# Scalable Object Relation Encoding for Better 3D Spatial Reasoning in Large Language Models

Shengli Zhou[1], Minghang Zheng[2], Feng Zheng[1], Yang Liu[2]✉
[1] Southern University of Science and Technology [2] Peking University

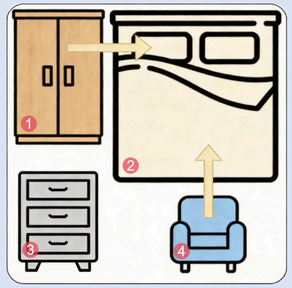
CVPR JUNE 3-7, 2026

Project Page | Paper | Video | Code

## 3D Vision-Language (3D VL) Tasks

▶ Input 3D point cloud + text and output textual response.
▶ Core: spatial reasoning, locate targets using spatial relations.

Examples:
[**3D Visual Grounding**] Locate the bed next to the right of the nightstand and next to the chair.
[Output] <obj002>
[**3D Visual Question-Answering**] What is the color of the chair next to the bed in the corner of the room?
[Output] Blue.




## Challenges of Injecting 3D Scenes into LLMs

**How to effectively represent 3D positions in the scene?**

▶ Encode absolute coordinates?
- Carries little inherent meaning.
- Does not explicitly represent relative geometry.
- Premature feature fusion obstructs position extraction.

▶ Encode relative coordinates?
- Encode all C(n, 2) relations for n objects as input tokens?
  Quadratic complexity, does not scale.
- Encoding partial spatial relations?
  Erroneous pruning affects accuracy.

## Method

▶ **QuatRoPE**: a novel 3D positional encoding
- **Scalable**: O(n) tokens for input
  Applies positional embedding on tokens for object features.
- **Complete**: O($n^2$) relations for inference
  Utilizes the dot product in the attention layer to calculate the relative position between the corresponding objects.
- **Accurate**: Correctly reflects relative position

M-RoPE's axis-wise encoding (left fig.)? Attention increases when coordinates on a certain axis approximate. ✗

Ours (right fig.): encodes position as a holistic vector via quaternion. ✓

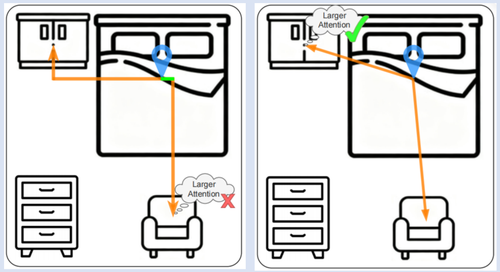
- Goal: $\langle f(\vec{q},\vec{m}), f(\vec{k},\vec{n})\rangle = g(\vec{q},\vec{k},\vec{m}-\vec{n})$

(Absolute Positions) (Relative Position)

- Approximate solution:

$$\begin{cases} f(\vec{q},\vec{m}) = Q(\vec{m})\,\vec{q}\,Q^{-1}(\vec{m}) \\ Q(\vec{m}) = Q_z(m_z)\,Q_y(m_y)\,Q_x(m_x) \\ Q_x(m_x) = \cos[m_x\theta_x(1)/2] + \hat{i}\sin[m_x\theta_x(1)/2] \\ Q_y(m_y) = \cos[m_y\theta_y(1)/2] + \hat{j}\sin[m_y\theta_y(1)/2] \\ Q_z(m_z) = \cos[m_z\theta_z(1)/2] + \hat{k}\sin[m_z\theta_z(1)/2] \end{cases}$$

where Q's are rotation matrices and θ's are rotation frequencies.




▶ **IGRE**: prevent interference between QuatRoPE and language RoPE.
- Extend dedicated dimensions: QuatRoPE rotation / zero-padding.

(tokens for object features) (other tokens)

**Isolated** **Gated**

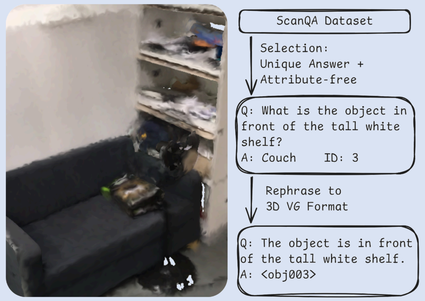
- Preserves LLMs' abilities + enables spatial awareness.

## Attribute-free Spatial Reasoning Benchmark

Previous 3D-VL tasks: spatial relations often entangle with category / attribute cues in input text, models can solve by detection. ✗

**ASR benchmark**: omits category / attribute cues, exclusively evaluates spatial reasoning ability. ✓

Pipeline example:




## Experiment Results

Table 1. QuatRoPE achieves consistent gains on existing 3D VL benchmarks.

| Model | ScanRefer Acc@0.25 | ScanRefer Acc@0.5 | ScanRefer Multi@0.25 | ScanRefer Multi@0.5 | Multi3DRef F1@0.25 | Multi3DRef F1@0.5 | SQA3D EM@1 |
|---|---|---|---|---|---|---|---|
| ScanRefer [5] | 39.0 | 26.1 | 32.1 | 21.3 | – | – | – |
| 3DJCG [4] | 49.6 | 37.3 | 41.4 | 30.8 | – | 26.6 | – |
| Vil3DRef [7] | 47.9 | 37.7 | 40.3 | 30.7 | – | – | – |
| D3Net [6] | – | 37.9 | – | 30.1 | – | 32.2 | – |
| VPP-Net [24] | 55.7 | 43.3 | 50.3 | 39.0 | – | – | – |
| AugRefer [29] | 55.7 | 44.0 | 50.0 | 39.1 | – | – | – |
| M3DRef-CLIP [35] | – | 44.7 | – | 36.8 | 42.8 | 38.4 | – |
| MA2TransVG [31] | 57.9 | 45.7 | 53.8 | 41.4 | – | – | – |
| 3D-VisTA [37] | 50.6 | 45.8 | 43.7 | 39.1 | – | – | 48.5 |
| 3DSyn [32] | 52.3 | 46.2 | – | – | – | – | – |
| TSP3D [12] | 56.5 | 46.7 | – | – | – | – | – |
| PQ3D [38] | – | 51.2 | – | 46.2 | – | 50.1 | 47.1 |
| BridgeQA [20] | – | – | – | – | – | – | 52.9 |
| Scene-LLM [9] | – | – | – | – | – | – | 53.6 |
| Chat-Scene-7B [14] | 55.5 | 50.2 | 47.8 | 42.9 | 57.1 | 52.4 | 54.6 |
| **Chat-Scene-7B + QuatRoPE (Ours)** | 57.8 | 52.2 | 51.1 | 45.7 | 59.5 | 54.8 | 54.7 |
| 3DGraphLLM-7B [34] | 57.0 | 51.3 | – | – | 60.1 | 55.4 | 53.1 |
| **3DGraphLLM-7B + QuatRoPE (Ours)** | **58.2** | 52.5 | **54.3** | **49.2** | **60.6** | **56.0** | **55.2** |

Table 2. QuatRoPE achieves large and consistent gains on the ASR benchmark, directly verifying that our method can boost models' spatial reasoning ability.

| Model | LLM | Acc @ 0.25 | Gain | Acc @ 0.5 | Gain |
|---|---|---|---|---|---|
| Chat-Scene [14] | Llama-3.2-1B-Instruct | 22.92 | – | 22.92 | – |
| **Chat-Scene + QuatRoPE (Ours)** | Llama-3.2-1B-Instruct | 27.38 | 4.46 (19.48%) | 27.38 | 4.46 (19.48%) |
| 3DGraphLLM [34] | Llama-3.2-1B-Instruct | 25.89 | – | 25.60 | – |
| **3DGraphLLM + QuatRoPE (Ours)** | Llama-3.2-1B-Instruct | 29.76 | 3.87 (14.94%) | 29.76 | 4.17 (16.28%) |
| 3DGraphLLM [34] | Llama-3-8B-Instruct | 37.50 | – | 36.90 | – |
| **3DGraphLLM + QuatRoPE (Ours)** | Llama-3-8B-Instruct | 41.96 | 4.46 (11.90%) | 41.96 | 5.06 (13.71%) |