Trends in Multi-modal Large Language Models for 3D Point Cloud Understanding

Seok Bin Son, Soohyun Park and Joongheon Kim
Department of Electrical and Computer Engineering, Korea University, Seoul, Republic of Korea
E-mails: {lydiasb,joongheon}@korea.ac.kr, soohyun.park@sookmyung.ac.kr

***Abstract*—Recent advances in multi-modal large language models (LLMs) have demonstrated strong reasoning capabilities in 2D vision-language tasks, motivating the extension of these capabilities to 3D domains. Point clouds, as a compact and precise representation of 3D geometry, have become a key modality for enabling high-level spatial understanding. However, the irregular and unordered structure of point clouds introduces challenges for efficient processing and cross-modal integration. Recent research addressing these challenges can be categorized into four main paradigms: (i) direct point cloud encoding, (ii) point cloud–based multi-modal alignment, (iii) point cloud and semantic information fusion for upsampling, and (iv) multi-view image–based 3D processing. Representative methods within each paradigm employ distinct architectural choices to balance geometric fidelity, computational efficiency, and multi-modal capability. This taxonomy provides a structured perspective on the evolving landscape of point cloud–LLM integration, highlighting design trade-offs and offering insights into potential future directions in multi-modal 3D scene understanding.**

***Index Terms*—Point Cloud, Large Language Model, Multi-modal Large Language Model, LLM**

I. Introduction

The ability to understand and reason about 3D environments is essential for applications such as autonomous driving, robotic manipulation, and immersive mixed reality. Among various 3D representations, point clouds have become a dominant choice due to their direct acquisition from sensors like LiDAR and RGB-D cameras, as well as their precise preservation of spatial geometry. However, the irregular and unordered nature of point clouds makes them challenging to process efficiently, particularly when integrating with large language models (LLMs) for high-level reasoning. LLMs have recently achieved remarkable progress in language understanding and reasoning across diverse domains [1]–[3]. Building on this progress, multi-modal LLMs in 2D vision-language tasks have further advanced, sparking growing interest in extending such capabilities to the 3D domain, as shown in Fig. 1. Unlike 2D images, point clouds lack a regular grid structure. This absence complicates the direct use of standard convolutional or transformer architectures. Moreover, 3D understanding often benefits from multi-modal information, including text, images, audio, and even video. These complementary modalities help capture both the geometric and semantic aspects of a scene. As a result, recent research has explored diverse strategies to bridge point cloud representations with LLMs.

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2025-00561377). *(Corresponding author: Joongheon Kim)*

Current methods can be broadly grouped into four major paradigms. Direct point cloud encoding processes raw XYZ and RGB features with specialized 3D encoders to preserve geometric fidelity and produce token sequences compatible with LLMs [4]–[6]. Point cloud–based multi-modal alignment embeds 3D data into a shared semantic space with other modalities through contrastive learning, enabling cross-modal retrieval and reasoning [7]. Point cloud and semantic information fusion for upsampling integrates LLM-generated semantic cues with geometric features to enhance sparse point clouds into high-resolution reconstructions [8]. Multi-view image–based 3D processing bypasses direct point cloud handling by constructing 3D-aware features from multi-view RGB and depth images, leveraging pretrained 2D LMM architectures [9]. Each paradigm reflects a different design choice in balancing geometric accuracy, computational efficiency, and multi-modal capability. Direct encoding excels at detailed spatial reasoning but faces scalability challenges in large-scale scenes [4]–[6]. Multi-modal alignment supports flexible cross-domain queries but may lose fine-grained geometric detail [7]. Semantic-guided upsampling improves reconstruction quality but depends on reliable semantic generation [8]. Multi-view image approaches benefit from existing 2D infrastructure yet require high-quality multi-view captures [9].

This trend in integrating point clouds with LLMs highlights the convergence of geometric processing and multi-modal reasoning. Understanding these categories and their trade-offs is crucial for guiding future work in 3D scene understanding that combines spatial precision with semantic richness.

II. 3D Multi-modal Large Language Model

A. Direct Point Cloud Encoding Approaches

Direct point cloud encoding processes raw 3D data, consisting of XYZ coordinates and RGB color values, directly with a 3D encoder. This avoids intermediate representations and preserves both geometric detail and visual cues. The main challenge in this approach is handling large point sets, where efficient sampling and token reduction are required for scalability. LL3DA applies this method to point clouds from datasets such as ScanNet and ARKitScenes [4]. It uses a PointNet++ backbone, preceded by preprocessing steps that normalize coordinates, apply farthest point sampling (FPS) for uniform coverage, and add sinusoidal positional encoding to

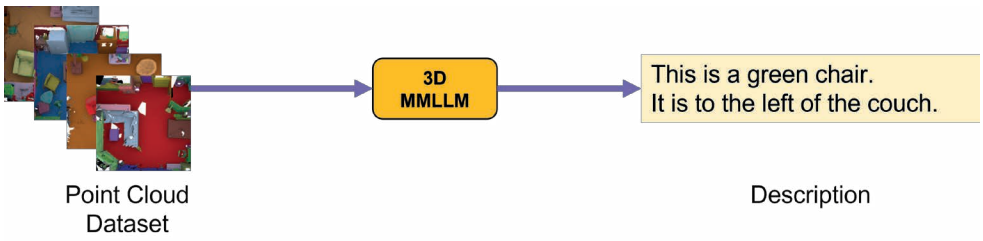

Fig. 1: 3D MMLLM pipeline: multi-view 3D inputs yield grounded scene captions describing attributes and spatial relations.

preserve spatial structure. The encoder extracts multi-scale local and global features, which are then transformed into LLM-compatible tokens. These tokens, paired with textual inputs, are processed by a multi-modal Transformer for tasks like 3D visual question answering, object grounding, and scene summarization. 3UR-LLM follows the same paradigm but adopts a Sparse Convolution architecture for large-scale scenes [5]. The point cloud is voxelized into a sparse grid, with RGB and XYZ attributes averaged within each voxel. A MinkowskiNet-based sparse convolutional encoder processes the non-empty voxels, producing features augmented with positional embeddings. Voxel pooling then reduces token counts to fit LLM context limits, enabling efficient processing of dense, complex environments. JM3D-LLM also uses a sparse convolutional backbone for direct XYZ+RGB encoding, but integrates multi-modal cues after feature extraction [6]. Following voxelization and FPS, the 3D features are fused with multi-view image embeddings from a CLIP-based visual encoder and text embeddings from an LLM text encoder via cross-attention. This fusion enriches the geometric representation with visual texture and semantic information, supporting reasoning that combines spatial configuration with linguistic context. Overall, direct encoding approaches share core design elements such as spatially aware preprocessing, sampling or voxel pooling, and LLM-compatible token generation. They differ mainly in the choice of encoder backbone and in whether additional multi-modal fusion is applied, leading to varying strengths in scalability, geometric fidelity, and semantic richness.

B. Point Cloud-based multi-modal Alignment Approaches

Point cloud-based multi-modal alignment encodes 3D data into a shared embedding space with other modalities such as images, text, audio, and video. The objective is to make embeddings of the same object or scene close in latent space through contrastive learning, while pushing apart those of unrelated samples. Point-Bind follows this approach using a Point-BERT-based Transformer encoder to process point clouds [7]. The model extracts high-level geometric features, projects them into a fixed-dimensional space, and aligns them with embeddings from image, text, audio, and video encoders. Training uses paired multi-modal data with a contrastive loss to enforce cross-modal consistency. This alignment allows flexible cross-modal retrieval and reasoning. For example, the model can retrieve a 3D shape from a text prompt or match a 3D object to its audio description. By grounding 3D geometry in a unified semantic space, such methods enable richer interaction between 3D data and diverse media.

C. Point Cloud and Semantic Information Fusion for Upsampling

Upsampling approaches enhance sparse point clouds by combining geometric features with semantic cues from LLMs. This allows reconstruction of high-resolution geometry with both structural detail and semantic coherence. PULLM first extracts multi-scale geometric features using hierarchical grouping and feature aggregation [8]. A PointLLM module generates a textual description of the scene, encoded into a semantic embedding. The geometric and semantic features are merged in the feature-aware translator (FAT), aligning the modalities in a shared space. An Adaptive B-spline convolution then refines the fused features, preserving sharp edges and smooth surfaces during upsampling. The result is a dense, high-quality point cloud that reflects precise geometry and meaningful scene semantics.

D. Multi-view Image-based 3D Processing Approaches

Multi-view image-based methods build 3D representations from RGB images and depth maps taken from multiple viewpoints, without directly processing raw point clouds. This allows the reuse of pretrained 2D Large multi-modal Model (LMM) architectures while adding spatial context. LLaVA-3D extracts image patches from multi-view RGB inputs and assigns each patch a 3D coordinate using camera pose and depth [9]. These coordinates become 3D positional embeddings that are combined with visual features to create 3D-aware tokens. To reduce redundancy, voxelization pooling or FPS is applied before feeding tokens into a multi-modal Transformer. A Grounding Decoder then predicts 3D bounding boxes for object localization and grounding. By embedding spatial coordinates into 2D visual tokens, this approach extends vision-language models to spatial reasoning tasks such as 3D object grounding and scene understanding, without requiring direct point cloud encoding.

III. Concluding Remarks

The integration of point clouds with LLMs has emerged as a promising direction for advancing 3D scene understanding. Existing methods can be classified into four paradigms: direct point cloud encoding, multi-modal alignment, semantic-guided upsampling, and multi-view image–based processing. Each paradigm adopts a different strategy to balance geometric detail, scalability, and multi-modal reasoning. Direct encoding

methods maintain high geometric fidelity but face limitations in processing large-scale scenes. Multi-modal alignment enables flexible cross-domain retrieval while potentially sacrificing fine-grained spatial accuracy. Semantic-guided upsampling improves reconstruction quality by incorporating language-driven cues, whereas multi-view image approaches leverage pretrained 2D architectures but depend on high-quality multi-view input. No single paradigm fully resolves all challenges in 3D–LLM integration. Future advancements are likely to benefit from hybrid designs that combine geometric precision, semantic richness, and computational efficiency. The development of unified benchmarks and large-scale multi-modal 3D datasets will be essential for fair evaluation and progress. A clear understanding of the strengths and limitations of current approaches can inform the design of next-generation frameworks capable of precise, semantically rich, and efficient 3D reasoning.

REFERENCES

[1] H. Ahn, S. Oh, G. S. Kim, S. Jung, S. Park, and J. Kim, "Hallucination-aware generative pretrained transformer for cooperative aerial mobility control," *arXiv preprint arXiv:2504.10831*, 2025.

[2] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang *et al.*, "A survey on evaluation of large language models," *ACM transactions on intelligent systems and technology*, vol. 15, no. 3, pp. 1–45, 2024.

[3] E. J. Roh and J. Kim, "Quantum-amplitude embedded adaptation for parameter-efficient fine-tuning in large language models," in *Proc. ACM International Conference on Information and Knowledge Management (CIKM)*, Seoul, Korea, November 2025.

[4] S. Chen, X. Chen, C. Zhang, M. Li, G. Yu, H. Fei, H. Zhu, J. Fan, and T. Chen, "LL3DA: visual interactive instruction tuning for omni-3d understanding, reasoning, and planning," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, (CVPR)*, Seattle, WA, USA, June 2024, pp. 26 418–26 428.

[5] H. Xiong, Y. Zhuge, J. Zhu, L. Zhang, and H. Lu, "3UR-LLM: An end-to-end multimodal large language model for 3d scene understanding," *IEEE Transactions on Multimedia*, vol. 27, pp. 2899–2911, May 2025.

[6] J. Ji, H. Wang, C. Wu, Y. Ma, X. Sun, and R. Ji, "JM3D & JM3D-LLM: Elevating 3d representation with joint multi-modal cues," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 4, pp. 2475–2492, December 2025.

[7] Z. Guo, R. Zhang, X. Zhu, Y. Tang, X. Ma, J. Han, K. Chen, P. Gao, X. Li, H. Li *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.

[8] Z. Zhang, R. Liu, X. Liu, Y. Zhu, Y. Yang, C. Wang, and J. Zhang, "PULLM: A multimodal framework for enhanced 3d point cloud upsampling using large language models," in *Proc. of the ACM/SIGAPP Symposium on Applied Computing, (SAC)*, Catania International Airport, Catania, Italy, April 2025, pp. 1223–1230.

[9] C. Zhu, T. Wang, W. Zhang, J. Pang, and X. Liu, "Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness," *arXiv preprint arXiv:2409.18125*, 2024.