ARTICLE
https://doi.org/10.26599/AIR.2026.9150017
CAAI Artificial Intelligence Research

L3DM: An Open-Source Large 3D Multi-Modality Model on MindSpore with Text-Geometry Alignment

Shangkun Sun[1,†], Junhong Lin[1,†], Yuqi Ye[1,†], Ge Li[1], Wei Gao[1,*]

ABSTRACT

We present an efficient and scalable MindSpore-based framework for three-dimensional (3D) scene understanding, namely Large 3D Multi-Modality Model (L3DM), with a focus on text-geometry alignment through MindSpore-native multimodal feature integration and key–value (KV) cache adaptation. Our work delivers three key contributions to the community. First, we provide a dedicated MindSpore performance optimization for multi-modal feature inputs and KV cache, resolving prior bottlenecks and enabling stable, high-performance operation where it was previously challenging. Second, we explore zero-shot Chain-of-Thought reasoning for 3D multi-view understanding, providing valuable insights into its limitations for complex spatial reasoning tasks. Third and foremost, we offer the first complete, open-source, and scalable 3D foundation model implementation on the MindSpore platform. This contribution fills a critical gap by providing the community with a fully functional training and inference codebase, pre-trained weights, and pipelines for a multimodal 3D large model—resources that were entirely absent before. Overall, the proposed L3DM method establishes a new, stable, and accurate benchmark solution for 3D understanding on MindSpore, with impressive empirical validation on the ScanQA benchmark. Code is available at https://openi.pcl.ac.cn/OpenAIRobotics/L3DM.

KEYWORDS

3D Scene Understanding, Chain-of-Thought, Large Language Models, Multi-Modality Learning and Reasoning, MindSpore Implementation

Just Accepted

1 Introduction

The burgeoning field of multimodal artificial intelligence (AI) increasingly demands models capable of sophisticated and unified understanding that seamlessly bridges the visual and linguistic domains [1, 2, 3, 4, 5]. Beyond simple captioning or retrieval, the frontier is advancing toward systems that can engage in deep, contextual reasoning—jointly perceiving rich visual content, interpreting it through the lens of natural language, and generating responses that are coherent, actionable, and aligned with nuanced human intentions and goals. While remarkable progress has been made in aligning two-dimensional (2D) imagery or video with text, enabling applications such as visual question answering and detailed description generation, the comprehension of and interaction with complex, three-dimensional (3D) environments through natural language remains a significant and open challenge. This challenge arises because 3D scenes encapsulate far richer structural and relational information than their 2D projections. They introduce fundamental complexities such as inherent spatial continuity, multi-view consistency, long-range dependencies, and the need to reason about occlusion and geometry from partial observations. Large language models (LLMs) for 3D scenes [6, 7, 8] have emerged as a promising paradigm to address this gap, aiming to construct foundational models specifically architected for scalable 3D scene understanding and cross-modal reasoning. At their core, these models seek to achieve precise and semantically grounded text-to-geometry alignment. This involves building representations that allow a model to interpret intricate 3D structures, infer spatial relationships between objects, and understand scenes holistically in direct response to natural language queries or instructions. By strengthening this core alignment, the field aims to unlock transformative capabilities in several next-generation applications. This includes advancing embodied AI, where agents must navigate and manipulate the physical world; enhancing virtual and augmented reality with intuitive, language-driven interaction; and developing intelligent systems capable of reliable, high-level reasoning and planning in real-

1 Guangdong Provincial Key Laboratory of Ultra High Definition Immersive Media Technology, School of Electronic and Computer Engineering, Peking University, Shenzhen, 518055, China.
† Equal contribution: Shangkun Sun, Junhong Lin, and Yuqi Ye contributed equally to this work.
* Corresponding author: Wei Gao (e-mail: gaowei@pku.edu.cn).
This work was supported by National Science and Technology Major Project (2024ZD01NL00101), National Natural Science Foundation of China (62671011), Guangdong Provincial Key Laboratory of Ultra High Definition Immersive Media Technology (2024B1212010006), Shenzhen Science and Technology Program (SYSPG20241211173440004), and sponsored by CAAI-MindSpore Open Fund, developed on OpenI Community (CAAIXSJLJJ-2023-MindSpore-07).

world 3D environments. Ultimately, the goal is to move from systems that merely see 3D data to those that can understand and reason about it in a language-embedded, interactive loop.

Effectively aligning textual descriptions with 3D geometric data presents notable and distinct challenges when compared to modalities like 2D images or video. Firstly, the intrinsic gap between dense, continuous 3D representations—such as point clouds, meshes, or multi-view projections—and discrete linguistic symbols tends to be wider and more structurally complex. This divergence extends beyond semantic content to fundamental differences in data organization and statistical distributions, making the learning of a shared embedding space more difficult. Secondly, 3D data is often characterized by inherent sparsity, frequent occlusions, and strong viewpoint dependence. These properties complicate consistent feature extraction, robust cross-modal matching, and integrated scene understanding, even for advanced neural architectures. Such difficulties become more pronounced in cluttered environments containing numerous small or visually similar objects, or when the available observations are partial. Thirdly, in the 3D domain, the semantic disconnect between features captured by geometric encoders—focused on spatial layouts, primitive shapes, and local relationships—and those encoded by language models—oriented toward abstract concepts and syntactic structures—appears especially acute. This misalignment can hinder the deep fusion necessary for fine-grained reasoning, often resulting in a mismatch between high-level textual descriptions and low-level geometric cues, thereby limiting a model's capacity to reliably ground language in the underlying 3D scene structure.

Previous approaches to 3D-text alignment have largely drawn inspiration from techniques successful in 2D vision-language tasks [9, 10, 11, 12, 13], often by extending established multimodal architectures. Common strategies [6, 7, 8] involve employing cross-attention mechanisms [14] or adapter modules such as querying transformers (Q-Formers) [15] to project a limited set of learnable features from one modality into the representation space of the other. While these methods enable a shared representation for downstream tasks and can capture broad correspondences between language and geometry, they may face difficulties in addressing the specific complexities of 3D data. In particular, they may not fully address the inherent semantic gap between dense geometric features and discrete linguistic concepts, which can sometimes lead to suboptimal fusion and reduced robustness in cluttered or highly diverse scenes. Additionally, such approaches often lack mechanisms to explicitly enhance a model's intrinsic, multi-view spatial reasoning capability without relying heavily on extensive task-specific fine-tuning, detailed annotations, or manually crafted supervision. Consequently, scaling these methods to large and varied 3D datasets can present challenges, and their performance may encounter limitations in more demanding geometric understanding scenarios.

To address these challenges, we present the Large 3D Multi-Modality Model (L3DM), a scalable framework based on MindSpore for 3D-language understanding, designed to better align with the requirements of contemporary multimodal applications. MindSpore's strengths in static graph compilation and automated parallelization are particularly beneficial for processing computationally demanding 3D data, allowing for more efficient large-scale training under common hardware constraints. During development, we observed that existing support for multimodal input pipelines within the MindSpore ecosystem could be further refined, especially in 3D settings where direct adoption of certain model designs led to noticeable performance gaps compared to implementations in other frameworks such as PyTorch. In response, we undertook a series of incremental adjustments—first diagnosing specific bottlenecks in multimodal embedding and inference-time key–value (KV) cache handling, and then introducing modest but targeted optimizations in these areas. These modifications, though relatively lightweight, helped narrow the performance discrepancy and improve the stability and accuracy of 3D scene representation during inference. Alongside these engineering adaptations, we also explored whether zero-shot chain-of-thought (CoT) [16] reasoning could be effectively integrated to support step-wise reasoning in complex 3D environments without task-specific fine-tuning. Our preliminary experiments, however, suggested that its applicability in L3DM remains limited—likely because the structured, spatially-grounded reasoning required in 3D contexts may not be fully captured by general-purpose CoT prompting without more explicit geometric supervision or tailored reasoning modules. Collectively, these efforts aim to offer a more robust, reproducible, and scalable open-source implementation for 3D-language modeling within the MindSpore environment. We view L3DM not as a definitive solution, but as an adaptable foundation that may facilitate future research and practical applications—such as embodied navigation, interactive scene editing, and detailed 3D visual question answering—while also contributing modestly to the growing toolkit for multimodal geometric reasoning.

The contributions of L3DM can be divided into three levels. First, the framework-level contribution concerns the adaptation of MindSpore for continuous multimodal feature input and KV cache handling in hybrid 3D-text sequences. Second, the model-level design integrates a frozen 3D encoder, a visual expert, and an LLM for text-geometry alignment. The visual expert follows the established design of the embodied multimodal generalist agent LEO [17], while low-rank adaptation (LoRA) is employed as a standard parameter-efficient adaptation strategy; neither is claimed as a novel component of L3DM. Third, the implementation-level contribution provides reproducible training and inference pipelines for 3D multi-modal modeling on MindSpore.

In summary, L3DM introduces a comprehensive MindSpore-based model for geometrically-grounded multimodal intelligence, delivering three core contributions that advance both methodological innovation and practical applicability. Each contribution is designed not only to push the boundaries of 3D scene understanding but also to actively support and enrich the open-source research ecosystem:

- Refinements to multi-modal input processing in 3D scenarios and optimizations to KV cache, developed as an enhancement to existing MindSpore functionalities. This part of our work attempts to address a practical limitation observed in prior implementations within the MindSpore ecosystem—namely, persistent accuracy issues in 3D scene representation during inference. By refining the multi-modal input processing design and correspondingly optimizing the KV cache mechanism,

Under Review

we seek to improve the robustness and precision of the computational graph in 3D reasoning settings. These adjustments, while built upon available MindSpore operators, help stabilize performance and correct earlier inaccuracies, leading to more reliable and consistent outcomes. We are pleased to share these modest improvements with the community, in hopes that they may contribute to more solid and reproducible results for others working at the intersection of 3D understanding and multimodal learning.

- An exploratory study on zero-shot Chain-of-Thought reasoning for multi-view 3D scene understanding. In this portion of our work, we gently examine the applicability of zero-shot CoT in geometrically complex settings. Our observations suggest that its effectiveness within L3DM appears limited— an outcome that may reflect the inherent structural reasoning demands of 3D scenes, which might not be fully met by general-purpose zero-shot CoT methods. Rather than presenting a definitive critique, we offer these preliminary findings as a shared reference, in the hope that they might inform future designs of more tailored supervision or reasoning strategies for 3D multimodal models.

- An open-source MindSpore implementation of a scalable 3D foundational model. With the goal of supporting reproducible research, we provide a publicly available implementation built on the MindSpore framework. By leveraging its native support for distributed training, we aim to offer a stable and efficient codebase that facilitates larger-scale 3D pretraining and multimodal alignment experiments. We hope that releasing this implementation—along with detailed documentation and pre-trained weights—might lower the barrier to entry for other researchers and encourage further community efforts in scalable 3D multimodal learning.

2 Related Work

Large Language Models (LLMs). Recent advancements in LLMs, such as Vicuna and Large Language Model Meta AI (LLaMA), have demonstrated remarkable capabilities in natural language understanding and generation through extensive pre-training and instruction fine-tuning. However, these models primarily rely on static text or 2D visual inputs, limiting their ability to interact with physical 3D environments. Emerging approaches integrate LLMs with geometric modeling techniques, such as Conformal Geometric Algebra (CGA), to enable controllable 3D scene editing and object repositioning.

Vision-and-Language Question Answering Tasks. Multimodal large language models (MM-LLMs) have achieved strong performance in vision-and-language question answering (QA) tasks, yet challenges persist in geometric reasoning and precise arithmetic calculations. Methods such as modal-enhanced semantic modeling (MESM) [18] improve video–text alignment by enhancing semantic representations at the frame–word and segment–sentence levels. Additionally, embodied AI frameworks combined with LLMs have advanced robustness analysis for QA systems in dynamic environments.

Open-sourced Frameworks. MindSpore is a full-stack, open-source AI computing framework developed by Huawei. It is designed to support end-to-end model development, training, and deployment across diverse hardware platforms, including Ascend AI processors, graphics processing units (GPUs), and central processing units (CPUs). A key characteristic of MindSpore is its native support for automatic differentiation and parallel execution, which simplifies the implementation of complex models. MindSpore prioritizes static graph optimization and automated parallelism for efficient large-scale distributed training, while PyTorch emphasizes a dynamic, define-by-run paradigm favored for its flexibility and debugging ease during research prototyping. Its static graph optimization (in Graph mode) enhances computational efficiency and memory usage, making it particularly suitable for large-scale distributed training scenarios common in modern foundational model research.

2.1 Large Language Models

In recent years, the emergence and rapid advancement of LLMs [19, 20, 21, 22] have demonstrated exceptional versatility and remarkable performance across a diverse and expanding spectrum of natural language processing tasks, encompassing not only fundamental capabilities such as text generation [23] and complex multistep reasoning [24] but also extending to a wide variety of other critical applications. Beyond these core inherent abilities, LLMs have consistently exhibited strong and robust performance in numerous other tasks, including but not limited to text summarization, engaging in multi-turn dialogue, machine translation between different languages, and sophisticated semantic parsing, where they demonstrate a notable capacity to flexibly adapt to varying input formats and diverse output stylistic requirements even with minimal or highly limited task-specific customization.

These powerful models fundamentally leverage massive-scale pre-training conducted on immensely diverse and extensive textual corpora, primarily employing sophisticated self-supervised training methodologies [25] to effectively internalize complex linguistic patterns, deep semantic regularities, and broad, comprehensive world knowledge. Through extensive exposure to heterogeneous data sources spanning multiple domains and genres, they progressively acquire a robust, generalizable contextual understanding and demonstrate an impressive ability to generalize to varied and previously unseen domains, even in scenarios where explicit supervision is significantly limited or entirely absent during the adaptation phase.

A critical and influential breakthrough in this rapidly evolving domain has been the development and widespread adoption of instruction tuning, where pioneering methods such as InstructGPT [26] systematically refine LLMs' alignment with nuanced human intentions by carefully fine-tuning on curated, high-quality datasets composed of task-specific instructions and exemplary responses. This meticulous process encourages models to follow natural language commands more faithfully and consistently, substantially reduce the generation of harmful, biased, or irrelevant responses, and more accurately capture subtle user preferences and intents during interactive sessions. Such advanced techniques significantly enhance the model's generalization capabilities to entirely unseen and novel

Just Accepted

problems, thereby enabling modern models to successfully tackle complex, often multi-modal challenges that lie beyond their original pre-training scope and to effectively handle open-ended, creative instructions that substantially differ from the specific prompts encountered during their initial training phases.

Although contemporary state-of-the-art models such as Generative Pre-trained Transformer 4 (GPT-4) [21] excel in a wide array of demanding cognitive tasks, including abstract reasoning, strategic planning, and nuanced language understanding, they inherently lack explicit, built-in mechanisms to perceive, interpret, and understand complex 3D environments, which represents a critical and substantial gap for numerous real-world applications in fields such as robotics [27], augmented reality [28], and embodied AI [29]. Hence, our work strategically builds upon these powerful foundation language models to directly address a key fundamental limitation of conventional LLMs: their inherent reliance on predominantly 2D-centric representations that systematically neglect essential spatial, physical, and geometric reasoning aspects, and we actively seek to extend their core capabilities systematically toward richer, geometry-aware multimodal understanding.

2.2 3D Scenes Question Answering

The task of 3D question answering (3D-QA) [30] represents a significant evolution beyond conventional visual question answering (VQA) [31], demanding that intelligent systems develop a holistic understanding of three-dimensional environments through natural language interfaces. Unlike traditional VQA frameworks that process static two-dimensional projections, 3D-QA necessitates sophisticated volumetric reasoning capabilities, requiring models to interpret spatial configurations, geometric relationships, and topological arrangements of objects within a continuous 3D coordinate space. This paradigm shift introduces fundamental computational challenges related to depth perception, occlusion handling, and the inference of spatial hierarchies that extend far beyond the planar constraints inherent in single-image analysis.

Whereas conventional 2D VQA systems leverage established computer vision techniques [32]–including region proposal networks, cross-modal attention mechanisms, and convolutional feature extraction—to establish correspondences between linguistic concepts and visual elements, these approaches fundamentally lack the capacity to model physical space volumetrically. Specifically, methodologies such as the mask region-based convolutional neural network (Mask R-CNN) [33] for instance segmentation and relationship detection operate effectively within the limited context of image planes but fail to capture the metric depth, surface normals, and spatial occupancy patterns that characterize real-world 3D scenes. Consequently, 3D-QA architectures must integrate heterogeneous data streams from multiple sensor modalities, aggregating information across diverse viewpoint configurations and reconciling partial observations into coherent volumetric representations that support robust spatial reasoning.

The ScanQA dataset [34] exemplifies these heightened requirements by explicitly evaluating systems' capacity for nuanced spatial comprehension within indoor environments reconstructed from red, green, blue, and depth (RGB-D) data. Unlike earlier VQA 360 datasets that primarily address panoramic image understanding while disregarding critical depth cues and volumetric structural information, ScanQA demands precise reasoning about object layouts, spatial extents, and geometric constraints inherent in physical spaces. Furthermore, the task formulation differs substantively from related 3D vision-language tasks such as ScanRefer [8], which focuses primarily on localizing specific objects through free-form descriptive text by predicting axis-aligned 3D bounding boxes. While ScanRefer typically involves single-object grounding scenarios, 3D-QA as instantiated in ScanQA introduces significantly greater complexity by enabling multiple object references within individual questions, thereby requiring sophisticated compositional reasoning and multi-entity grounding capabilities.

Additional benchmark datasets have progressively expanded the scope and complexity of 3D-QA research. The Interactive Question Answering Dataset (IQUAD) [35] introduces interactive querying paradigms within the Allen Institute for AI's The House Of inteRactions (AI2-THOR) simulated environment, necessitating that models resolve ambiguous questions through active object manipulation and state changes rather than passive observation. Complementing this direction, situated question answering in 3D scenes (SQA3D) [36] advances the field by integrating situated reasoning with embodied scene understanding, establishing connections between linguistic queries, navigational trajectories, and agent-centric spatial reasoning within reconstructed 3D environments. These developmental trajectories collectively underscore the field's transition from static, image-centric approaches toward dynamic, spatially aware systems capable of grounding language in volumetric representations.

Despite these advances, substantial challenges persist regarding computational scalability, the robust fusion of multi-view observations under varying lighting and occlusion conditions, and the achievement of fine-grained cross-modal alignment between abstract linguistic concepts and noisy, reconstructed 3D geometry. In the present work, we employ the ScanQA dataset as our primary evaluation framework to rigorously validate our proposed methodology, assessing both the spatial reasoning capabilities and the generalization performance of our model across diverse, realistic indoor scenarios.

2.3 3D Vision-Language Model

With the development of LLMs, 3D vision-language (3D-VL) models[37] have gradually become a research hotspot and started to address question-answering tasks in 3D scenes. These models aim to extend the expressive power of language models by coupling them with 3D perception modules that can interpret point clouds, meshes, and other geometric representations, thereby grounding linguistic descriptions in physical structure. Recent advancements aim to unify 3D-VL understanding through promptable frameworks, such as 3D-LLM[7] and LEO[17], which attempt to bridge this divide by integrating point-cloud processing and geometric algebra into language frameworks. By designing architectures that accept both textual inputs and 3D scene encodings, these methods seek to enable flexible instruction following, interactive scene exploration, and embodied reasoning in complex environments.

Just Accepted

The 3D vision and text alignment model (3D-VisTA)[6] and 3D vision-language pre-training (3DVLP)[38] introduce pre-trained transformers that align 3D point clouds with text embeddings, while SceneVerse[39] scales this approach to support multi-modal instructions in large-scale 3D environments with diverse scene layouts and object categories. Single-stage 3D visual grounding via referred point progressive selection (3D-SPS)[40] leverages cross-modal attention mechanisms to iteratively refine object localization based on language guidance, progressively narrowing down candidate regions until a precise grounding is achieved. Others, such as ViL3DRel[41], explicitly encode spatial relationships (e.g., distances, orientations) into their neural networks to better capture relational information that is crucial for geometric reasoning and object interaction. More recently, 3UR-LLM [42] presents an end-to-end 3D multimodal large language model that directly processes 3D point clouds and employs a dedicated 3D compressor to integrate spatial and textual information into compact tokens. In contrast, L3DM does not introduce a new 3D compression architecture, but primarily focuses on continuous 3D feature integration and multimodal execution within the MindSpore ecosystem. Beyond 3D scene understanding, Context-Infused Trajectories [43] highlights the importance of contextual information preservation for multimodal reasoning in video object segmentation, providing a complementary perspective on maintaining modality-specific context during reasoning. However, many of these approaches rely heavily on task-specific design choices or complex optimization strategies, limiting their generalization to broader 3D-VL scenarios and making them difficult to scale or adapt to new tasks without significant engineering effort. Besides, these approaches often struggle with fine-grained object referencing and cross-modal alignment, especially when interpreting spatial relations or physics-based queries that require consistent reasoning across views and time. Different from these architecture-centric approaches, L3DM primarily focuses on MindSpore-native multimodal execution, including continuous 3D feature ingestion and multimodal-prefix-aware KV cache handling, while also exploring zero-shot reasoning for 3D scene understanding.

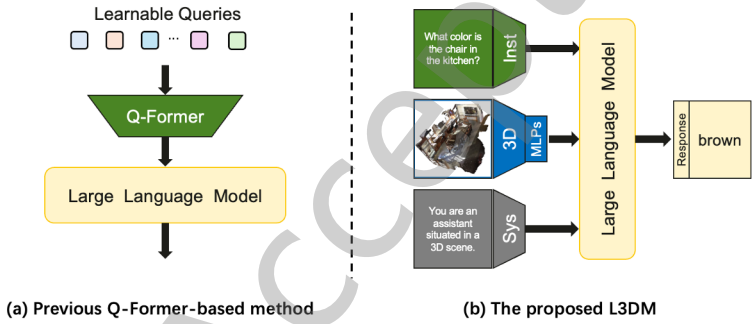


Fig. 1 Overview of the proposed L3DM. Diagram abbreviations: instruction (Inst), system prompt (Sys), and multilayer perceptrons (MLPs).

3 Method

3.1 Overview

We present L3DM, a MindSpore-based framework for 3D scene understanding and reasoning that integrates a pre-trained 3D encoder with a large language model. The core objective is to achieve effective alignment between dense, high-dimensional 3D geometric features and the discrete token space of an LLM, enabling the model to perform complex multimodal tasks such as 3D question answering and scene description. Our framework builds upon and extends prior work in 3D vision-language modeling [17, 6, 7, 8]. As illustrated in Fig. 1, the pipeline begins with processing raw 3D point cloud or mesh data using a specialized, frozen 3D encoder to extract robust and generalizable 3D feature representations. These features are then projected into the LLM's latent embedding space via a lightweight, learnable visual expert module. This design choice is critical: unlike approaches that employ query-based transformers (e.g., Q-Former [15]) for feature compression, our projection does not impose a strict token length limit, thereby minimizing information loss that can occur when summarizing rich 3D contexts into a fixed number of tokens. During training, only the parameters of the visual expert module and a subset of layers within the LLM are optimized, while the 3D encoder remains frozen to preserve its pre-trained geometric knowledge.

In the development process, we identify significant inefficiencies and potential vulnerabilities in the handling of multimodal data and autoregressive inference within the native MindSpore framework. These shortcomings, particularly concerning continuous feature input and KV cache management for multimodal sequences, could lead to substantial accuracy degradation and unstable training. Consequently, we implement targeted improvements to the framework's core components, which are detailed in the following subsection. These enhancements ensure stable and efficient training and inference for 3D large models on MindSpore.

3.2 Optimizing Multimodal Feature Processing and KV Cache

A key technical contribution of this work is the optimization of the underlying deep learning framework to natively support the unique demands of 3D vision-language models, specifically addressing two major limitations inherent in the standard MindSpore framework that previously hindered efficient multimodal development: first, its fundamental design, originally tailored for processing discrete token sequences typical in natural language processing, lacks

Just Accepted

native handling of continuous multi-modal feature vectors, forcing researchers to adopt cumbersome workarounds such as serializing continuous 3D point cloud features into token-like representations or maintaining separate processing branches with manual synchronization, which not only complicates model architecture but also introduces computational overhead and disrupts end-to-end gradient flow; second, its KV cache mechanism, optimized for auto-regressive generation with homogeneous token sequences, fails to accommodate hybrid sequences that interleave text tokens with continuous multi-modal embeddings—a common requirement in tasks like 3D captioning or visual question answering where visual and textual contexts must be jointly attended to—leading to inefficient memory allocation, misaligned cache entries, and redundant computations when handling variable-length multi-modal inputs; by directly optimizing the framework to overcome these limitations, this work enables seamless integration of continuous 3D features with discrete text processing within a unified execution environment, thereby streamlining model development and improving runtime efficiency for 3D vision-language tasks without altering the framework's core application programming interfaces (APIs) or introducing external dependencies.

Multimodal Feature Input Enhancement. Standard transformer implementations in MindSpore expect input_ids composed of vocabulary indices. To enable vision-language models (VLMs) to process the continuous output from our 3D encoder, we extend the **Transformers** library within MindSpore. Specifically, we modify the model's forward pass to accept an additional **multimodal_features** argument. These features are processed through the learnable visual expert to obtain a sequence of feature embeddings. These embeddings are then prepended to the text token embeddings before being fed into the LLM's transformer blocks. This modification allows the model to seamlessly integrate 3D geometric information as a prefix context for language generation.

KV Cache Restructuring for Multimodal Sequences. During autoregressive generation, caching previous keys and values in the KV cache is essential for efficiency. However, the standard KV cache implementation in MindSpore does not correctly handle cases where the initial part of the sequence consists of non-textual feature embeddings. We restructure the cache management logic to account for the distinct nature of these multimodal prefix embeddings. The cache is now properly initialized with the keys and values computed from the multimodal feature sequence, and the subsequent text generation steps correctly attend to this cached multimodal context. This ensures that the attention mechanism has consistent and accurate access to the 3D scene information throughout the generation process, preventing errors and improving inference speed.

Cross-Modal Alignment Module. For projecting the 3D features into the LLM's space, we adopt and implement the effective design from LEO [17]. The visual expert is a compact multilayer perceptron (MLP) with residual connections. Let N denote the number of encoded 3D features, d_{v} the dimension of each feature, and d_{LLM} the LLM embedding dimension. With $\mathbb{R}$ denoting the real numbers, the input feature matrix is $\boldsymbol{V} \in \mathbb{R}^{N \times d_{\mathrm{v}}}$, with one feature per row. The aligned feature matrix $\boldsymbol{H} \in \mathbb{R}^{N \times d_{\mathrm{LLM}}}$ is computed as

$$\boldsymbol{H} = \mathrm{MLP}(\boldsymbol{V}) + \boldsymbol{V}\boldsymbol{W}_{\mathrm{p}},$$

where $\boldsymbol{W}_{\mathrm{p}} \in \mathbb{R}^{d_{\mathrm{v}} \times d_{\mathrm{LLM}}}$ is the learnable linear projection matrix. The MLP maps each row of $\boldsymbol{V}$ to the LLM embedding space, so both terms have the same dimensions as $\boldsymbol{H}$. The resulting feature sequence is ready for the LLM. This residual design facilitates gradient flow and enables dynamic refinement of features, promoting deeper fusion between modalities.

Through these low-level framework optimizations, we provide a robust and reliable MindSpore-based platform for developing and deploying multimodal 3D large models. These improvements are foundational, ensuring that the model's architectural innovations can be realized without being hampered by framework-level constraints.

These components serve different purposes and are not directly interchangeable ablation factors. Multimodal feature input support and KV cache restructuring are framework-level adaptations required for processing continuous 3D embeddings and hybrid 3D-text sequences in MindSpore; removing them would disable the corresponding multimodal-input or autoregressive-caching functionality rather than yield a directly comparable model variant. The visual expert serves as the model-level alignment module and follows the design of LEO, while LoRA is employed as a standard parameter-efficient adaptation strategy. Therefore, the reported results characterize the complete system and should not be attributed to any single component.

3.3 Exploration on Zero-Shot CoT

Motivation and Theoretical Background To augment the reasoning capabilities of L3DM without incurring additional training costs or architectural modifications, we conduct a systematic and comprehensive investigation into the integration of zero-shot CoT prompting [16, 44]. The underlying theoretical premise of this approach is to leverage the inherent in-context learning capabilities and implicit reasoning patterns acquired during the pretraining of large language models. By explicitly guiding the LLM to decompose complex 3D reasoning tasks into intermediate, logical steps before synthesizing a final answer, we aim to achieve multiple objectives: (1) improved accuracy through step-by-step verification and error correction opportunities, (2) enhanced interpretability through transparent and inspectable reasoning traces, and (3) better handling of compositional queries that require integrating multiple sources of information.

The success of zero-shot CoT in textual reasoning domains—such as mathematical problem solving, commonsense reasoning [44], and symbolic manipulation—provides strong motivation for its application to multimodal 3D understanding. In these domains, CoT prompting has been shown to elicit emergent reasoning capabilities that are not explicitly present in standard prompting paradigms. The key insight is that by forcing the model to generate intermediate reasoning steps, we effectively increase the computational depth allocated to the problem and provide additional context for the final answer generation.

However, extending this paradigm to 3D scene understanding introduces unique challenges that distinguish it from

Just Accepted

purely textual reasoning. Unlike symbolic mathematics or textual commonsense reasoning, 3D understanding requires grounding abstract concepts in continuous geometric spaces, reasoning about spatial relationships under projection and viewpoint transformations, and integrating information across multiple sensory modalities. These requirements raise fundamental questions about the compatibility of linguistic reasoning chains with geometric perception tasks.

Prompting Strategy Design. We design a comprehensive suite of explicit prompting templates that reformulate user queries into structured reasoning workflows. Rather than presenting the model with direct instructions such as *"Describe the scene"*, we engineer prompts that enforce systematic analysis:

> *"Let's think step by step. First, identify the major objects present in the 3D scene and estimate their semantic categories. Second, analyze and describe the spatial relationships between these objects, including relative distances, orientations, and support relations. Third, examine the overall scene layout and architectural context. Finally, synthesize a coherent scene description based on these structured observations."*

These instructional prompts are concatenated with the original user query, and the resulting composite text undergoes tokenization before being fed into the LLM alongside the encoded 3D features. The 3D features are extracted from our point cloud encoder and projected into the LLM's embedding space to enable multimodal fusion.

Quantitative Evaluation and Empirical Findings. Contrary to our initial expectations, extensive quantitative evaluation reveals that this zero-shot CoT approach fails to deliver consistent performance improvements across standard 3D understanding benchmarks. As presented in Table 2, we observe marginal or negative impacts on key evaluation metrics when comparing the CoT-enhanced variant against our baseline L3DM model. Generic zero-shot CoT primarily decomposes reasoning at the language level but does not directly improve perceptual grounding in continuous 3D geometry. If the model initially misidentifies an object or incorrectly estimates a spatial relation, the generated reasoning chain may propagate or reinforce this error. Such limitations are likely to be more pronounced for viewpoint-dependent relations, occluded or visually similar objects, object counting, and questions involving multiple spatial relations. These observations suggest that effective 3D reasoning may require geometry-aware intermediate representations rather than language-only prompting.

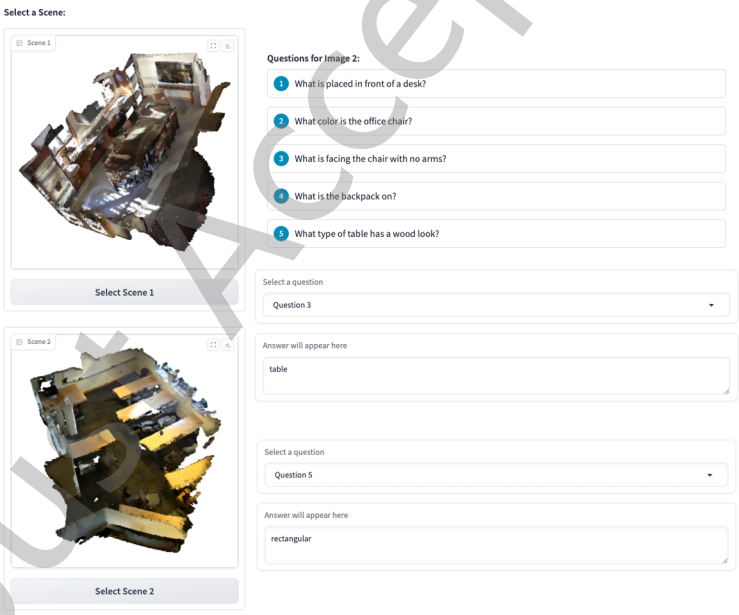


Fig. 2 Demo cases of the open-sourced L3DM implementation.

4 Experiments

4.1 Environmental Setup

We conduct experiments on the ScanQA dataset [34], following 3D-VisTA [6] in utilizing Mask3D-generated object proposals instead of ground-truth object segments for evaluation. We adopt comprehensive metrics including consensus-based image description evaluation (CIDEr), bilingual evaluation understudy using up to four-grams (BLEU-4), metric for evaluation of translation with explicit ordering (METEOR), recall-oriented understudy for gisting evaluation (ROUGE), and top-1 exact-match accuracy (EM@1). For concise notation, we denote CIDEr as C, BLEU-4 as B-4, METEOR as M, and ROUGE as R throughout this paper. For baseline comparisons, we select state-of-the-art task-specific fine-tuned models 3D-VisTA [6] and 3D-LLM [7]. We select ScanQA as the primary benchmark because it directly evaluates generative 3D question answering and text-geometry alignment, which are the main focus of L3DM. The current evaluation should not be interpreted as comprehensive evidence of cross-task generalization. SQA3D and embodied reasoning datasets involve situated reasoning, while ScanRefer and Multi3DRefer emphasize explicit 3D grounding and localization. Evaluation

Just Accepted

Table 1 Quantitative results on ScanQA. Columns denote consensus-based image description evaluation (CIDEr, C), bilingual evaluation understudy using up to four-grams (BLEU-4, B-4), metric for evaluation of translation with explicit ordering (METEOR, M), recall-oriented understudy for gisting evaluation (ROUGE, R), and top-1 exact-match accuracy (EM@1). ScanRefer+MCAN combines ScanRefer with a modular co-attention network (MCAN).

Method	Framework	C	B-4	M	R	EM@1
ScanRefer+MCAN [8]	PyTorch	55.4	7.9	11.5	30.0	18.6
3D-VisTA [6]	PyTorch	69.6	10.4	13.9	35.7	22.4
3D-LLM [7]	PyTorch	69.4	12.0	14.5	35.7	20.5
L3DM	MindSpore	91.3	14.1	18.2	44.1	21.4

Table 2 Comparison of L3DM with and without zero-shot chain-of-thought (CoT) reasoning on ScanQA. Metric abbreviations follow Table 1.

Method	C	B-4	M	R	EM@1
L3DM with zero-shot CoT	90.3	13.4	17.9	44.0	21.5
L3DM without zero-shot CoT	91.3	14.1	18.2	44.1	21.4

on these tasks remains an important direction for future work. All methods are evaluated on ScanQA using the reported standard metrics, and L3DM follows 3D-VisTA in using Mask3D-generated object proposals. However, the compared methods differ in implementation framework, backbone model, and training strategy. Therefore, Table 1 should be interpreted as a benchmark-level comparison rather than a strictly controlled framework comparison.

4.2 Implementation Details

In our implementation, which follows the methodology established by LEO [17], we employ the seven-billion-parameter Vicuna-7B [45] to serve as the foundational backbone language model responsible for processing input token sequences. Vicuna-7B, a large language model fine-tuned from LLaMA on approximately 70,000 user-shared conversations, is chosen for its strong performance in instruction-following tasks and its relatively compact size, which strikes an effective balance between computational efficiency and representational capacity for embodied 3D vision-language reasoning. To preserve the extensive pretrained knowledge inherent within the language model while still enabling effective adaptation to the specific characteristics of 3D vision-language tasks, we incorporate LoRA [46]. Specifically, let k and d denote the input and output dimensions of a linear layer, respectively, and let $\boldsymbol{W}_0 \in \mathbb{R}^{d\times k}$ be its pretrained weight matrix. For an input vector $\boldsymbol{x} \in \mathbb{R}^k$, the adapted output vector is $\boldsymbol{h} \in \mathbb{R}^d$. LoRA represents the weight update $\Delta\boldsymbol{W} \in \mathbb{R}^{d\times k}$ as $\Delta\boldsymbol{W} = \boldsymbol{B}\boldsymbol{A}$, where $\boldsymbol{A} \in \mathbb{R}^{r\times k}$ and $\boldsymbol{B} \in \mathbb{R}^{d\times r}$ are trainable low-rank factors, and $r \ll \min(d, k)$ is the adaptation rank. For the unit update scaling used here, the forward pass is

$$\boldsymbol{h} = \boldsymbol{W}_0\boldsymbol{x} + \Delta\boldsymbol{W}\boldsymbol{x} = (\boldsymbol{W}_0 + \boldsymbol{B}\boldsymbol{A})\boldsymbol{x}.$$

During training, only $\boldsymbol{A}$ and $\boldsymbol{B}$ are updated, while $\boldsymbol{W}_0$ remains frozen, thereby significantly reducing the number of trainable parameters and memory footprint while still allowing the model to adapt to domain-specific multimodal inputs. All computational operations, encompassing both the training and inference phases, are carried out utilizing 16-bit brain floating-point (bfloat16) precision to ensure numerical stability and computational efficiency. For the specific configuration of LoRA, we set the adaptation rank to $r = 16$ and the LoRA scaling parameter to $\alpha = 16$. The multiplier α/r on the low-rank update therefore equals one, as used in the forward expression above; the optimization process for these introduced parameters is subsequently driven by the adaptive moment estimation with decoupled weight decay (AdamW) algorithm, with hyperparameters configured according to LEO [17].

4.3 Quantitative Results

As clearly illustrated in Table 1, the results from our comprehensive evaluation study consistently highlight the distinct performance advantages demonstrated by our proposed L3DM framework when compared against a range of existing baseline methods. The proposed methodology achieves the highest scores among the compared methods on CIDEr, BLEU-4, METEOR, and ROUGE, attaining 91.3, 14.1, 18.2, and 44.1, respectively. For EM@1, L3DM achieves 21.4%, which is higher than ScanRefer+MCAN and 3D-LLM but lower than the 22.4% achieved by 3D-VisTA. The experimental results further reveal that the L3DM framework consistently demonstrates stronger cross-modal generalization capacities when compared to more specialized pipeline architectures, with this advantage being particularly pronounced in the CIDEr and METEOR evaluation metrics. This collective evidence serves to validate the overall effectiveness inherent in our architectural design, which successfully balances the dual objectives of achieving parameter-efficient adaptation while simultaneously maintaining strong multimodal alignment.

Just Accepted

5 Conclusion

In this work, we have presented L3DM, an open-source solution built specifically upon the MindSpore architecture to address the challenges of text-geometry alignment in 3D scene understanding. L3DM, which stands for Large 3D Multi-Modality Model, is designed to bridge the representational gap between discrete linguistic tokens and continuous 3D geometric features extracted from point clouds, thereby enabling more coherent and contextually grounded reasoning about three-dimensional environments. To ensure optimal performance, we tackled the inherent inefficiencies in MindSpore regarding multimodal feature ingestion and KV cache, providing a refined implementation that successfully boosts system throughput; these optimizations, as detailed in the preceding sections, involve enhancing the framework's tensor operators to natively process continuous feature vectors from 3D point clouds without serialization overhead and extending the KV cache mechanism to support hybrid sequences that interleave text tokens with multimodal embeddings, collectively resulting in reduced memory fragmentation, improved cache utilization, and accelerated autoregressive generation during both training and inference. As part of our investigation, we also explored the potential application of zero-shot CoT reasoning for handling multi-view 3D projections, motivated by the observation that step-by-step reasoning protocols have demonstrated notable success in improving logical consistency and interpretability for complex question-answering tasks in purely textual domains. While this particular exploration was initially undertaken with the aim of enhancing the model's overall reasoning capabilities, our subsequent analysis revealed that its applicability within the broader L3DM framework remains somewhat limited, a situation most likely attributable to the inherently structured and complex spatial reasoning demands posed by 3D scenes. These demands fundamentally necessitate the implementation of more explicitly guided and structured approaches than what zero-shot CoT methodologies are typically able to provide.

This study has several limitations. The current evaluation is mainly conducted on ScanQA and does not establish generalization across situated reasoning, explicit 3D grounding, or embodied tasks. In addition, the framework-level adaptations, alignment module, and LoRA strategy are not evaluated through a fully controlled component-wise study. Cross-framework comparisons should also be interpreted cautiously because the compared systems differ in backbone and training configuration. Future work will extend L3DM to broader 3D vision-language tasks and investigate geometry-aware reasoning mechanisms beyond language-only zero-shot CoT prompting.

Just Accepted

References

[1] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

[2] Xiaohong Liu, Xiongkuo Min, Guangtao Zhai, Chunyi Li, Tengchuan Kou, Wei Sun, Haoning Wu, Yixuan Gao, Yuqin Cao, Zicheng Zhang, et al. Ntire 2024 quality assessment of ai-generated content challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6337–6362, 2024.

[3] Shangkun Sun, Yuanqi Chen, Yu Zhu, Guodong Guo, and Ge Li. Skflow: Learning optical flow with super kernels. *Advances in Neural Information Processing Systems*, 35:11313–11326, 2022.

[4] Bowen Qu, Xiaoyu Liang, Shangkun Sun, and Wei Gao. Exploring aigc video quality: A focus on visual harmony, video-text consistency and domain distribution gap. *arXiv preprint arXiv:2404.13573*, 2024.

[5] Shangkun Sun, Bowen Qu, Xiaoyu Liang, Songlin Fan, and Wei Gao. Ie-bench: Advancing the measurement of text-driven image editing for human perception alignment. *arXiv preprint arXiv:2501.09927*, 2025.

[6] Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3d-vista: Pre-trained transformer for 3d vision and text alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2911–2921, 2023.

[7] 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.

[8] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European conference on computer vision*, pages 202–221. Springer, 2020.

[9] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[10] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[11] Shangkun Sun, Jiaming Liu, Thomas H Li, Huaxia Li, Guoqing Liu, and Wei Gao. Streamflow: Streamlined multi-frame optical flow estimation for video sequences. *Advances in Neural Information Processing Systems*, 2024.

[12] Shangkun Sun, Xiaoyu Liang, Songlin Fan, Wenxu Gao, and Wei Gao. Ve-bench: Subjective-aligned benchmark suite for text-driven video editing quality assessment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7105–7113, 2025.

[13] Shangkun Sun, Xiaoyu Liang, Bowen Qu, and Wei Gao. Content-rich aigc video quality assessment via intricate text alignment and motion-aware consistency. *arXiv preprint arXiv:2502.04076*, 2025.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[15] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[16] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[17] Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. *arXiv preprint arXiv:2311.12871*, 2023.

[18] Zhihang Liu, Jun Li, Hongtao Xie, Pandeng Li, Jiannan Ge, Sun-Ao Liu, and Guoqing Jin. Towards balanced alignment: Modal-enhanced semantic modeling for video moment retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 3855–3863, 2024.

[19] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[21] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[22] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[23] Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. A survey of knowledge-enhanced text generation. *ACM Computing Surveys*, 54(11s):1–38, 2022.

[24] Zayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. Musr: Testing the limits of chain-of-thought with multistep soft reasoning. *arXiv preprint arXiv:2310.16049*, 2023.

[25] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020.

[26] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[27] David B Camarillo, Thomas M Krummel, and J Kenneth Salisbury Jr. Robotic technology in surgery: past, present, and future. *The American Journal of Surgery*, 188(4):2–15, 2004.

[28] Julie Carmigniani and Borko Furht. Augmented reality: an overview. *Handbook of augmented reality*, pages 3–46, 2011.

[29] Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(2):230–244, 2022.

[30] Shuquan Ye, Dongdong Chen, Songfang Han, and Jing Liao. 3d question answering. *IEEE Transactions on Visualization and Computer Graphics*, 30(3):1772–1786, 2022.

[31] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[32] Yuanqi Chen, Shangkun Sun, Ge Li, Wei Gao, and Thomas H Li. Closing the gap between theory and practice during alternating optimization for gans. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[33] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[34] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19129–19139, 2022.

[35] Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. Iqa: Visual question answering in interactive environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4089–4098, 2018.

[36] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. *arXiv preprint arXiv:2210.07474*, 2022.

[37] Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15768–15780, 2025.

[38] Taolin Zhang, Sunan He, Tao Dai, Zhi Wang, Bin Chen, and Shu-Tao Xia. Vision-language pre-training with object contrastive learning for 3d scene understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7296–7304, 2024.

[39] 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.

[40] Junyu Luo, Jiahui Fu, Xianghao Kong, Chen Gao, Haibing Ren, Hao Shen, Huaxia Xia, and Si Liu. 3d-sps: Single-stage 3d visual grounding via referred point progressive selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16454–16463, 2022.

[41] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Language conditioned spatial relation reasoning for 3d object grounding. *Advances in neural information processing systems*, 35:20522–20535, 2022.

[42] Haomiao Xiong, Yunzhi Zhuge, Jiawen Zhu, Lu Zhang, and Huchuan Lu. 3ur-llm: An end-to-end multimodal large language model for 3d scene understanding. *IEEE Transactions on Multimedia*, 27:2899–2911, 2025.

[43] Yunzhi Zhuge, Sitong Gong, Lu Zhang, Qi Xu, Wenda Zhao, Jin Zhan, and Huchuan Lu. Context-infused trajectories: Enhancing context and frame consistency in reasoning video object segmentation. *IEEE Transactions on Image Processing*, 2026.

[44] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[45] Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023.

[46] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Just Accepted

Author biographies

Shangkun Sun received the B.S. degree from the School of Data and Computer Science, Sun Yatsen University, Guangzhou, China, in 2021. He is currently pursuing the Ph.D. degree with the School of Electronic and Computer Engineering, Shenzhen Graduate School, Peking University, Shenzhen, China. His research interests include generative models and multi-modality learning.

Junhong Lin received the B.S. and M.S. degrees from the College of Physics and Information Engineering, Fuzhou University, Fuzhou, China, in 2021 and 2024, respectively. He is currently working toward the Ph.D. degree with the School of Electronic and Computer Engineering, Shenzhen Graduate School, Peking University, Shenzhen, China. His research interests include computer vision and spatial intelligence.

Yuqi Ye is currently a Master student with the School of Electronic and Computer Engineering, Peking University, China. Prior to that, he received the B.S. degree from Yingcai Honors College, University of Electronic Science and Technology of China, Chengdu, China, in 2024. His research interests are multimodal large models and vision-language-action models.

Ge Li (Member, IEEE) received the Ph.D. degree from the Department of Electrical Engineering, Auburn University, AL, USA, in 1999. He joined the School of Electronic and Computer Engineering, Peking University, China, in 2014, as a Full Professor. His general research interests include multimedia coding, image/video analysis, and signal processing.

Wei Gao (Senior Member, IEEE) received the Ph.D. degree in computer science from City University of Hong Kong, Hong Kong, in February 2017. In 2016, he was a Visiting Scholar at the University of California at Los Angeles (UCLA), CA, USA. From 2017 to 2019, he worked as a Postdoctoral Fellow at City University of Hong Kong, and a Research Fellow at Nanyang Technological University, Singapore. He joined the School of Electronic and Computer Engineering, Shenzhen Graduate School, Peking University, Shenzhen, China, in 2019, where he is currently a Tenured Associate Professor. His research interests include multimedia coding, multimedia processing, and artificial intelligence, especially 3D visual data coding, processing, and applications. He is an Elected Member of IEEE Multimedia Systems and Applications Technical Committee (MSA-TC), IEEE Visual Signal Processing and Communications Technical Committee (VSPC-TC), and APSIPA Image, Video, and Multimedia Technical Committee (IVM-TC). He is currently an Associate Editor of IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and IEEE TRANSACTIONS ON MULTIMEDIA. He was an organizer of workshops and special sessions, and a tutorial speaker at several international conferences. He is also the Leader to establish several open source projects, including OpenPointCloud, OpenAICoding, OpenAIDriving, and OpenDatasets. He has authored or co-authored four books published by Springer Nature, including "AI-based 3D Point Cloud Coding", "AI-based Image and Video Coding", "Deep Learning for 3D Point Clouds", and "Point Cloud Compression".

Just Accepted