LEMON: A Unified and Scalable 3D Multimodal Model

Understanding 3D world is fundamental for embodied agents, enabling interaction, manipulation, and navigation in the physical world. While large multimodal models (LMMs) have achieved impressive progress in 2D vision-language domains — demonstrated by Flamingo (Alayrac et al., 2022), GPT-4V (OpenAI, 2023) and many open-sourced models (Xiong et al., 2024, Yang et al., 2025, Wang et al., 2025) — scaling to 3D data remains challenging due to point clouds' irregular structure, sparsity, and high dimensionality. Despite robust 3D understanding being crucial for robotics (Fang et al., 2023, Zhu et al., 2024, Qi et al., 2025), AR/VR systems, and spatial AI (Chen et al., 2024, Cheng et al., 2024, Zheng et al., 2024), and the emergence of 3D foundation models like Point-BERT (Yu et al., 2022) and ULIP (Xue et al., 2022), current efforts fall short of general-purpose 3D understanding analogous to 2D LMMs. Existing 3D LMMs typically use modular designs with separate 3D and language encoders (Liu et al., 2023, Zhou et al., 2023), but face fundamental challenges: (1) limited 3D pretraining datasets with narrow objectives, (2) constrained 3D data scale versus billions of 2D images, (3) architectural imbalance creating representational bottlenecks, and (4) frozen encoders preventing end-to-end optimization and generalization to novel 3D structures.

We propose **LEMON**, a unified transformer architecture that directly embeds both 3D geometry and natural language into a shared token space. Rather than relying on separate encoders, **LEMON** treats 3D point cloud patches and language tokens as a unified sequence for joint processing. Each 3D patch is mapped to the language embedding space via a learnable linear projector, and structured using modality-specific and spatial separator tokens. This design allows cohesive processing of spatial and linguistic information while eliminating modality-specific encoders and cross-modal alignment mechanisms, improving scalability of 3D multimodal models. To our knowledge, **LEMON** is the first architecture that unifies point cloud and language processing at the token level within a single transformer for general-purpose 3D reasoning.

To address the challenges of sparse and irregular 3D data, we introduce a dynamic patchification and tokenization strategy. Point clouds are partitioned into patches via a recursive 3D spatial scheme, ensuring uniform patch sizes while preserving geometric structure. Specialized separator tokens encode spatial hierarchy, allowing transformers to operate over structured sequences. We design a three-stage training curriculum: (1) object recognition using large-scale 3D object data; (2) object-level captioning and grounding; and (3) scene-level spatial question answering. This curriculum supports progressive scaling, transitioning from object-level to complex scene reasoning.

We evaluate **LEMON** across a suite of 3D multimodal tasks, including generative object classification, caption generation, embodied interaction QA, and spatial scene understanding. Our model consistently outperforms prior state-of-the-art baselines in each domain, while exhibiting more favorable scaling behavior as model and data size increase. **LEMON**'s unified architecture reduces parameter redundancy, simplifies the training pipeline, and enables joint spatial-linguistic reasoning, paving the way toward general-purpose 3D multimodal systems for embodied AI, robotics, and beyond.