

How to Enable LLM with 3D Capacity? A Survey of Spatial Reasoning in LLM

Jirong Zha^{1*}, Yuxuan Fan^{2*}, Xiao Yang², Chen Gao^{3†}, Xinlei Chen^{1†}

¹Shenzhen International Graduate School, Tsinghua University

²The Hong Kong University of Science and Technology (Guang Zhou)

³BNRist, Tsinghua University

zhajirong23@mails.tsinghua.edu.cn, {yfan546, xyang856}@connect.hkust-gz.edu.cn,
chgao96@gmail.com, chen.xinlei@sz.tsinghua.edu.cn

Abstract

3D spatial understanding is essential in real-world applications such as robotics, autonomous vehicles, virtual reality, and medical imaging. Recently, Large Language Models (LLMs), having demonstrated remarkable success across various domains, have been leveraged to enhance 3D understanding tasks, showing potential to surpass traditional computer vision methods. In this survey, we present a comprehensive review of methods integrating LLMs with 3D spatial understanding. We propose a taxonomy that categorizes existing methods into three branches: image-based methods deriving 3D understanding from 2D visual data, point cloud-based methods working directly with 3D representations, and hybrid modality-based methods combining multiple data streams. We systematically review representative methods along these categories, covering data representations, architectural modifications, and training strategies that bridge textual and 3D modalities. Finally, we discuss current limitations, including dataset scarcity and computational challenges, while highlighting promising research directions in spatial perception, multi-modal fusion, and real-world applications.

1 Introduction

Large Language Models (LLMs) have evolved from basic neural networks to advanced transformer models, originally excelling at language tasks by learning from vast text datasets. Recent advancements, however, have extended these models beyond pure linguistic processing to encompass multimodal ability. Their ability to capture complex patterns and relationships [Chen *et al.*, 2024a] now holds promise for spatial reasoning tasks [Ma *et al.*, 2024b]. By applying these enhanced models to challenges such as understanding 3D object relationships and spatial navigation, we open up new opportunities for advancing fields like robotics, computer vision, and augmented reality [Gao *et al.*, 2024].

At the same time, 3D data and 3D modeling techniques have seen significant developments [Ma *et al.*, 2024c], finding extensive applications in virtual and augmented reality, robotics, autonomous vehicles, gaming, medical imaging, and

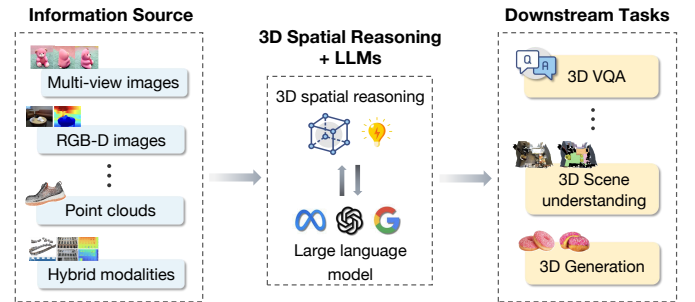


Figure 1: Large Language Models can acquire 3D spatial reasoning capabilities through various input sources including multi-view images, RGB-D images, point clouds, and hybrid modalities, enabling the processing and understanding of three-dimensional information.

more. Unlike traditional two-dimensional images, 3D data provides a richer view of objects and environments, capturing essential spatial relationships and geometry. Such information is critical for tasks like scene reconstruction, object manipulation, and autonomous navigation, where merely text-based descriptions or 2D representations may fall short of conveying the necessary depth or spatial context.

LLMs help Spatial Understanding. Integrating the powerful language comprehension of LLMs with the spatial realism of 3D data creates the potential for highly capable, context-aware systems. From a linguistic perspective, real-world descriptions often reference physical arrangement, orientation, or manipulations of objects in space. Text alone can be imprecise or ambiguous about size, shape, or relative positioning unless one can integrate a robust spatial or visual understanding. Consequently, there is growing interest in enhancing LLMs with a “3D capacity” that enables them to interpret, reason, and even generate three-dimensional representations based on natural language prompts. Such an integrated approach opens up exciting prospects: robots that can follow language instructions more effectively by grounding their commands in 3D context, architects who quickly prototype 3D layouts from textual descriptions, game designers who generate immersive environments for narrative-based experiences, and many other creative applications yet to be envisioned.

Motivation. Although LLMs have been increasingly applied in 3D-related tasks, and [Ma *et al.*, 2024b] provided a system-

atic overview of this field, the rapid advancement of this domain has led to numerous new developments in recent months, necessitating an up-to-date survey that captures these recent breakthroughs. Integrating 3D capacity into LLMs faces several key challenges: (1) the scarcity of high-quality 3D datasets compared to abundant text corpora; (2) the fundamental mismatch between sequential text data and continuous 3D spatial structures, requiring specialized architectural adaptations; and (3) the intensive computational requirements for processing 3D data at scale. While early attempts at combining language and 3D have shown promise, current approaches often remain limited in scope, scalability, and generalization capability. Most existing solutions are domain-specific and lack the broad applicability characteristic of text-based LLMs.

Contribution. The contributions of this work are summarized in the following three aspects: (1) *A structured taxonomy.* We provide a timely and comprehensive survey that distinguishes itself from [Ma *et al.*, 2024b] by presenting a novel perspective on LLM applications in 3D-related tasks: our work constructs a structured taxonomy that categorizes existing research into three primary groups (Figure 2) and offers a forward-looking analysis of the latest breakthroughs, thereby underscoring our unique contributions and the significance of our approach in advancing the field. (2) *A comprehensive review.* Building on the proposed taxonomy, we systematically review the current research progress on LLMs for spatial reasoning tasks. (3) *Future directions.* We highlight the remaining limitations of existing works and suggest potential directions for future research.

2 Preliminary

2.1 Large Language Models

Large Language Models (LLMs) have evolved from early word embeddings to context-aware models. Generative transformers like GPT series have further advanced text generation and few-shot learning. However, these models often struggle with spatial reasoning due to their focus on textual patterns, prompting efforts to integrate external spatial knowledge [Fu *et al.*, 2024].

Vision-Language Models (VLMs) extend LLMs by aligning visual data with text. Some early works leverage co-attentional architectures and contrastive learning, while later models such as BLIP [Li *et al.*, 2022] refine these techniques with larger datasets. Yet, most VLMs process only 2D data, limiting their ability to capture detailed 3D spatial configurations. Integrating 3D context via depth maps, point clouds, or voxels remains challenging, motivating ongoing research toward more robust spatial intelligence.

2.2 3D Data Structures

3D data has different structures, which are essential for understanding the three-dimensional world, and common methods include point clouds, voxel grids, polygonal meshes, neural fields, hybrid representations, and 3D Gaussian splatting. Point clouds represent shapes using discrete points, typically denoted as

$$P = \{p_i \in \mathbb{R}^3 \mid i = 1, \dots, N\},$$

which are storage-efficient but lack surface topology. Voxel grids partition space into uniform cubes, with each voxel $V(i, j, k)$ storing occupancy or distance values, providing detailed structure at the expense of increased memory usage at higher resolutions. Polygonal meshes compactly encode complex geometries through a set of vertices $\{v_i\}$ and faces $\{F_j\}$, though their unstructured and non-differentiable nature poses challenges for integration with neural networks. Neural fields offer an implicit approach by modeling 3D shapes as continuous and differentiable functions, such as

$$f_\theta : \mathbb{R}^3 \rightarrow (c, \sigma),$$

which maps spatial coordinates to color c and density σ . Hybrid representations combine these neural fields with traditional volumetric methods (e.g., integrating f_θ with voxel grids) to achieve high-quality, real-time rendering. Meanwhile, 3D Gaussian splatting enhances point clouds by associating each point p_i with a covariance matrix Σ_i and color c_i , efficiently encoding radiance information for rendering. Each method has its unique strengths and trade-offs, making them suitable for various applications in 3D understanding and generation.

2.3 Proposed taxonomy

We propose a taxonomy that classifies 3D-LLM research into three main categories based on input modalities and integration strategies, as shown in Figure 1: Image-based spatial reasoning encompasses approaches that derive 3D understanding from 2D images. This includes multi-view methods that reconstruct 3D scenes, RGB-D images providing explicit depth information, monocular 3D perception inferring depth from single views, and medical imaging applications. While these approaches benefit from readily available image data and existing vision models, they may struggle with occlusions and viewpoint limitations. Point cloud-based spatial reasoning works directly with 3D point cloud data through three alignment strategies: (1) Direct alignment that immediately connects point features with language embeddings, (2) Step-by-step alignment that follows sequential stages to bridge modalities, and (3) Task-specific alignment customized for particular spatial reasoning requirements. These methods maintain geometric fidelity but face challenges in handling unstructured 3D data. Hybrid modality-based spatial reasoning combines multiple data streams through either tightly or loosely coupled architectures. Tightly coupled approaches integrate modalities through shared embeddings or end-to-end training, while loosely coupled methods maintain modular components with defined interfaces between them. This enables leveraging complementary strengths across modalities but increases architectural complexity.

This taxonomy provides a structured framework for understanding the diverse technical approaches in the field while highlighting the distinct challenges and trade-offs each branch must address. Figure 2 presents a detailed breakdown of representative works in each category.

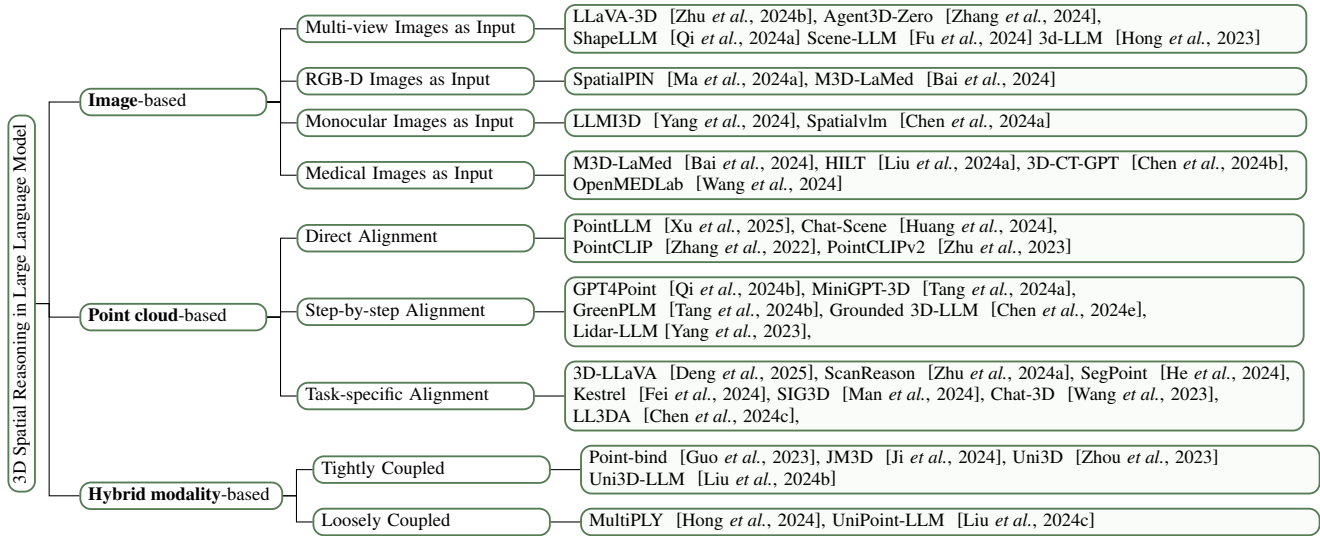


Figure 2: A Taxonomy of Models for Spatial Reasoning with LLMs: Image-based, Point Cloud-based, and Hybrid Modality-based Approaches and Their Subdivisions.

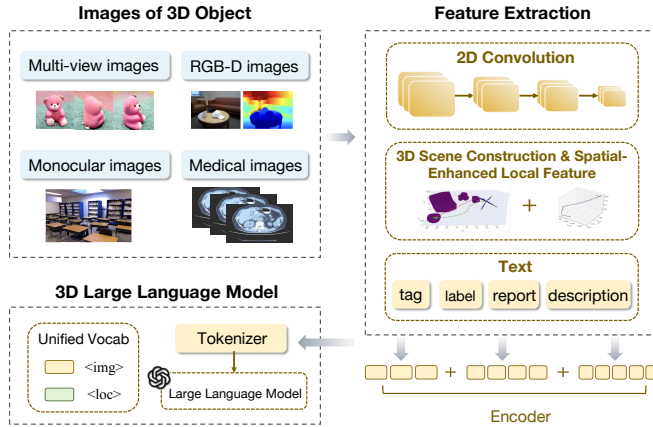


Figure 3: An overview of image-based approaches.

3 Recent Advances of Spatial Reasoning in LLM

3.1 Image-based Spatial Reasoning

Image-based spatial reasoning methods can be categorized based on their input modalities: multi-view images, monocular images, RGB-D images, and 3D medical images shown in Figure 3. Each modality offers unique advantages for enhancing 3D understanding in Large Language Models (LLMs). Multi-view images provide spatial data from different perspectives, monocular images extract 3D insights from a single view, RGB-D images incorporate depth information, and 3D medical images address domain-specific challenges in healthcare. These categories highlight the strengths and challenges of each approach in improving spatial reasoning capabilities.

Multi-view Images as input. Some real-world scenarios can support the collection of multi-view images, such as multi-UAV systems [Chen *et al.*, 2024d]. Several studies explore

multi-view images to enhance LLMs’ spatial understanding. LLaVA-3D [Zhu *et al.*, 2024b] leverages multi-view images and 3D positional embeddings to create 3D Patches, achieving state-of-the-art 3D spatial understanding while maintaining 2D image understanding capabilities. Agent3D-Zero [Zhang *et al.*, 2024] utilizes multiple images from different viewpoints, enabling VLMs to perform robust reasoning and understand spatial relationships, achieving zero-shot scene understanding. ShapeLLM [Qi *et al.*, 2024a] also uses multi-view image input, with robustness to occlusions. Scene-LLM [Fu *et al.*, 2024] uses multi-view images to build 3D feature representations, incorporating scene-level and egocentric 3D information to support interactive planning. SpatialPIN [Ma *et al.*, 2024a] enhances VLM’s spatial reasoning by decomposing, understanding and reconstructing explicit 3D representations from multi-view images and generalizes to various 3D tasks. LLMI3D [Yang *et al.*, 2024] extracts spatially enhanced local features from high-resolution images using CNNs and a depth predictor and uses ViT to obtain tokens from low-resolution images. It employs a spatially enhanced cross-branch attention mechanism to effectively mine spatial local features of objects and uses geometric projection to handle. Extracting multi-view features results in huge computational overhead and ignores the essential geometry and depth information. Additionally, plain texts often lead to ambiguities especially in cluttered and complex 3D environments [Chen *et al.*, 2024c]. ConceptGraphs [Gu *et al.*,] proposes a graph-structured representation for 3D scenes that operates with an open vocabulary, which is developed by utilizing 2D foundation models and integrating their outputs into a 3D format through multiview association.

Monocular Image as input. LLMI3D [Yang *et al.*, 2024] uses a single 2D image for 3D perception, enhancing performance through spatial local feature mining, 3D query token decoding, and geometry-based 3D reasoning. It uses a depth predictor and CNN to extract spatial local features and uses

learnable 3D query tokens for geometric coordinate regression. It combines black-box networks and white-box projection to address changes in camera focal lengths.

RGB-D Image as Input. Depth is estimated in SpatialPIN [Ma *et al.*, 2024a] by ZoeDepth when finding field of view (FOV) through perspective fields and provided for 3D-scene understanding and reconstruction. M3D-LaMed [Bai *et al.*, 2024] pre-trains the 3D medical vision encoder with medical image slices along depth and introduces end-to-end tuning to integrate 3D information into LLM.

3D Medical Image as input. Unlike previous research focused on 2D medical images, integrating multi-modal other information such as textual descriptions, M3D-LaMed [Bai *et al.*, 2024] is specifically designed for 3D CT images by analyzing spatial features. It demonstrates excellent performance across multiple tasks, including image-text retrieval, report generation, visual question answering, localization, and segmentation. In order to generate radiology reports automatically, a brand-new framework [Liu *et al.*, 2024a] is proposed to employs low-resolution (LR) visual tokens as queries to extract information from high-resolution (HR) tokens, ensuring that detailed information is retained across HR volumes while minimizing computational costs by processing only the HR-informed LR visual queries. 3D-CT-GPT [Chen *et al.*, 2024b], based medical visual language model, is tailored for the generation of radiology reports from 3D CT scans, with a focus on chest CTs. OpenMEDLab [Wang *et al.*, 2024] comprises and publishes a variety of medical foundation models to process multi-modal medical data including Color Fundus Photography (CFP), Optical Coherence Tomography (OCT), endoscopy videos, CT&MR volumes and other pathology images.

Discussion. Image-based spatial reasoning methods offer significant advantages, such as easy data acquisition and integration with pre-trained 2D models. Multi-view images provide rich spatial information, while depth estimation enhances scene understanding. However, challenges remain, including limited depth from single views, scale uncertainty, occlusion, and viewpoint dependency. These methods also face issues with visual hallucinations, generalization to novel scenes, and high computational costs. Future research should focus on improving multi-view integration and depth estimation to address these limitations.

3.2 Recent Advances of Point Cloud-based Spatial Reasoning

As shown in Figure 4, point cloud-based spatial reasoning has advanced significantly in recent years, employing three main alignment methods: Direct, Step-by-step, and Task-specific Alignment. These methods are essential for integrating point cloud data with language models to enable effective spatial reasoning. Direct Alignment establishes immediate connections between point cloud features and language model embeddings, while Step-by-step Alignment follows a sequential process through multiple stages. Task-specific Alignment is customized for particular spatial reasoning requirements. The choice of method depends on specific application needs and constraints.

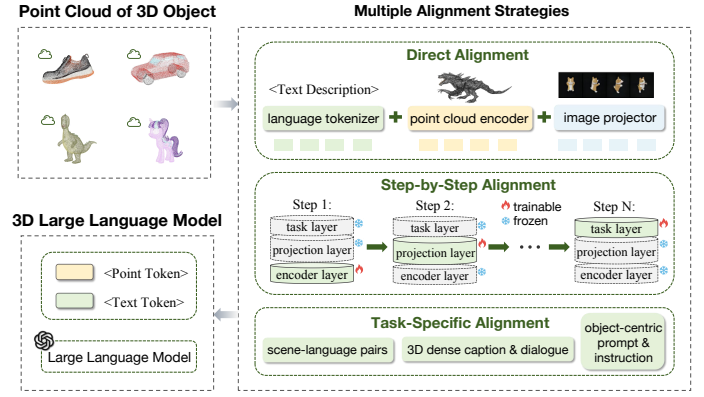


Figure 4: An overview of point cloud-based approaches.

Direct Alignment. Direct alignment methods create direct connections between point cloud data and language models. PointCLIP [Zhang *et al.*, 2022] was a pioneer, projecting point clouds into multi-view depth maps and using CLIP’s pre-trained visual encoder for feature extraction, which was then aligned with textual features through a hand-crafted template. This approach showed promising results in zero-shot and few-shot classification tasks by transferring 2D knowledge to the 3D domain. PointCLIP V2 [Zhu *et al.*, 2023] improved the projection quality with a realistic projection module and used GPT-3 for generating 3D-specific text descriptions, achieving better performance in zero-shot classification, part segmentation, and object detection. Chat-Scene [Huang *et al.*, 2024] introduced object identifiers to facilitate object referencing during user-assistant interactions, representing scenes through object-centric embeddings. PointLLM [Xu *et al.*, 2025] advanced the field by integrating a point cloud encoder with a powerful LLM, effectively fusing geometric, appearance, and linguistic information, and overcoming data scarcity with automated generation. These methods demonstrate the potential for effective 3D point cloud understanding through language models, enabling improved spatial reasoning and human-AI interaction.

Step-by-step Alignment. Step-by-step alignment has gained popularity in integrating point cloud features with language models. Notable approaches include GPT4Point [Qi *et al.*, 2024b], which uses a Bert-based Point-QFormer for point-text feature alignment, followed by object generation. Grounded 3D-LLMs [Chen *et al.*, 2024e] first aligns 3D scene embeddings with textual descriptions via contrastive pre-training, then fine-tunes with referent tokens. LiDAR-LLMs [Yang *et al.*, 2023] employ a three-stage process: cross-modal alignment, object-centric learning, and high-level instruction fine-tuning. MiniGPT-3D [Tang *et al.*, 2024a] follows a four-stage strategy, from point cloud projection to advanced model enhancements using Mixture of Query Experts. GreenPLM [Tang *et al.*, 2024b] uses a three-stage method that aligns a text encoder with an LLM using large text data, followed by point-LLM alignment with 3D data. These step-by-step approaches highlight the gradual improvement of spatial reasoning in 3D contexts, offering valuable insights for future research.

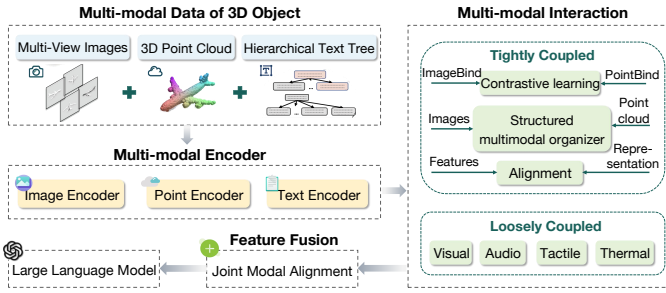


Figure 5: An overview of hybrid modality-based approaches.

Task-specific Alignment. Task-specific alignment customizes models for specific spatial reasoning tasks to improve performance and generalization. SceneVerse [Jia *et al.*, 2024] introduces a large 3D vision-language dataset and Grounded Pre-training for Scenes (GPS), using multi-level contrastive alignment for unified scene-text alignment, achieving state-of-the-art results in tasks like 3D visual grounding and question answering. LL3DA [Chen *et al.*, 2024c] presents a dialogue system that integrates textual instructions and visual interactions, excelling in complex 3D environments. Chat-3D [Wang *et al.*, 2023] proposes a three-stage training scheme to align 3D scene representations with language models, capturing spatial relations with limited data. VisProg [Yuan *et al.*, 2024] introduces visual programming for zero-shot open-vocabulary 3D grounding, leveraging LLMs to generate and execute programmatic representations. These task-specific approaches highlight the importance of adapting models to complex spatial relationships, enabling robust performance even with limited data or zero-shot tasks.

Discussion. The three alignment approaches, Direct, Step-by-step, and Task-specific, each have distinct strengths and challenges. Direct alignment offers efficiency and quick results but struggles with complex spatial relationships. Step-by-step alignment improves feature integration at the cost of higher computational resources and training time. Task-specific alignment excels in specialized tasks but may lack broader applicability.

3.3 Hybrid Modality-based Spatial Reasoning

Hybrid modality-based spatial reasoning integrates point clouds, images, and LLMs through Tightly Coupled and Loosely Coupled approaches, as shown in Figure 5. The Tightly Coupled approach fosters close integration, enabling seamless interaction and high performance, while the Loosely Coupled approach promotes modularity, allowing independent operation of components for greater scalability and flexibility at the cost of reduced real-time interaction.

Tightly Coupled. Several recent works have explored tightly integrated approaches for spatial reasoning across point clouds, images and language modalities: Point-Bind [Guo *et al.*, 2023] proposes a joint embedding space to align point clouds with images and text through contrastive learning. It leverages ImageBind to construct unified representations that enable tasks like zero-shot classification, open-world understanding and multi-modal generation. The tight coupling allows Point-Bind to reason about point clouds using both visual

and linguistic cues. JM3D [Ji *et al.*, 2024] introduces a Structured Multimodal Organizer that tightly fuses multi-view images and hierarchical text trees with point clouds. This coupled architecture enables detailed spatial understanding by leveraging complementary information across modalities. The Joint Multi-modal Alignment further enhances the synergistic relationships between visual and linguistic features. Uni3D [Zhou *et al.*, 2023] employs a unified transformer architecture that directly aligns point cloud features with image-text representations. By tightly coupling the modalities through end-to-end training, it achieves strong performance on tasks like zero-shot classification and open-world understanding. The shared backbone enables efficient scaling to billion-parameter models. Uni3D-LLM [Liu *et al.*, 2024b] extends this tight coupling to LLMs through an LLM-to-Generator mapping block. This enables unified perception, generation and editing of point clouds guided by natural language. The tight integration allows leveraging rich semantic knowledge from LLMs while maintaining high-quality 3D understanding.

Loosely Coupled. Loosely coupled approaches maintain greater independence between different modalities while still enabling interaction through well-defined interfaces. MultiPLY [Hong *et al.*, 2024] proposes a multisensory embodied LLM that handles multiple input modalities (visual, audio, tactile, thermal) through separate encoders. The modalities are processed independently and communicate through action tokens and state tokens. This decoupled design allows the system to process each modality with specialized encoders optimized for that data type, while enabling scalability and modularity in the system architecture. Similarly, UniPoint-LLM [Liu *et al.*, 2024c] introduces a Multimodal Universal Token Space (MUTS) that loosely connects point clouds and images through independent encoders and a shared mapping layer. This modular design allows easy integration of new modalities and simplified training by only requiring alignment between new modalities and text, rather than pairwise alignment between all modalities. The main benefits of loosely coupled architectures include greater modularity and flexibility in system design, easier integration of new modalities, and independent scaling of different components. However, this approach may result in less optimal joint representation learning, reduced real-time interaction capabilities, and potential information loss between modalities compared to tightly coupled approaches.

Discussion. The choice between tightly and loosely coupled approaches presents important tradeoffs in multimodal spatial reasoning systems. Tightly coupled approaches like Point-Bind and JM3D offer stronger joint representation learning and real-time interaction capabilities through end-to-end training and shared feature spaces. This makes them particularly suitable for applications requiring detailed spatial understanding and precise control. However, they can be more complex to train and scale, and adding new modalities may require significant architectural changes. In contrast, loosely coupled approaches like MultiPLY and UniPoint-LLM provide greater modularity and flexibility, making them easier to extend and maintain. They allow independent optimization of different components and simplified training procedures,

	Model	Data Source	Alignment Type	Pre-training	Fine-tuning	Task	Code
Image - based	LLaVA-3D [Zhu <i>et al.</i> , 2024b]	Multi-view Images	-	✓	✓	3D VQA, 3D Scene Understanding	code
	Agent3D-Zero [Zhang <i>et al.</i> , 2024]	Multi-view Images	-	✓	✗	3D VQA, 3D Semantic Segmentation	✗
	ShapeLLM [Qi <i>et al.</i> , 2024a]	Multi-view Images	-	✓	✓	3D Object Classification, 3D Scene Captioning	code
	Scene-LLM [Fu <i>et al.</i> , 2024]	Multi-view Images	-	✓	✓	3D VQA, Dense Captioning	✗
	SpatialPIN [Ma <i>et al.</i> , 2024a]	RGB-D Images	-	✓	✗	3D Motion Planning, Task Video Generation	✗
	LLM3D [Yang <i>et al.</i> , 2024]	Monocular Images	-	✓	✓	3D Grounding, 3D VQA	✗
	Spatialvlm [Chen <i>et al.</i> , 2024a]	Monocular Images	-	✓	✓	Dense Reward Annotator, Spatial Data Generation	code
	M3D-LaMed [Bai <i>et al.</i> , 2024]	Medical Images	-	✓	✓	3D VQA, 3D VLP	code
	HILT [Liu <i>et al.</i> , 2024a]	Medical Images	-	✓	✓	3DHRG	✗
	3D-CT-GPT [Chen <i>et al.</i> , 2024b]	Medical Images	-	✓	✓	Radiology Report Generation, 3D VQA	✗
Point Cloud - based	OpenMEDLab [Wang <i>et al.</i> , 2024]	Medical Images	-	✓	✓	Medical Imaging	code
	PointLLM [Xu <i>et al.</i> , 2025]	Point Cloud	Direct Alignment	✓	✓	3D Object Classification, 3D Object Captioning	code
	Chat-Scene [Huang <i>et al.</i> , 2024]	Point Cloud	Direct Alignment	✓	✓	3D Visual Grounding, 3D Scene Captioning	code
	PointCLIP [Zhang <i>et al.</i> , 2022]	Point Cloud	Direct Alignment	✓	✓	3D Point Cloud Classification	code
	PointCLIPv2 [Zhu <i>et al.</i> , 2023]	Point Cloud	Direct Alignment	✓	✓	3D Point Cloud Classification	code
	GPT4Point [Qi <i>et al.</i> , 2024b]	Point Cloud	Step-by-step Alignment	✓	✓	3D Object Understanding	code
	MiniGPT-3D [Tang <i>et al.</i> , 2024a]	Point Cloud	Step-by-step Alignment	✓	✓	3D Object Classification, 3D Object Captioning	code
	GreenPLM [Tang <i>et al.</i> , 2024b]	Point Cloud	Step-by-step Alignment	✓	✓	3D Object Classification	code
	Grounded 3D-LLM [Chen <i>et al.</i> , 2024c]	Point Cloud	Step-by-step Alignment	✓	✓	3D Object Detection, 3D VQA	code
	Lidar-LLM [Yang <i>et al.</i> , 2023]	Point Cloud	Step-by-step Alignment	✓	✓	3D Captioning, 3D Grounding	code
	3D-LLaVA [Deng <i>et al.</i> , 2025]	Point Cloud	Task-specific Alignment	✓	✓	3D VQA, 3D Captioning	code
	ScanReason [Zhu <i>et al.</i> , 2024a]	Point Cloud	Task-specific Alignment	✓	✓	3D Reasoning Grounding	code
	SegPoint [He <i>et al.</i> , 2024]	Point Cloud	Task-specific Alignment	✓	✓	3D Instruction Segmentation	✗
	Kestrel [Fei <i>et al.</i> , 2024]	Point Cloud	Task-specific Alignment	✓	✓	Part-Aware Point Grounding	✗
	SIG3D [Man <i>et al.</i> , 2024]	Point Cloud	Task-specific Alignment	✓	✓	Situation Estimation	code
Hybrid - based	Chat-3D [Wang <i>et al.</i> , 2023]	Point Cloud	Task-specific Alignment	✓	✓	3D VQA	code
	LL3DA [Chen <i>et al.</i> , 2024c]	Point Cloud	Task-specific Alignment	✓	✓	3D Dense Captioning	code
	Point-bind [Guo <i>et al.</i> , 2023]	Point cloud, Image	Tightly Coupled	✓	✓	3D Cross-modal Retrieval, Any-to-3D Generation	code
	JM3D [Ji <i>et al.</i> , 2024]	Point cloud, Image	Tightly Coupled	✓	✓	Image-3D Retrieval, 3D Part Segmentation	code
	Uni3D [Zhou <i>et al.</i> , 2023]	Point cloud, Image	Tightly Coupled	✓	✓	Zero-shot Shape Classification	code
Hybrid - based	Uni3D-LLM [Liu <i>et al.</i> , 2024b]	Point cloud, Image	Tightly Coupled	✓	✓	3D VQA	✗
	MultiPLY [Hong <i>et al.</i> , 2024]	Point cloud, Image	Loosely Coupled	✓	✓	Object retrieval	code
	UniPoint-LLM [Liu <i>et al.</i> , 2024c]	Point cloud, Image	Loosely Coupled	✓	✓	3D generation, 3D VQA	✗

Table 1: Taxonomy of Large Language Models with spatial reasoning capability. This table presents a comprehensive comparison of various 3D vision-language models categorized by their input modalities (image-based, point cloud-based, and hybrid-based), showing their data sources, alignment types, training strategies (pre-training and fine-tuning), primary tasks, and code availability. The models are organized into three main categories based on their input type: image-based models, point cloud-based models, and hybrid models that utilize both modalities.

but may sacrifice some performance in tasks requiring fine-grained cross-modal understanding. The optimal choice ultimately depends on specific application requirements. Tightly coupled architectures may be preferred for specialized high-performance systems, while loosely coupled designs better suit general-purpose platforms prioritizing extensibility and maintainability. Future work may explore hybrid approaches that combine the benefits of both paradigms, potentially using adaptive coupling mechanisms that adjust based on task demands.

4 Applications

A key research focus leverages LLMs to enhance robotic embodied intelligence, enabling machines to interpret natural language commands for real-world tasks. This includes robotic control, navigation, and manipulation, where LLMs parse instructions, generate action plans, and adapt to dynamic environments, for instance, guiding robots to locate objects in cluttered spaces using text-based prompts [Zhao *et al.*, 2025].

3D Scene Understanding. Advanced 3D scene analysis integrates multimodal data (e.g., images, point clouds, text) for tasks like open-vocabulary segmentation, semantic mapping, and spatial reasoning. Central to this is 3D visual question answering (3D-VQA) [Zhan *et al.*, 2025], requiring models to interpret queries about object attributes, spatial relationships, or contextual roles within scenes. Context-aware systems further account for user perspectives to deliver precise responses.

Cross-Domain Applications. In healthcare, LLMs analyze volumetric medical scans (e.g., CT) for lesion detection and

automated diagnostics. Autonomous driving systems utilize 3D-capable LLMs to interpret traffic scenes, aiding object detection and tracking and path planning [Zha *et al.*, 2025; Zha *et al.*, 2024]. Design-oriented applications include generating indoor layouts from textual requirements, while educational tools employ interactive 3D environments to teach spatial concepts.

5 Challenges and Future Directions

Table 1 summarizes the models that leverage LLMs to assist graph-related tasks according to the proposed taxonomy. Based on the above review and analysis, we believe that there is still much space for further enhancement in this field. Recent advances in integrating LLMs with three-dimensional (3D) data have demonstrated considerable promise. However, numerous challenges must still be overcome to realize robust and practical 3D-aware LLMs. Below, we summarize these obstacles and then outline potential pathways to address them, highlighting key research directions for the future.

5.1 Challenges

Weak Spatial Reasoning and Representation. Multimodal LLMs (MLLMs) exhibit limited acuity in 3D spatial understanding, struggling with fine-grained relationships (e.g., front/back distinctions, occluded object localization) and precise geometric outputs (distances, angles). These issues stem partly from mismatches between unstructured point clouds and sequence-based LLM architectures, where high-dimensional

3D data incur prohibitive token counts or oversimplified encodings.

Data and Evaluation Gaps. Progress in 3D-aware LLMs is hindered by the scarcity of high-quality 3D-text paired datasets. Unlike the abundant resources for 2D images and video, the 3D domain lacks standardized, richly annotated datasets crucial for training robust models. Existing benchmarks focus mainly on discriminative tasks like classification and retrieval, emphasizing category differentiation rather than generating rich and descriptive 3D scene outputs. Consequently, evaluations often rely on subjective metrics (e.g., human or GPT-based judgments) that can lack consistency. Advancing the field requires developing objective, comprehensive benchmarks that assess both open-vocabulary generation and the spatial plausibility of descriptions relative to the underlying 3D structure.

Multimodal Integration and Generalization. Fusing 3D data (e.g., point clouds) with other modalities like 2D imagery, audio, or text poses significant challenges due to their distinct structural characteristics. The conversion and alignment of high-dimensional 3D data with lower-dimensional representations can lead to a loss of intricate details, diluting the original 3D richness. Moreover, current models often struggle with open-vocabulary recognition, limiting their ability to identify or describe objects outside of their training data, especially when encountering unseen scenes or novel objects. This difficulty is further compounded by the variability of natural language, from colloquial expressions to domain-specific terminology, and by noisy inputs. Thus, more sophisticated multimodal integration techniques and generalization strategies are needed to preserve geometric fidelity while accommodating diverse, unpredictable inputs.

Complex Task Definition. While 3D-aware LLMs excel in controlled settings, they lack frameworks for nuanced language-context inference in dynamic environments. Task decomposition and scalable encoding methods are needed to balance geometric fidelity with computational tractability, particularly for interactive applications requiring real-time spatial reasoning.

5.2 Future Directions

Enhancing 3D Perception and Representations. Addressing spatial reasoning gaps requires richer 3D-text datasets (e.g., from robotics, gaming, autonomous driving) and model architectures that encode geometric relationships. Multi-view data and robust depth cues can improve orientation, distance, and occlusion estimation. Compact 3D tokens and refined encoding/decoding methods may bridge unstructured point clouds with sequence-based models, enabling fine-grained spatial understanding and generation.

Multi-Modal Fusion and Instruction Understanding. Tighter integration of modalities (point clouds, images, text, audio) via unified latent spaces or attention mechanisms could preserve subtle geometric and semantic details. Enhanced instruction processing, including hierarchical task decomposition, contextual interpretation, and robustness to dialects/terminology, would improve compositional reasoning in 3D environments and broaden real-world applicability. Fur-

thermore, by leveraging these integrated representations, models can more adeptly adapt to complex instructions and novel scenarios, ultimately paving the way for more robust and versatile 3D reasoning systems.

Cross-Scene Generalization and Robust Evaluation. Open-vocabulary 3D understanding demands large-scale pre-training on diverse scenes and transfer/lifelong learning paradigms for adapting to novel objects or environments. This understanding extends beyond predefined categories to generalize to unseen objects and scenes. For instance, models need to comprehend “an old rocking chair” even if this specific type of chair never appeared in the training data.

Expanding Applications for Autonomous Systems. 3D-aware LLMs hold potential in robotics (navigation, manipulation), medical imaging (lesion detection), architectural design, and interactive education. Future systems may integrate environmental constraints, user perspectives, and object affordances for autonomous planning and decision-making in dynamic 3D contexts.

Collectively, these challenges and potential directions underscore the field’s rapid evolution and its equally significant open questions. Moving forward, more robust 3D-specific data resources, better model architectures, and more refined evaluation protocols will be essential to unlock the full potential of LLMs in three-dimensional settings, and ultimately bring intelligent and multimodal understanding closer to real-world deployment.

6 Conclusion

The integration of LLMs with 3D data is a dynamic research area. This survey categorized 3D-LLM research into image-based, point cloud-based, and hybrid modality-based spatial reasoning. It reviewed state-of-the-art methods, their applications in multiple fields, and associated challenges. Notably, image-based methods have data-related advantages but face issues like depth information shortage. Point cloud-based methods offer precise 3D details but encounter data-handling difficulties. Hybrid methods combine strengths yet struggle with data alignment. Applications are diverse, but challenges such as weak spatial perception, data scarcity, and evaluation problems exist. Future research should focus on enhancing 3D perception, improving multi-modal fusion, expanding generalization, developing evaluation metrics, enhancing instruction understanding, optimizing 3D representations, and exploring continuous learning. By addressing these, we can unlock the full potential of 3D-aware LLMs for real-world deployment and industry advancement.

Acknowledgments

This paper was supported by the National Key R&D program of China (2022YFC3300703), the Natural Science Foundation of China under Grant 62371269, Beijing National Research Center for Information Science and Technology (No. BNR2023RC01003, BNR2023TD03006).

References

- [Bai *et al.*, 2024] Fan Bai, Yuxin Du, Tiejun Huang, Max Q-H Meng, and Bo Zhao. M3d: Advancing 3d medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*, 2024.
- [Chen *et al.*, 2024a] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465, 2024.
- [Chen *et al.*, 2024b] Hao Chen, Wei Zhao, Yingli Li, Tianyang Zhong, Yisong Wang, Youlan Shang, Lei Guo, Junwei Han, Tianming Liu, et al. 3d-ct-gpt: Generating 3d radiology reports through integration of large vision-language models. *arXiv preprint arXiv:2409.19330*, 2024.
- [Chen *et al.*, 2024c] Sijin Chen, Xin Chen, Chi Zhang, Ming-sheng Li, Gang Yu, Fei, et al. Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26428–26438, 2024.
- [Chen *et al.*, 2024d] Xuecheng Chen, Zijian Xiao, Yuhan Cheng, Chen-Chun Hsia, Haoyang Wang, Jingao Xu, Susu Xu, Fan Dang, Xiao-Ping Zhang, Yunhao Liu, and Xinlei Chen. Soscheduler: Toward proactive and adaptive wildfire suppression via multi-uav collaborative scheduling. *IEEE Internet of Things Journal*, 11(14):24858–24871, 2024.
- [Chen *et al.*, 2024e] Yilun Chen, Shuai Yang, Haifeng Huang, Tai Wang, Ruiyuan Lyu, et al. Grounded 3d-llm with referent tokens. *arXiv preprint arXiv:2405.10370*, 2024.
- [Deng *et al.*, 2025] Jiajun Deng, Tianyu He, Li Jiang, Tianyu Wang, Feras Dayoub, and Ian Reid. 3d-llava: Towards generalist 3d llms with omni superpoint transformer. *arXiv preprint arXiv:2501.01163*, 2025.
- [Fei *et al.*, 2024] Junjie Fei, Mahmoud Ahmed, Jian Ding, Eslam Mohamed Bakr, et al. Kestrel: Point grounding multimodal llm for part-aware 3d vision-language understanding. *arXiv preprint arXiv:2405.18937*, 2024.
- [Fu *et al.*, 2024] Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. Scene-llm: Extending language model for 3d visual understanding and reasoning. *arXiv preprint arXiv:2403.11401*, 2024.
- [Gao *et al.*, 2024] Chen Gao, Baining Zhao, Weichen Zhang, Jinzhu Mao, Jun Zhang, Zhiheng Zheng, Fanhang Man, et al. Embodiedcity: A benchmark platform for embodied agent in real-world city environment. *arXiv preprint arXiv:2410.09604*, 2024.
- [Gu *et al.*,] Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In *2024 IEEE International Conference on Robotics and Automation*, pages 5021–5028.
- [Guo *et al.*, 2023] Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming Han, et al. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. *arXiv preprint arXiv:2309.00615*, 2023.
- [He *et al.*, 2024] Shuting He, Henghui Ding, Xudong Jiang, and Bihan Wen. Segpoint: Segment any point cloud via large language model. In *European Conference on Computer Vision*, pages 349–367. Springer, 2024.
- [Hong *et al.*, 2023] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems*, 36:20482–20494, 2023.
- [Hong *et al.*, 2024] Yining Hong, Zishuo Zheng, Peihao Chen, Yian Wang, Junyan Li, and Chuang Gan. Multiply: A multisensory object-centric embodied large language model in 3d world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26406–26416, 2024.
- [Huang *et al.*, 2024] Haifeng Huang, Yilun Chen, Zehan Wang, Rongjie Huang, Runsen Xu, Tai Wang, Luping Liu, Xize Cheng, Yang Zhao, Jiangmiao Pang, et al. Chat-scene: Bridging 3d scene and large language models with object identifiers. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- [Ji *et al.*, 2024] Jiayi Ji, Haowei Wang, Changli Wu, Yiwei Ma, Xiaoshuai Sun, and Rongrong Ji. Jm3d & jm3d-llm: Elevating 3d representation with joint multi-modal cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
- [Jia *et al.*, 2024] Baoxiong Jia, Yixin Chen, Huangyue Yu, Yan Wang, Xuesong Niu, Tengyu Liu, Qing Li, and Siyuan Huang. Sceneverse: Scaling 3d vision-language learning for grounded scene understanding. In *European Conference on Computer Vision*, pages 289–310. Springer, 2024.
- [Li *et al.*, 2022] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.
- [Liu *et al.*, 2024a] Che Liu, Zhongwei Wan, Yuqi Wang, Hui Shen, Haozhe Wang, Kangyu Zheng, Mi Zhang, and Rossella Arcucci. Benchmarking and boosting radiology report generation for 3d high-resolution medical images. *arXiv preprint arXiv:2406.07146*, 2024.
- [Liu *et al.*, 2024b] Dingning Liu, Xiaoshui Huang, Yuenan Hou, Zhihui Wang, Zhenfei Yin, Yongshun Gong, Peng Gao, and Wanli Ouyang. Uni3d-llm: Unifying point cloud perception, generation and editing with large language models. *arXiv preprint arXiv:2402.03327*, 2024.
- [Liu *et al.*, 2024c] Dingning Liu, Xiaoshui Huang, Zhihui Wang, Zhenfei Yin, Peng Gao, Yujiao Wu, Yuenan Hou, et al. Pointmllm: Aligning multi-modality with llm for point cloud understanding, generation and editing. 2024.

- [Ma *et al.*, 2024a] Chenyang Ma, Kai Lu, Ta-Ying Cheng, Niki Trigoni, and Andrew Markham. Spatialpin: Enhancing spatial reasoning capabilities of vision-language models through prompting and interacting 3d priors. *arXiv preprint arXiv:2403.13438*, 2024.
- [Ma *et al.*, 2024b] Xianzheng Ma, Yash Bhalgat, Brandon Smart, Shuai Chen, Xinghui Li, Jian Ding, Jindong Gu, et al. When llms step into the 3d world: A survey and meta-analysis of 3d tasks via multi-modal large language models. *arXiv preprint arXiv:2405.10255*, 2024.
- [Ma *et al.*, 2024c] Yuexin Ma, Tai Wang, Xuyang Bai, Huitong Yang, Yuenan Hou, Yaming Wang, Yu Qiao, et al. Vision-centric bev perception: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
- [Man *et al.*, 2024] Yunze Man, Liang-Yan Gui, and Yu-Xiong Wang. Situational awareness matters in 3d vision language reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13678–13688, 2024.
- [Qi *et al.*, 2024a] Zekun Qi, Runpei Dong, Shaochen Zhang, Haoran Geng, Chunrui Han, Zheng Ge, Li Yi, and Kaisheng Ma. Shapellm: Universal 3d object understanding for embodied interaction. In *European Conference on Computer Vision*, pages 214–238. Springer, 2024.
- [Qi *et al.*, 2024b] Zhangyang Qi, Ye Fang, Zeyi Sun, Xiaoyang Wu, Tong Wu, Jiaqi Wang, Dahua Lin, and Hengshuang Zhao. Gpt4point: A unified framework for point-language understanding and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26417–26427, 2024.
- [Tang *et al.*, 2024a] Yuan Tang, Xu Han, Xianzhi Li, Qiao Yu, Yixue Hao, Long Hu, et al. Minigpt-3d: Efficiently aligning 3d point clouds with large language models using 2d priors. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6617–6626, 2024.
- [Tang *et al.*, 2024b] Yuan Tang, Xu Han, Xianzhi Li, Qiao Yu, Jinfeng Xu, Yixue Hao, Long Hu, and Min Chen. More text, less point: Towards 3d data-efficient point-language understanding. *arXiv preprint arXiv:2408.15966*, 2024.
- [Wang *et al.*, 2023] Zehan Wang, Haifeng Huang, Yang Zhao, Ziang Zhang, and Zhou Zhao. Chat-3d: Data-efficiently tuning large language model for universal dialogue of 3d scenes. *arXiv preprint arXiv:2308.08769*, 2023.
- [Wang *et al.*, 2024] Xiaosong Wang, Xiaofan Zhang, Guotai Wang, Junjun He, Zhongyu Li, et al. Openmedlab: An open-source platform for multi-modality foundation models in medicine. *arXiv preprint arXiv:2402.18028*, 2024.
- [Xu *et al.*, 2025] Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, et al. Pointllm: Empowering large language models to understand point clouds. In *European Conference on Computer Vision*, pages 131–147. Springer, 2025.
- [Yang *et al.*, 2023] Senqiao Yang, Jiaming Liu, Ray Zhang, Mingjie Pan, Zoey Guo, Xiaoqi Li, Zehui Chen, Peng Gao, Yandong Guo, and Shanghang Zhang. Lidar-llm: Exploring the potential of large language models for 3d lidar understanding. *arXiv preprint arXiv:2312.14074*, 2023.
- [Yang *et al.*, 2024] Fan Yang, Sicheng Zhao, Yanhao Zhang, Haoxiang Chen, Hui Chen, Wenbo Tang, Haonan Lu, et al. Llm3d: Empowering llm with 3d perception from a single 2d image. *arXiv preprint arXiv:2408.07422*, 2024.
- [Yuan *et al.*, 2024] Zhihao Yuan, Jinke Ren, Chun-Mei Feng, Hengshuang Zhao, Shuguang Cui, et al. Visual programming for zero-shot open-vocabulary 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20623–20633, 2024.
- [Zha *et al.*, 2024] Jirong Zha, Nan Zhou, Zhenyu Liu, Tao Sun, and Xinlei Chen. Diffusion-based filter for fast and accurate collaborative tracking with low data transmission. *Authorea Preprints*, 2024.
- [Zha *et al.*, 2025] Jirong Zha, Yuxuan Fan, Kai Li, Han Li, Chen Gao, Xinlei Chen, and Yong Li. Dimm: Decoupled multi-hierarchy kalman filter for 3d object tracking. *arXiv preprint arXiv:2505.12340*, 2025.
- [Zhan *et al.*, 2025] Weichen Zhan, Zile Zhou, Zhiheng Zheng, Chen Gao, Jinqiang Cui, Yong Li, Xinlei Chen, et al. Open3DVQA: A benchmark for comprehensive spatial reasoning with multimodal large language model in open space. *arXiv preprint arXiv:2503.11094*, 2025.
- [Zhang *et al.*, 2022] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, et al. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8552–8562, 2022.
- [Zhang *et al.*, 2024] Sha Zhang, Di Huang, Jiajun Deng, Shixiang Tang, Wanli Ouyang, et al. Agent3d-zero: An agent for zero-shot 3d understanding. In *European Conference on Computer Vision*, pages 186–202. Springer, 2024.
- [Zhao *et al.*, 2025] Baining Zhao, Jianjie Fang, Zichao Dai, Ziyu Wang, Jirong Zha, Weichen Zhang, Chen Gao, Yue Wang, Jinqiang Cui, Xinlei Chen, et al. Urbanvideo-bench: benchmarking vision-language models on embodied intelligence with video data in urban spaces. *arXiv preprint arXiv:2503.06157*, 2025.
- [Zhou *et al.*, 2023] Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3d: Exploring unified 3d representation at scale. *arXiv preprint arXiv:2310.06773*, 2023.
- [Zhu *et al.*, 2023] Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyu Guo, Ziyao Zeng, Zipeng Qin, Shanghang Zhang, and Peng Gao. Pointclip v2: Prompting clip and gpt for powerful 3d open-world learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2639–2650, 2023.
- [Zhu *et al.*, 2024a] Chenming Zhu, Tai Wang, Wenwei Zhang, Kai Chen, and Xihui Liu. Scanreason: Empowering 3d visual grounding with reasoning capabilities. In *European Conference on Computer Vision*, pages 151–168. Springer, 2024.
- [Zhu *et al.*, 2024b] Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Llava-3d: A simple yet effective pathway to empowering llms with 3d-awareness. *arXiv preprint arXiv:2409.18125*, 2024.