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

Shengli Zhou[1] Minghang Zheng[2] Feng Zheng[1] Yang Liu[2,3]✉
[1]Department of Computer Science and Engineering, Southern University of Science and Technology
[2]Wangxuan Institute of Computer Technology, Peking University
[3]State Key Laboratory of General Artificial Intelligence, Peking University
zhousl2022@mail.sustech.edu.cn, {minghang, yangliu}@pku.edu.cn, f.zheng@ieee.org

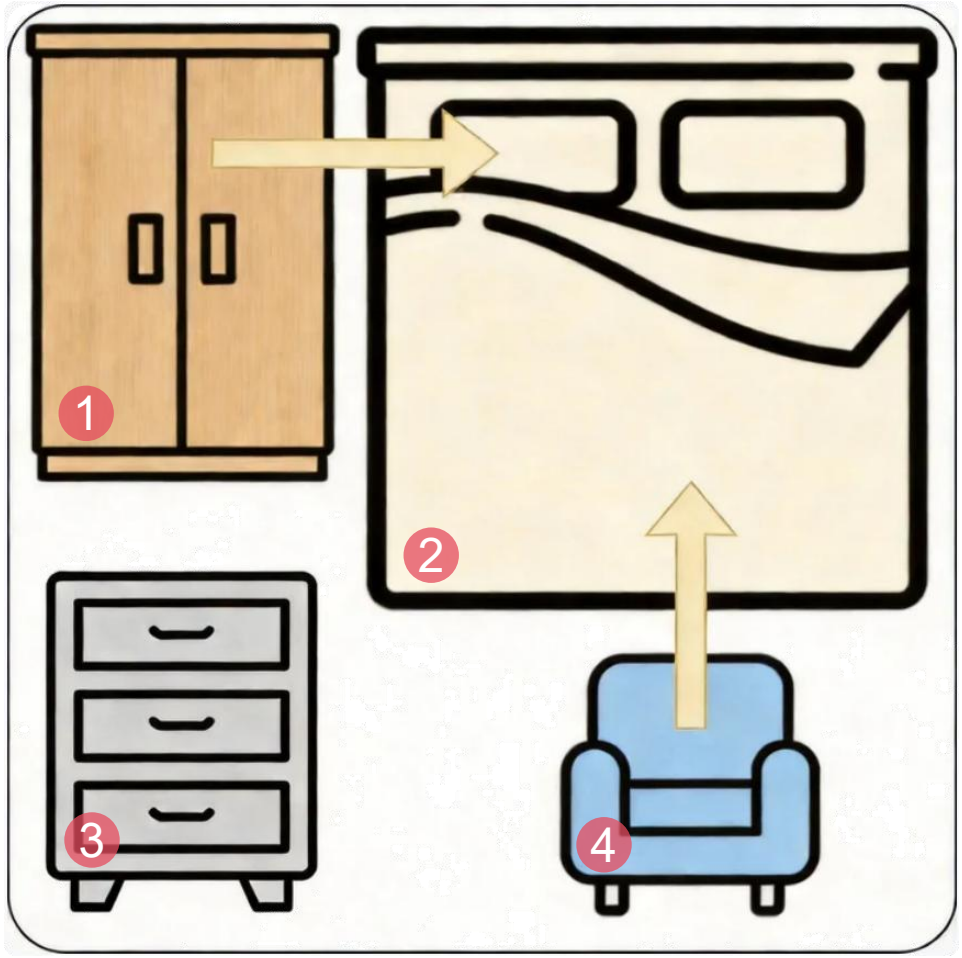
3D Vision-Language (3D VL) Tasks

CVPR JUNE 3-7, 2026 DENVER COLORADO

- Input: point cloud (3D scene) + text.
- Output: text (response).
- Core ability: spatial reasoning, locate target objects 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?
- [Output] Blue.



Challenge 1
Data Scarcity

CVPR JUNE 3-7, 2026 DENVER COLORADO

Insufficient 3D vision-language paired data

↓

Unable to train models with strong spatial reasoning ability from scratch

↓

Inject 3D scene into LLMs (i.e., 3D LLM),
utilize pretrained reasoning ability

How to effectively represent 3D positions in the scene?

Challenge 2
Scene Representation

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 and calculation.
- Encode relative coordinates?
  - Encode all $\binom{n}{2}$ relations for n objects? Quadratic complexity, does not scale.
  - Encoding partial spatial relations? Erroneous pruning affects accuracy.

Method: Core Idea

CVPR JUNE 3-7, 2026 DENVER COLORADO

Scalable: O(n) tokens for input

Apply positional embedding on tokens for object feature.

Complete: $O(n^2)$ relations for inference

Utilize the dot product in the attention layer to calculate the relative position between the objects for the query / key vector.

Accurate: Correctly reflect relative position between objects

Encode position as a holistic vector, avoiding inflated attention scores from small coordinate differences on single axes.

Method: QuatRoPE

- **Goal:** Absolute Coordinates $\xrightarrow{\text{Dot Product}}$ Relative Positions

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

Absolute Position Relative Position

- **An approximate solution:**

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

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

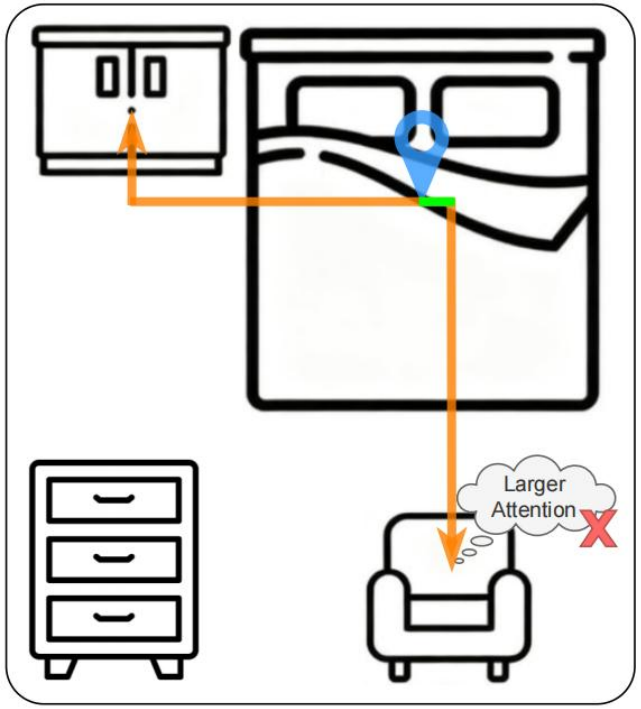
Scalable

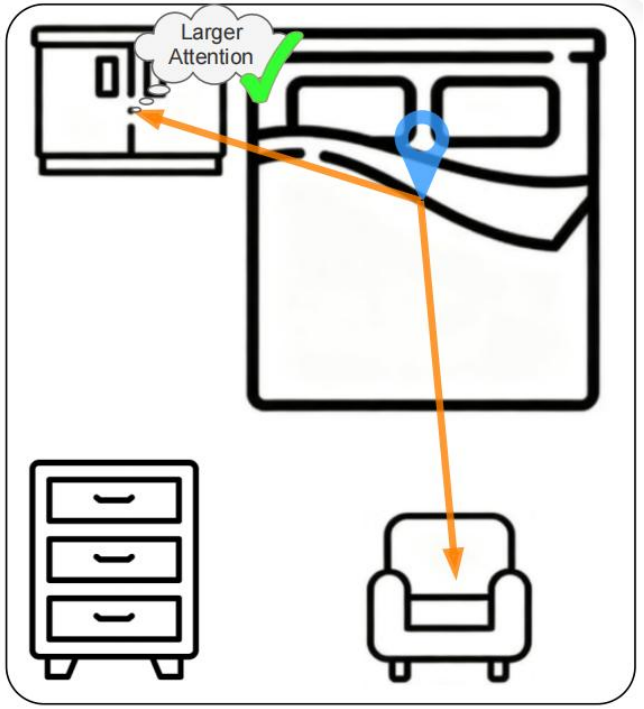
Complete

Accurate

Method: QuatRoPE

- How to effectively represent spatial relation via attention scores?
- Encode each coordinate independently? ✕
  - Attention increases when coordinates on a certain axis approximate.
- QuatRoPE: encode 3D position as a holistic vector via quaternion.





Scalable

Complete

Accurate

Method: IGRE

- We propose Isolated Gated RoPE Extension (IGRE) to prevent interference between QuatRoPE and the original (language) RoPE in LLMs (as they both rotate tokens) .
- Extend dedicated dimensions: QuatRoPE rotation / zero-padding.

(tokens for object features) (other tokens)

isolated

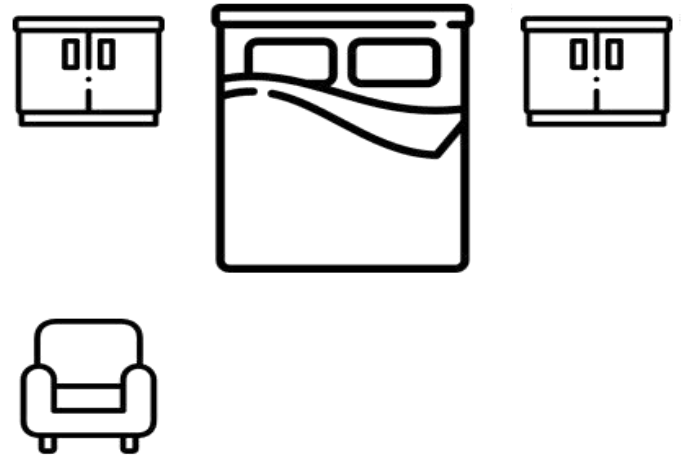
gated

- Preserves LLMs' abilities + enabling spatial awareness.

Attribute-free Spatial Reasoning (ASR) Benchmark

CVPR JUNE 3-7, 2026 DENVER COLORADO

- We propose the ASR benchmark to **exclusively evaluate spatial reasoning ability**.
- Previous 3D VL tasks: spatial relations often entangle with other cues (e.g., categories or attributes) in texts.
  - E.g., locate the wooden bed between the cabinets.
  - Can be solved via spatial reasoning or detection.



- **Attribute-free Spatial Reasoning (ASR)** benchmark: omits category / attribute cues
  - E.g., locate the object between the cabinets.
  - Follows the visual grounding format, eliminating the difference in language generation capability (the model only does multiple choice questions).

Metrics for Experiments

- ScanRefer (3D Visual Grounding): Acc @ 0.25, Acc @ 0.5, Multi @ 0.25, Multi @ 0.5
  - The accuracy on the entire dataset / a subset with other objects of the same type as the target object (which requires spatial reasoning).
  - The output is regarded correct when its IoU with ground truth exceeds 0.25 / 0.5.
- Multi3DRef (3D Visual Grounding): F1 @ 0.25, F1 @ 0.5
  - F1 score, the output is correct if its IoU with ground truth exceeds 0.25 / 0.5.
- SQA3D (3D Visual Question-Answering): EM @ 1
  - Accuracy, correct when top-confidence output exactly matches ground truth.

Comparative Experiment

- Consistent gains under all metrics: QuatRoPE effectively conveys spatial relations.

Model	Detector / Segmentation	ScanRefer				Multi3DRef		SQA3D
		Acc@0.25	Acc@0.5	Multi@0.25	Multi@0.5	F1@0.25	F1@0.5	EM@1
ScanRefer [5]	VoteNet	39.0	26.1	32.1	21.3	–	–	–
3DJCG [4]	VoteNet	49.6	37.3	41.4	30.8	–	26.6	–
Vil3DRef [7]	PointGroup	47.9	37.7	40.3	30.7	–	–	–
D3Net [6]	PointGroup	–	37.9	–	30.1	–	32.2	–
VPP-Net [24]	Group-free	55.7	43.3	50.3	39.0	–	–	–
AugRefer [29]	Group-free	55.7	44.0	50.0	39.1	–	–	–
M3DRef-CLIP [35]	PointGroup	–	44.7	–	36.8	42.8	38.4	–
MA2TransVG [31]	Group-free	57.9	45.7	53.8	41.4	–	–	–
3D-VisTA [37]	Mask3D	50.6	45.8	43.7	39.1	–	–	48.5
3DSyn [32]	Mask3D	52.3	46.2	–	–	–	–	–
TSP3D [12]	N/A	56.5	46.7	–	–	–	–	–
PQ3D [38]	PQ3D Promptable	–	51.2	–	46.2	–	50.1	47.1
BridgeQA [20]	VoteNet	–	–	–	–	–	–	52.9
Scene-LLM [9]	N/A	–	–	–	–	–	–	53.6
Chat-Scene-1B [14]	GT	50.7	50.3	42.7	42.3	53.3	52.9	50.7
Chat-Scene-1B + QuatRoPE (Ours)	GT	55.4	55.0	47.8	47.4	58.1	57.7	53.1
3DGraphLLM-1B [34]	GT	55.9	55.8	47.9	47.7	58.6	58.4	51.1
3DGraphLLM-1B + QuatRoPE (Ours)	GT	**58.3**	**58.2**	50.8	**50.6**	**60.7**	**60.5**	53.2
Chat-Scene-7B [14]	Mask3D	55.5	50.2	47.8	42.9	57.1	52.4	54.6
Chat-Scene-7B + QuatRoPE (Ours)	Mask3D	57.8	52.2	51.1	45.7	59.5	54.8	54.7
3DGraphLLM-7B [34]	Mask3D	57.0	51.3	–	–	60.1	55.4	53.1
3DGraphLLM-7B + QuatRoPE (Ours)	Mask3D	58.2	52.5	**54.3**	49.2	60.6	56.0	**55.2**

Experiment on ASR

- Consistent and large-margin 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%)

Ablation Studies

RoPE Composition Approach	ScanRefer				SQA3D
	Acc @ 0.25	Acc @ 0.5	Multi @ 0.25	Multi @ 0.5	EM @ 1
Baseline: Chat-Scene [14]					
None	50.72	50.33	42.69	42.29	50.72
Trans-Additive	53.12	52.79	45.48	45.14	52.96
IGRE (Ours)	**55.44**	**55.00**	**47.81**	**47.36**	**53.14**
Baseline: 3DGraphLLM [34]					
None	55.92	55.75	47.92	47.74	51.09
Trans-Additive	53.68	53.38	45.94	45.64	51.55
IGRE (Ours)	**58.30**	**58.15**	**50.77**	**50.60**	**53.20**

Explicit Positional Encoding Approach	ScanRefer				SQA3D
	Acc @ 0.25	Acc @ 0.5	Multi @ 0.25	Multi @ 0.5	EM @ 1
Baseline: Chat-Scene [14]					
None	50.72	50.33	42.69	42.29	50.72
Raw Coordinates	52.26	52.01	44.41	44.17	51.40
M-RoPE	54.30	53.92	46.44	46.10	51.55
QuatRoPE (Ours)	**55.44**	**55.00**	**47.81**	**47.36**	**53.14**
Baseline: 3DGraphLLM [34]					
None	55.92	55.75	47.92	47.74	51.09
Raw Coordinates	3.60	3.44	3.57	3.46	35.50
M-RoPE	57.69	57.48	50.07	49.83	53.14
QuatRoPE (Ours)	**58.30**	**58.15**	**50.77**	**50.60**	**53.20**

- IGRE is better than rotating with language RoPE simultaneously

👉 Effectively reduces interference

- QuatRoPE is better than M-RoPE and adding raw coordinates

👉 PE is crucial for 3D VL tasks

👉 It is better to encode positions as holistic vectors

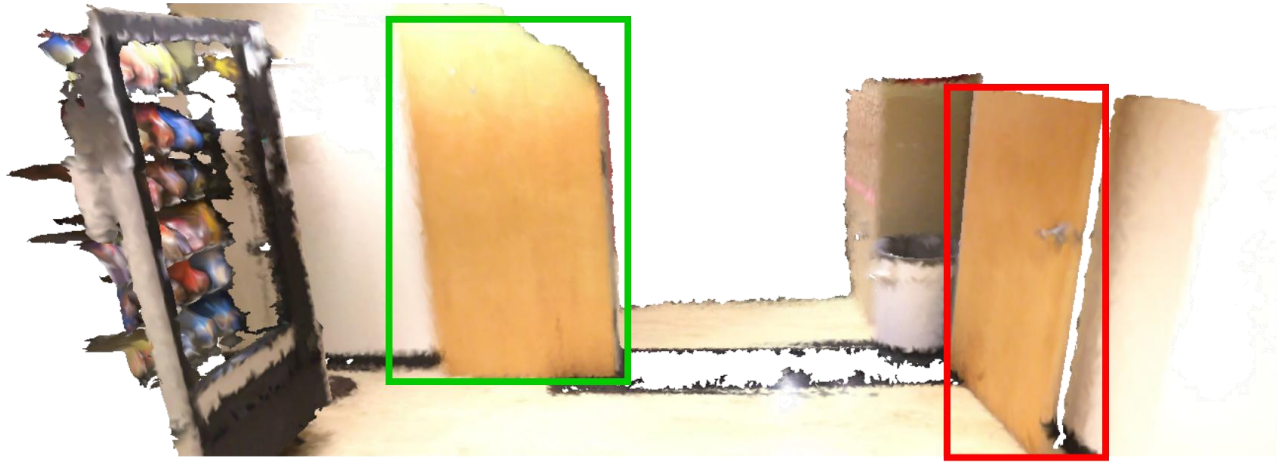
Qualitative Results

(a) This is a brown table. It is surrounded by quite a few matching chairs.

QuatRoPE: Chat-Scene:

(b) This is a tan wood door. ... It is to the right of a snack machine.

QuatRoPE: Chat-Scene:



By applying QuatRoPE, models can:

- Better align spatial relations with text descriptions
- Align with human implication (Maxim of Relation) better

Conclusion

- We propose QuatRoPE, a 3D positional encoding with linear-complexity input that provides pairwise object relative positions.
- We propose Isolated Gated RoPE Extension (IGRE) for minimizing the interference between QuatRoPE and language RoPE.
- We construct the ASR benchmark for exclusively evaluating 3D spatial reasoning.
- We achieve consistent and large-margin gains on ASR and multiple existing 3D VL benchmarks, validating the effectiveness.

Thank you!

Project Page

Paper

Code