JG, Natarajan S (eds) Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, AAAI Press, pp 1753–1760, DOI 10.1609/AAAI.V38I2.27943, URL https://doi.org/10.1609/aaai.v38i2.27943

Gao P, Han J, Zhang R, Lin Z, Geng S, Zhou A, Zhang W, Lu P, He C, Yue X, Li H, Qiao Y (2023) Llama-adapter V2: parameter-efficient visual instruction model. CoRR abs/2304.15010, DOI 10.48550/ARXIV.2304.15010, URL https://doi.org/10.48550/arXiv.2304.15010, 2304.15010

Geiger A, Lenz P, Urtasun R (2012) Are we ready for autonomous driving? the KITTI vision benchmark suite. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012, IEEE Computer Society, pp 3354–3361, DOI 10.1109/CVPR.2012.6248074, URL https://doi.org/10.1109/CVPR.2012.6248074

Hao S, Liu P, Zhan Y, Jin K, Liu Z, Song M, Hwang J, Wang G (2024a) Divotrack: A novel dataset and baseline method for cross-view multi-object tracking in diverse open scenes. International Journal of Computer Vision (IJCV) 132(4):1075–1090, DOI 10.1007/S11263-023-01922-7, URL https://doi.org/10.1007/s11263-023-01922-7

Hao T, Ding X, Feng J, Yang Y, Chen H, Ding G (2024b) Quantized prompt for efficient generalization of vision-language models. arXiv preprint arXiv:240710704

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, IEEE Computer Society, pp 770–778, DOI 10.1109/CVPR.2016.90, URL https://doi.org/10.1109/CVPR.2016.90

Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L, Chen W (2022) Lora: Low-rank adaptation of large language models. In: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022, OpenReview.net, URL https://openreview.net/forum?id=nZeVKeeFYf9

Huang K, Wu T, Su H, Hsu WH (2022) Monodtr: Monocular 3d object detection with depth-aware transformer. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, IEEE, pp 4002–4011, DOI 10.1109/CVPR52688.2022.00398, URL https://doi.org/10.1109/CVPR52688.2022.00398

Lang AH, Vora S, Caesar H, Zhou L, Yang J, Beijbom O (2019) Pointpillars: Fast encoders for object detection from point clouds. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, pp 12697–12705, DOI 10.1109/CVPR.2019.01298, URL http://openaccess.thecvf.com/content_CVPR_2019/html/Lang_PointPillars_Fast_Encoders_for_Object_Detection_From_Point_Clouds_CVPR_2019_paper.html

Lehner A, Gasperini S, Marcos-Ramiro A, Schmidt M, Navab N, Busam B, Tombari F (2024) 3d adversarial augmentations for robust out-of-domain predictions. International Journal of Computer Vision (IJCV) 132(3):931–963, DOI 10.1007/S11263-023-01914-7, URL https://doi.org/10.1007/s11263-023-01914-7

Li J, Li D, Savarese S, Hoi SCH (2023) BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: Krause A, Brunskill E, Cho K, Engelhardt B, Sabato S, Scarlett J (eds) International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA, PMLR, Proceedings of Machine Learning

Research, vol 202, pp 19730–19742, URL https://proceedings.mlr.press/v202/li23q.html

Li P, Zhao H, Liu P, Cao F (2020) RTM3D: real-time monocular 3d detection from object keypoints for autonomous driving. In: Vedaldi A, Bischof H, Brox T, Frahm J (eds) Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part III, Springer, Lecture Notes in Computer Science, vol 12348, pp 644–660, DOI 10.1007/978-3-030-58580-8_38, URL https://doi.org/10.1007/978-3-030-58580-8_38

Li Y, Zhang Y, Wang C, Zhong Z, Chen Y, Chu R, Liu S, Jia J (2024) Mini-gemini: Mining the potential of multi-modality vision language models. CoRR abs/2403.18814, DOI 10.48550/ARXIV.2403.18814, URL https://doi.org/10.48550/arXiv.2403.18814, 2403.18814

Li Z, Xi T, Zhang G, Liu J, He R (2021) Autodet: Pyramid network architecture search for object detection. International Journal of Computer Vision (IJCV) 129(4):1087–1105, DOI 10.1007/S11263-020-01415-X, URL https://doi.org/10.1007/s11263-020-01415-x

Liu H, Li C, Wu Q, Lee YJ (2023a) Visual instruction tuning. In: Oh A, Naumann T, Globerson A, Saenko K, Hardt M, Levine S (eds) Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, URL http://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html

Liu S, Zeng Z, Ren T, Li F, Zhang H, Yang J, Li C, Yang J, Su H, Zhu J, Zhang L (2023b) Grounding DINO: marrying DINO with grounded pre-training for open-set object detection. CoRR abs/2303.05499, DOI 10.48550/ARXIV.2303.05499, URL https://doi.org/10.48550/arXiv.2303.05499, 2303.05499

Liu Z, Mao H, Wu C, Feichtenhofer C, Darrell T, Xie S (2022) A convnet for the 2020s. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, IEEE, pp 11966–11976, DOI 10.1109/CVPR52688.2022.01167, URL https://doi.org/10.1109/CVPR52688.2022.01167

Lu H, Liu W, Zhang B, Wang B, Dong K, Liu B, Sun J, Ren T, Li Z, Yang H, Sun Y, Deng C, Xu H, Xie Z, Ruan C (2024) Deepseek-vl: Towards real-world vision-language understanding. CoRR abs/2403.05525, DOI 10.48550/ARXIV.2403.05525, URL https://doi.org/10.48550/arXiv.2403.05525, 2403.05525

Lu J, Batra D, Parikh D, Lee S (2019) Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Wallach HM, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox EB, Garnett R (eds) Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, pp 13–23, URL https://proceedings.neurips.cc/paper/2019/hash/c74d97b01eae257e44aa9d5bade97baf-Abstract.html

Lu Y, Ma X, Yang L, Zhang T, Liu Y, Chu Q, Yan J, Ouyang W (2021) Geometry uncertainty projection network for monocular 3d object detection. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, IEEE, pp 3091–3101, DOI 10.1109/ICCV48922.2021.00310, URL https://doi.org/10.1109/ICCV48922.2021.00310

Lyu M, Zhou J, Chen H, Huang Y, Yu D, Li Y, Guo Y, Guo Y, Xiang L, Ding G (2023) Box-level active detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023, IEEE, pp 23766–23775, DOI 10.1109/CVPR52729.2023.02276, URL https://doi.org/10.1109/CVPR52729.2023.02276

Mao J, Shi S, Wang X, Li H (2023) 3d object detection for autonomous driving: A comprehensive survey. International Journal of Computer Vision (IJCV) 131(8):1909–1963, DOI 10.1007/S11263-023-01790-1, URL https://doi.org/10.1007/s11263-023-01790-1

Mousavian A, Anguelov D, Flynn J, Kosecka J (2017) 3d bounding box estimation using deep learning and geometry. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, IEEE Computer Society, pp 5632–5640, DOI 10.1109/CVPR.2017.597, URL https://doi.org/10.1109/CVPR.2017.597

OpenAI (2023) GPT-4 technical report. CoRR abs/2303.08774, DOI 10.48550/ARXIV.2303.08774, URL https://doi.org/10.48550/arXiv.2303.08774, 2303.08774

Peng Z, Wang W, Dong L, Hao Y, Huang S, Ma S, Wei F (2023) Kosmos-2: Grounding multimodal large language models to the world. CoRR abs/2306.14824, DOI 10.48550/ARXIV.2306.14824, URL https://doi.org/10.48550/arXiv.2306.14824, 2306.14824

Pi R, Yao L, Gao J, Zhang J, Zhang T (2023) Perceptiongpt: Effectively fusing visual perception into LLM. CoRR abs/2311.06612, DOI 10.48550/ARXIV.2311.06612, URL https://doi.org/10.48550/

arXiv.2311.06612, 2311.06612

Qin Z, Wang J, Lu Y (2019) Monogrnet: A geometric reasoning network for monocular 3d object localization. In: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, AAAI Press, pp 8851–8858, DOI 10.1609/AAAI.V33I01.33018851, URL https://doi.org/10.1609/aaai.v33i01.33018851

Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I (2021) Learning transferable visual models from natural language supervision. In: Meila M, Zhang T (eds) Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, PMLR, Proceedings of Machine Learning Research, vol 139, pp 8748–8763, URL http://proceedings.mlr.press/v139/radford21a.html

Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, Chen M, Sutskever I (2021) Zero-shot text-to-image generation. In: Meila M, Zhang T (eds) Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, PMLR, Proceedings of Machine Learning Research, vol 139, pp 8821–8831, URL http://proceedings.mlr.press/v139/ramesh21a.html

Rodriguez AL, Mikolajczyk K (2023) DESC: domain adaptation for depth estimation via semantic consistency. International Journal of Computer Vision (IJCV) 131(3):752–771, DOI 10.1007/S11263-022-01718-1, URL https://doi.org/10.1007/s11263-022-01718-1

Rodriguez AL, Busam B, Mikolajczyk K (2023) Project to adapt: Domain adaptation for depth completion from noisy and sparse sensor data. International Journal of Computer Vision (IJCV) 131(3):796–812, DOI 10.1007/S11263-022-01726-1, URL https://doi.org/10.1007/s11263-022-01726-1

Sadhu A, Chen K, Nevatia R (2019) Zero-shot grounding of objects from natural language queries. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, IEEE, pp 4693–4702, DOI 10.1109/ICCV.2019.00479, URL https://doi.org/10.1109/ICCV.2019.00479

Shen L, He T, Guo Y, Ding G (2023) X-reid: Cross-instance transformer for identity-level person re-identification. CoRR abs/2302.02075, DOI 10.48550/ARXIV.2302.02075, URL https://doi.org/10.48550/arXiv.2302.02075, 2302.02075

Shi S, Jiang L, Deng J, Wang Z, Guo C, Shi J, Wang X, Li H (2023) PV-RCNN++: point-voxel feature set abstraction with local vector representation for 3d object detection. International Journal of Computer Vision (IJCV) 131(2):531–551, DOI 10.1007/S11263-022-01710-9, URL https://doi.org/10.1007/s11263-022-01710-9

Shi X, Ye Q, Chen X, Chen C, Chen Z, Kim T (2021) Geometry-based distance decomposition for monocular 3d object detection. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, IEEE, pp 15152–15161, DOI 10.1109/ICCV48922.2021.01489, URL https://doi.org/10.1109/ICCV48922.2021.01489

Singh A, Hu R, Goswami V, Couairon G, Galuba W, Rohrbach M, Kiela D (2022) FLAVA: A foundational language and vision alignment model. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, IEEE, pp 15617–15629, DOI 10.1109/CVPR52688.2022.01519, URL https://doi.org/10.1109/CVPR52688.2022.01519

Song S, Lichtenberg SP, Xiao J (2015) SUN RGB-D: A RGB-D scene understanding benchmark suite. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, IEEE Computer Society, pp 567–576, DOI 10.1109/CVPR.2015.7298655, URL https://doi.org/10.1109/CVPR.2015.7298655

Stoiber M, Pfanne M, Strobl KH, Triebel R, Albu-Schäffer A (2022) SRT3D: A sparse region-based 3d object tracking approach for the real world. International Journal of Computer Vision (IJCV) 130(4):1008–1030, DOI 10.1007/S11263-022-01579-8, URL https://doi.org/10.1007/s11263-022-01579-8

Sundermeyer M, Marton Z, Durner M, Triebel R (2020) Augmented autoencoders: Implicit 3d orientation learning for 6d object detection. International Journal of Computer Vision (IJCV) 128(3):714–729, DOI 10.1007/S11263-019-01243-8, URL https://doi.org/10.1007/s11263-019-01243-8

Taori R, Gulrajani I, Zhang T, Dubois Y, Li X, Guestrin C, Liang P, Hashimoto TB (2023) Alpaca: A strong, replicable instruction-following model. Stanford Center for Research on Foundation Models https://crfm stanford edu/2023/03/13/alpaca html 3(6):7

Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F, Rodriguez A, Joulin A, Grave E, Lample G (2023) Llama: Open and efficient foundation language

models. CoRR abs/2302.13971, DOI 10.48550/ARXIV.2302.13971, URL https://doi.org/10.48550/arXiv.2302.13971, 2302.13971

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: Guyon I, von Luxburg U, Bengio S, Wallach HM, Fergus R, Vishwanathan SVN, Garnett R (eds) Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pp 5998–6008, URL https://proceedings.neurips.cc/paper/2017/hash

Vinyals O, Toshev A, Bengio S, Erhan D (2015) Show and tell: A neural image caption generator. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, IEEE Computer Society, pp 3156–3164, DOI 10.1109/CVPR.2015.7298935, URL https://doi.org/10.1109/CVPR.2015.7298935

Wang A, Chen H, Liu L, Chen K, Lin Z, Han J, Ding G (2024) Yolov10: Real-time end-to-end object detection. CoRR abs/2405.14458, DOI 10.48550/ARXIV.2405.14458, URL https://doi.org/10.48550/arXiv.2405.14458, 2405.14458

Wang P, Yang A, Men R, Lin J, Bai S, Li Z, Ma J, Zhou C, Zhou J, Yang H (2022) OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In: Chaudhuri K, Jegelka S, Song L, Szepesvári C, Niu G, Sabato S (eds) International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA, PMLR, Proceedings of Machine Learning Research, vol 162, pp 23318–23340, URL https://proceedings.mlr.press/v162/wang22al.html

Wang W, Chen Z, Chen X, Wu J, Zhu X, Zeng G, Luo P, Lu T, Zhou J, Qiao Y, Dai J (2023a) Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. In: Oh A, Naumann T, Globerson A, Saenko K, Hardt M, Levine S (eds) Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, URL http://papers.nips.cc/paper_files/paper/2023/hash

Wang W, Lv Q, Yu W, Hong W, Qi J, Wang Y, Ji J, Yang Z, Zhao L, Song X, Xu J, Xu B, Li J, Dong Y, Ding M, Tang J (2023b) Cogvlm: Visual expert for pretrained language models. CoRR abs/2311.03079, DOI 10.48550/ARXIV.2311.03079, URL https://doi.org/10.48550/arXiv.2311.03079, 2311.03079

Wang Y, Mao Q, Zhu H, Deng J, Zhang Y, Ji J, Li H, Zhang Y (2023c) Multi-modal 3d object detection in autonomous driving: A survey. International Journal of Computer Vision (IJCV) 131(8):2122–2152, DOI 10.1007/S11263-023-01784-Z, URL https://doi.org/10.1007/s11263-023-01784-z

Weng T, Xiao J, Pan H, Jiang H (2024) Partcom: Part composition learning for 3d open-set recognition. International Journal of Computer Vision (IJCV) 132(4):1393–1416, DOI 10.1007/S11263-023-01947-Y, URL https://doi.org/10.1007/s11263-023-01947-y

Wu Z, Gan Y, Wang L, Chen G, Pu J (2023) Monopgc: Monocular 3d object detection with pixel geometry contexts. In: IEEE International Conference on Robotics and Automation, ICRA 2023, London, UK, May 29 - June 2, 2023, IEEE, pp 4842–4849, DOI 10.1109/ICRA48891.2023.10161442, URL https://doi.org/10.1109/ICRA48891.2023.10161442

Xie Q, Lai Y, Wu J, Wang Z, Zhang Y, Xu K, Wang J (2021) Vote-based 3d object detection with context modeling and SOB-3DNMS. International Journal of Computer Vision (IJCV) 129(6):1857–1874, DOI 10.1007/S11263-021-01456-W, URL https://doi.org/10.1007/s11263-021-01456-w

Xiong Y, Chen H, Hao T, Lin Z, Han J, Zhang Y, Wang G, Bao Y, Ding G (2024) PYRA: parallel yielding re-activation for training-inference efficient task adaptation. CoRR abs/2403.09192, DOI 10.48550/ARXIV.2403.09192, URL https://doi.org/10.48550/arXiv.2403.09192, 2403.09192

Xu K, Ba J, Kiros R, Cho K, Courville AC, Salakhutdinov R, Zemel RS, Bengio Y (2015) Show, attend and tell: Neural image caption generation with visual attention. In: Bach FR, Blei DM (eds) Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015, JMLR.org, JMLR Workshop and Conference Proceedings, vol 37, pp 2048–2057, URL http://proceedings.mlr.press/v37/xuc15.html

Xu X, Chen H, Lin Z, Han J, Gong L, Wang G, Bao Y, Ding G (2024) Tad: A plug-and-play task-aware decoding method to better adapt llms on downstream tasks. In: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI

Yan L, Yan P, Xiong S, Xiang X, Tan Y (2024) Monocd: Monocular 3d object detection with complementary depths. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 10248–10257

Yang F, Xu X, Chen H, Guo Y, Han J, Ni K, Ding G (2022) Ground plane matters: Picking up ground plane prior in monocular 3d object detection. CoRR abs/2211.01556, DOI 10.48550/ARXIV.2211.01556, URL https://doi.org/10.48550/

arXiv.2211.01556, 2211.01556

Yang F, Xu X, Chen H, Guo Y, He Y, Ni K, Ding G (2023) Gpro3d: Deriving 3d bbox from ground plane in monocular 3d object detection. Neurocomputing 562:126894, DOI 10.1016/J.NEUCOM.2023.126894, URL https://doi.org/10.1016/j.neucom.2023.126894

Yang F, Chen H, He Y, Zhao S, Zhang C, Ni K, Ding G (2024a) Geometry-guided domain generalization for monocular 3d object detection. In: Wooldridge MJ, Dy JG, Natarajan S (eds) Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada, AAAI Press, pp 6467–6476, DOI 10.1609/AAAI.V38I6.28467, URL https://doi.org/10.1609/aaai.v38i6.28467

Yang S, Li G, Yu Y (2019) Dynamic graph attention for referring expression comprehension. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, IEEE, pp 4643–4652, DOI 10.1109/ICCV.2019.00474, URL https://doi.org/10.1109/ICCV.2019.00474

Yang Z, Chen T, Wang L, Luo J (2020) Improving one-stage visual grounding by recursive sub-query construction. In: Vedaldi A, Bischof H, Brox T, Frahm J (eds) Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XIV, Springer, Lecture Notes in Computer Science, vol 12359, pp 387–404, DOI 10.1007/978-3-030-58568-6_23, URL https://doi.org/10.1007/978-3-030-58568-6_23

Yang Z, Yue J, Ghamisi P, Zhang S, Ma J, Fang L (2024b) Open set recognition in real world. International Journal of Computer Vision (IJCV) 132(8):3208–3231, DOI 10.1007/S11263-024-02015-9, URL https://doi.org/10.1007/s11263-024-02015-9

Yin H, Xu X, Lu S, Chen X, Xiong R, Shen S, Stachniss C, Wang Y (2024) A survey on global lidar localization: Challenges, advances and open problems. International Journal of Computer Vision (IJCV) 132(8):3139–3171, DOI 10.1007/S11263-024-02019-5, URL https://doi.org/10.1007/s11263-024-02019-5

Zhan Y, Yuan Y, Xiong Z (2024) Mono3dvg: 3d visual grounding in monocular images. In: Wooldridge MJ, Dy JG, Natarajan S (eds) Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada, AAAI Press, pp 6988–6996, DOI 10.1609/AAAI.V38I7.28525, URL https://doi.org/10.1609/aaai.v38i7.28525

Zhang A, Ji W, Chua T (2023a) Next-chat: An LMM for chat, detection and segmentation. CoRR abs/2311.04498, DOI 10.48550/ARXIV.2311.04498, URL https://doi.org/10.48550/arXiv.2311.04498, 2311.04498

Zhang Q, Hou J, Qian Y, Chan AB, Zhang J, He Y (2022) Reggeonet: Learning regular representations for large-scale 3d point clouds. International Journal of Computer Vision (IJCV) 130(12):3100–3122, DOI 10.1007/S11263-022-01682-W, URL https://doi.org/10.1007/s11263-022-01682-w

Zhang R, Han J, Zhou A, Hu X, Yan S, Lu P, Li H, Gao P, Qiao Y (2023b) Llama-adapter: Efficient fine-tuning of language models with zero-init attention. CoRR abs/2303.16199, DOI 10.48550/ARXIV.2303.16199, URL https://doi.org/10.48550/arXiv.2303.16199, 2303.16199

Zhang R, Qiu H, Wang T, Guo Z, Cui Z, Qiao Y, Li H, Gao P (2023c) Monodetr: Depth-guided transformer for monocular 3d object detection. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023, IEEE, pp 9121–9132, DOI 10.1109/ICCV51070.2023.00840, URL https://doi.org/10.1109/ICCV51070.2023.00840

Zhang Y, Lu J, Zhou J (2021) Objects are different: Flexible monocular 3d object detection. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021, Computer Vision Foundation / IEEE, pp 3289–3298, DOI 10.1109/CVPR46437.2021.00330, URL https://openaccess.thecvf.com/content/CVPR2021/html/Zhang_Objects_Are_Different_Flexible_Monocular_3D_Object_Detection_CVPR_2021_paper.html

Zhang Y, Hou J, Yuan Y (2024) A comprehensive study of the robustness for lidar-based 3d object detectors against adversarial attacks. International Journal of Computer Vision (IJCV) 132(5):1592–1624, DOI 10.1007/S11263-023-01934-3, URL https://doi.org/10.1007/s11263-023-01934-3

Zhao L, Teng Y, Wang L (2024) Logit normalization for long-tail object detection. International Journal of Computer Vision (IJCV) 132(6):2114–2134, DOI 10.1007/S11263-023-01971-Y, URL https://doi.org/10.1007/s11263-023-01971-y

Zhao S, Li B, Xu P, Yue X, Ding G, Keutzer K (2021) MADAN: multi-source adversarial domain aggregation network for domain adaptation. Int J Comput Vis 129(8):2399–2424, DOI 10.1007/S11263-021-01479-3, URL https://doi.org/10.1007/s11263-021-01479-3

Zhou Y, Barnes C, Lu J, Yang J, Li H (2019) On the continuity of rotation representations in neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, pp 5745–5753, DOI 10.1109/CVPR.2019.00589, URL http://openaccess.thecvf.com/content_CVPR_2019/html/Zhou_On_the_Continuity_of_Rotation_Representations_in_Neural_Networks_CVPR_2019_paper.html

Zhu D, Chen J, Shen X, Li X, Elhoseiny M (2023) Minigpt-4: Enhancing vision-language understanding with advanced large language models. CoRR abs/2304.10592, DOI 10.48550/ARXIV.2304.10592, URL https://doi.org/10.48550/arXiv.2304.10592, 2304.10592