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

Supplementary Material

1. Mathematical Derivation for QuatRoPE

In this section, we provide a detailed mathematical derivation for QuatRoPE.

Let $\vec{m}$ and $\vec{n}$ be the absolute 3D coordinates of the objects corresponding to query vector $\vec{q}$ and key vector $\vec{k}$, and $f(\vec{x}, \vec{p})$ be the function for rotating the query or key vector $\vec{x}$ with a corresponding 3D position $\vec{p}$. Since the attention score should only relate to the relative position (i.e., $\vec{m} - \vec{n}$), the rotation function f should satisfy:

$$\left\langle f(\vec{q}, \vec{m}), f(\vec{k}, \vec{n}) \right\rangle = g(\vec{q}, \vec{k}, \vec{m} - \vec{n}) \tag{1}$$

In QuatRoPE, we embed the coordinates as a holistic vector by applying quaternion rotations to query and key vectors. Formally, the rotation function can be expressed as:

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

where Q's are rotation matrices and θ's are unary functions.

Through Equation (2), we transform the requirement in QuatRoPE (i.e., converting absolute coordinates to relative positions via dot products) into deriving θ's that satisfy Equation (1). To solve the equation, we transform the dot product into the real part of the product of the rotation functions to yield a form with multiplication between the rotation matrices (i.e., $Q^{-1}(\vec{m})$ and $Q(\vec{n})$).

$$\begin{aligned} &\left\langle f(\vec{q}, \vec{m}), f(\vec{k}, \vec{n}) \right\rangle \\ =&\Re[f(\vec{q}, \vec{m}) f^*(\vec{k}, \vec{n})] \\ =&\Re[Q(\vec{m})\ \vec{q}\, Q^{-1}(\vec{m})\ \overline{Q(\vec{n})\ \vec{k}\, Q^{-1}(\vec{n})}] \\ =&\Re[Q(\vec{m})\ \vec{q}\, Q^{-1}(\vec{m})\ Q(\vec{n})\ \vec{k}^*\ Q^{-1}(\vec{n})] \end{aligned} \tag{3}$$

where $\vec{k}^*$ denotes the conjugate of quaternion $\vec{k}$, and $\Re$ denotes the real part of the quaternion. To pair every $Q(\vec{m})$ with $Q(\vec{n})$, according to the real-part invariance of quaternion rotation, after left multiplying $Q^{-1}(\vec{m})$ and right multiplying $Q(\vec{m})$, Equation (3) yields:

$$\begin{aligned} &\left\langle f(\vec{q}, \vec{m}), f(\vec{k}, \vec{n}) \right\rangle \\ =&\Re[Q(\vec{m})\ \vec{q}\, Q^{-1}(\vec{m})\ Q(\vec{n})\ \vec{k}^*\ Q^{-1}(\vec{n})\ Q(\vec{m})\ Q^{-1}(\vec{m})] \\ =&\Re[\vec{q}\, Q^{-1}(\vec{m})\ Q(\vec{n})\ \vec{k}^*\ Q^{-1}(\vec{n})\ Q(\vec{m})] \end{aligned} \tag{4}$$

According to Equation (1), $\left\langle f(\vec{q}, \vec{m}), f(\vec{k}, \vec{n}) \right\rangle$ should only relate to $\vec{m} - \vec{n}$, the following equation should hold:

$$\Re[\vec{q}\, Q^{-1}(\vec{m})\ Q(\vec{n})\ \vec{k}^*\ Q^{-1}(\vec{n})\ Q(\vec{m})] = g(\vec{q}, \vec{k}, \vec{m} - \vec{n}) \tag{5}$$

Thus,

$$Q(\vec{m} - \vec{n}) = Q^{-1}(\vec{n})\ Q(\vec{m}) \tag{6}$$

i.e.,

$$\begin{aligned} &Q_z(m_z - n_z)\ Q_y(m_y - n_y)\ Q_x(m_x - n_x) \\ =&Q_x^{-1}(n_x)\ Q_y^{-1}(n_y)\ Q_z^{-1}(n_z)\ Q_z(m_z)\ Q_y(m_y)\ Q_x(m_x) \end{aligned} \tag{7}$$

When $\vec{m} = \vec{n} = \vec{0}$, we have $Q(\vec{0})\ Q(\vec{0}) = Q(\vec{0})$. Thus, $Q(\vec{0}) = 1$. According to Equation (2),

$$\begin{aligned} 1 =&Q(\vec{0}) \\ =&Q_z(0)\ Q_y(0)\ Q_x(0) \\ =&\left[\cos\left(\frac{\theta_z(0)}{2}\right) + \hat{k} \sin\left(\frac{\theta_z(0)}{2}\right)\right] \\ &\left[\cos\left(\frac{\theta_y(0)}{2}\right) + \hat{j} \sin\left(\frac{\theta_y(0)}{2}\right)\right] \\ &\left[\cos\left(\frac{\theta_x(0)}{2}\right) + \hat{i} \sin\left(\frac{\theta_x(0)}{2}\right)\right] \end{aligned} \tag{8}$$

Consider the real part of the equation above, we have:

$$
\begin{aligned}
1 =& \Re\left\{\left[\cos\left(\frac{\theta_z(0)}{2}\right)+\hat{k}\sin\left(\frac{\theta_z(0)}{2}\right)\right]\right. \\
&\left[\cos\left(\frac{\theta_y(0)}{2}\right)+\hat{j}\sin\left(\frac{\theta_y(0)}{2}\right)\right] \\
&\left.\left[\cos\left(\frac{\theta_x(0)}{2}\right)+\hat{i}\sin\left(\frac{\theta_x(0)}{2}\right)\right]\right\} \\
=& \cos\left(\frac{\theta_z(0)}{2}\right)\cos\left(\frac{\theta_y(0)}{2}\right)\cos\left(\frac{\theta_x(0)}{2}\right) \\
&+\hat{k}\hat{j}\hat{i}\sin\left(\frac{\theta_z(0)}{2}\right)\sin\left(\frac{\theta_y(0)}{2}\right)\sin\left(\frac{\theta_x(0)}{2}\right) \\
=& \cos\left(\frac{\theta_x(0)}{2}\right)\cos\left(\frac{\theta_y(0)}{2}\right)\cos\left(\frac{\theta_z(0)}{2}\right) \\
&+\sin\left(\frac{\theta_x(0)}{2}\right)\sin\left(\frac{\theta_y(0)}{2}\right)\sin\left(\frac{\theta_z(0)}{2}\right)
\end{aligned} \tag{9}
$$

Also, since the imaginary part of Equation (8) is 0, either all cosines or all sines are equal to 0. Therefore

$$
\cos\left(\frac{\theta_x(0)}{2}\right)=\cos\left(\frac{\theta_y(0)}{2}\right)=\cos\left(\frac{\theta_z(0)}{2}\right)=1 \tag{10}
$$

or

$$
\sin\left(\frac{\theta_x(0)}{2}\right)=\sin\left(\frac{\theta_y(0)}{2}\right)=\sin\left(\frac{\theta_z(0)}{2}\right)=1 \tag{11}
$$

Consider the first solution, let $m_x = m_y = n_x = n_y = 0$, Equation (7) yields:

$$
\begin{aligned}
&Q_z(m_z-n_z)\,Q_y(0-0)\,Q_x(0-0) \\
=&Q_x^{-1}(0)\,Q_y^{-1}(0)\,Q_z^{-1}(n_z)\,Q_z(m_z)\,Q_y(0)\,Q_x(0)
\end{aligned} \tag{12}
$$

i.e.,

$$
Q_z(m_z-n_z)=Q_z^{-1}(n_z)\,Q_z(m_z) \tag{13}
$$

For Equation (13), the left-hand side

$$
\begin{aligned}
&Q_z(m_z-n_z) \\
=&\cos\left(\frac{\theta_z(m_z-n_z)}{2}\right)+\sin\left(\frac{\theta_z(m_z-n_z)}{2}\right)\hat{k}
\end{aligned} \tag{14}
$$

while the right-hand side

$$
\begin{aligned}
&Q_z^{-1}(n_z)\,Q_z(m_z) \\
=&\left[\cos\left(\theta_z(n_z)/2\right)-\hat{k}\sin\left(\theta_z(n_z)/2\right)\right] \\
&\left[\cos\left(\theta_z(m_z)/2\right)+\hat{k}\sin\left(\theta_z(m_z)/2\right)\right] \\
=&\left[\cos\left(\theta_z(n_z)/2\right)\cos\left(\theta_z(m_z)/2\right)\right. \\
&+\sin\left(\theta_z(n_z)/2\right)\sin\left(\theta_z(m_z)/2\right)] \\
&+\left[\cos\left(\theta_z(n_z)/2\right)\sin\left(\theta_z(m_z)/2\right)\right. \\
&\left.-\sin\left(\theta_z(n_z)/2\right)\cos\left(\theta_z(m_z)/2\right)\right]\hat{k} \\
=&\cos\left(\frac{\theta_z(m_z)-\theta_z(n_z)}{2}\right)+\sin\left(\frac{\theta_z(m_z)-\theta_z(n_z)}{2}\right)\hat{k}
\end{aligned} \tag{15}
$$

By Equation (14) and Equation (15), we have

$$
\theta_z(m_z)-\theta_z(n_z)=\theta_z(m_z-n_z) \tag{16}
$$

When $m_z = n_z$, Equation (16) yields:

$$
\begin{aligned}
\theta_z(0) &= \theta_z(m_z-n_z) \\
&= \theta_z(m_z)-\theta_z(n_z) \\
&= 0
\end{aligned} \tag{17}
$$

Then, for any $t \in \mathbb{Z}$, we have

$$
\begin{aligned}
\theta_z(t) =& \theta_z(t-1)+\theta_z(1) \\
=& \theta_z(t-2)+\theta_z(1)+\theta_z(1) \\
=& \cdots \\
=& \theta_z(0)+t\theta_z(1) \\
=& t\theta_z(1)
\end{aligned} \tag{18}
$$

Moreover, for any $t, p \in \mathbb{Z}$ and $(t, p) = 1$ (i.e., $\frac{t}{p} \in \mathbb{Q}$), we have

$$
\begin{aligned}
\theta_z(t) =& \theta_z\left(\frac{t(p-1)}{p}\right)+\theta_z\left(\frac{t}{p}\right) \\
=& \theta_z\left(\frac{t(p-2)}{p}\right)+\theta_z\left(\frac{t}{p}\right)+\theta_z\left(\frac{t}{p}\right) \\
=& \cdots \\
=& p\theta_z\left(\frac{t}{p}\right)
\end{aligned} \tag{19}
$$

and hence

$$
\theta_z\left(\frac{t}{p}\right)=\frac{1}{p}\theta_z(t)=\frac{t}{p}\theta_z(1) \tag{20}
$$

Also, since the embedding should be continuous with respect to the position, θ_z should be continuous, and the solution to θ_z is

Table 1. Comparison between fixed and learnable base vectors for rotation.

Model	Base Vector	ScanRefer		SQA3D	Multi3dRef	
		Acc @ 0.25	Acc @ 0.5	EM @ 1	F1 @ 0.25	F1 @ 0.5
Chat-Scene [2]	Fixed	55.44	55.00	53.14	58.09	57.72
	Learnable	54.47	54.14	52.84	57.96	57.74
3DGraphLLM [3]	Fixed	58.30	58.15	53.20	60.70	60.52
	Learnable	56.89	56.64	52.68	60.67	60.51

Table 2. Comparison between different frequencies.

Frequency	ScanRefer		SQA3D	Multi3dRef	
	Acc @ 0.25	Acc @ 0.5	EM @ 1	F1 @ 0.25	F1 @ 0.5
0.3 (Default)	**58.30**	**58.15**	**60.70**	**60.52**	**53.20**
0.1 (Small)	54.55	54.39	58.02	57.90	51.99
1.0 (Large)	53.41	53.14	56.28	55.99	52.18

$$\theta_z(z) = z\theta_z(1), z \in \mathbb{R} \tag{21}$$

Let $n_z = m_z = 0$, according to Equation (7), we have

$$\begin{aligned} &Q_y(m_y - n_y)\, Q_x(m_x - n_x) \\ =&Q_x^{-1}(n_x)\, Q_y^{-1}(n_y)\, Q_y(m_y)\, Q_x(m_x) \end{aligned} \tag{22}$$

Similarly, the above equation yields

$$\theta_y(y) = y\theta_y(1), y \in \mathbb{R} \tag{23}$$

Again, let $n_y = m_y = n_z = m_z = 0$, Equation (22) yields

$$Q_x(m_x - n_x) = Q_x^{-1}(n_x)\, Q_x(m_x) \tag{24}$$

and thus

$$\theta_x(x) = x\theta_x(1), x \in \mathbb{R} \tag{25}$$

In conclusion, an approximate solution for QuatRoPE is:

$$\begin{cases} f(\vec{q}, \vec{m}) = Q(\vec{m})\, \vec{q}\, Q^{-1}(\vec{m}) \\ Q(\vec{m}) = Q_z(m_z)\, Q_y(m_y)\, Q_x(m_x) \\ Q_x(m_x) = \cos\left[\dfrac{m_x\theta_x(1)}{2}\right] + \hat{i}\sin\left[\dfrac{m_x\theta_x(1)}{2}\right] \\ Q_y(m_y) = \cos\left[\dfrac{m_y\theta_y(1)}{2}\right] + \hat{j}\sin\left[\dfrac{m_y\theta_y(1)}{2}\right] \\ Q_z(m_z) = \cos\left[\dfrac{m_z\theta_z(1)}{2}\right] + \hat{k}\sin\left[\dfrac{m_z\theta_z(1)}{2}\right] \end{cases} \tag{26}$$

where $\theta_x(1)$, $\theta_y(1)$, and $\theta_z(1)$ are frequencies for quaternion rotations. According to Equation (1), as we perform rotation by $\vec{q} := f(\vec{q}, \vec{m})$ and $\vec{k} := f(\vec{k}, \vec{n})$ before each attention layer, the attention scores between object-related tokens reflect their relative positions. By such an approach, QuatRoPE can effectively convey relative positional information for LLMs to perform spatial reasoning.

2. Experimental Settings

2.1. Base Vector for Rotation

In IGRE, the quaternion rotation of QuatRoPE is applied to the base vector to obtain the positional embedding. In this section, we compare the performance between using $(1, 0, 0)$ as a fixed base vector and the strategy of using a learnable base vector. Then we train and evaluate these approaches on Chat-Scene-1B [2] and 3DGraphLLM-1B [3], and the results are shown in Table 1.

The results indicate that learnable base vectors do not achieve better results. Such outcomes may result from the difficulty of learning base vectors, as these vectors have a significant impact on subsequent layers. Therefore, in our model, we set the base vector as $(1, 0, 0)$, which is also more computationally efficient.

2.2. Choice of Rotation Frequency

In the experiments, rotation frequency is set to 0.3 (untuned, consistent across all datasets) to avoid two issues shown in Tab. 2: (a) Small frequencies lead to small rotation angles, weakening feature vector influence and hindering learning. (b) Large frequencies cause the "wrapping" problem—large coordinate differences may produce similar rotation angles, misleading the model with incorrect scene layouts.

Given the maximum coordinate difference of 10, frequency is set to $\frac{\pi}{10} \approx 0.3$, ensuring all rotations lie in the same semi-circle and larger coordinate differences corre-

spond to larger angle differences.

Additionally, the error introduced by the non-commutativity of the Euler angle decomposition sequence is proportional to the square of the frequency. Thus, selecting a small frequency (e.g., 0.3) also makes QuatRoPE closer to the requirement of Equation (1).

3. Qualitative Results

In this section, we provide additional qualitative results to illustrate the effectiveness of QuatRoPE. The qualitative results are obtained from Chat-Scene-1B's [2] predictions on the validation split of the ScanRefer dataset [1].

The cases in Tables 3 - 5 demonstrate that QuatRoPE can effectively provide precise relative positions between objects. By providing explicit spatial relations between objects, models can directly perceive the scene layout without extracting and calculating objects' positions from prematurely fused features. Such a method significantly reduces the cost of training models to learn spatial reasoning, enabling them to achieve better performance.

References

[1] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 202–221. Springer, 2020. 4

[2] Haifeng Huang, Yilun Chen, Zehan Wang, Rongjie Huang, Runsen Xu, Tai Wang, Luping Liu, Xize Cheng, Yang Zhao, Jiangmiao Pang, et al. Chat-scene: Bridging 3d scene and large language models with object identifiers. *Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada*, 2024. 3, 4, 5, 6, 7

[3] Tatiana Zemskova and Dmitry Yudin. 3dgraphllm: Combining semantic graphs and large language models for 3d scene understanding, 2024. 3

Table 3. Qualitative Results

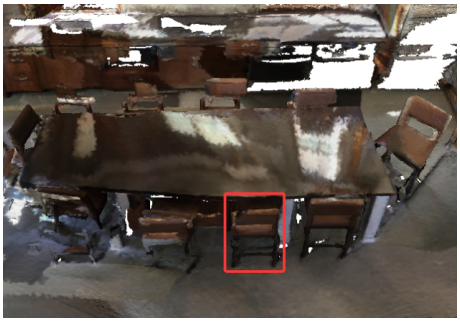
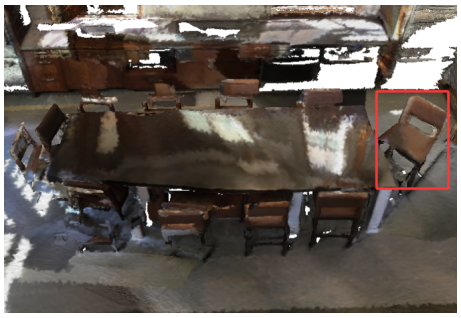
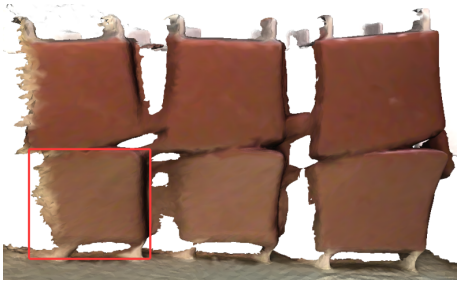
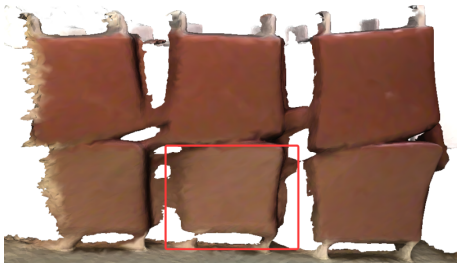
Description	Chat-Scene [2]	QuatRoPE (Ours)
This is a brown chair. It is turned toward the end of the table.		
Box-shaped footstool with a tarnished red color. There are 6 footstools stacked, 3 on the bottom row and 3 on the top. This is located on the bottom row in the middle.		
A blue towel that is hanging on the glass shower door. The towel is in the middle of the three towels hanging on the shower handle.		
This is a black office chair. It is facing the desk corner.		

Table 4. Qualitative Results (Continued)

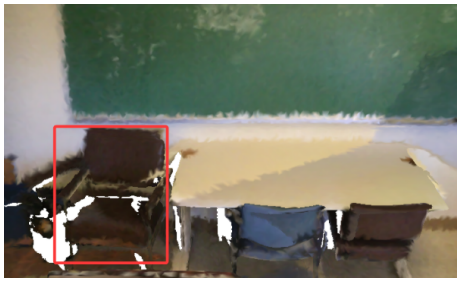
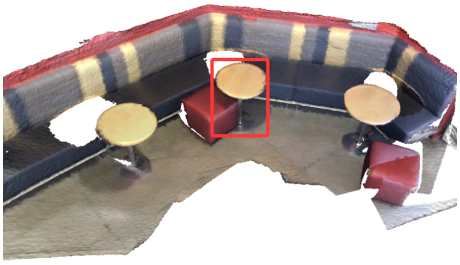
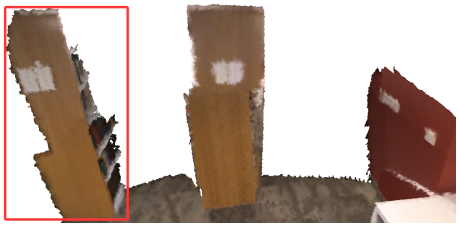
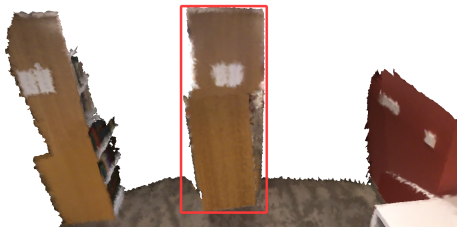
Description	Chat-Scene [2]	QuatRoPE (Ours)
It is a brown chair with armrests and four legs. It is directly under a blackboard.		
Case 1: This door appears to be the front door to the apartment. If you walk through the apartment and past the bathroom, you will encounter this door. The door is black and has a small window. Case 2: The door is rectangular in shape and has a small window on the upper portion. The door is located to the right of the bath area. Chat-Scene fails under both cases.		
The small rounded table. The table is next to the couch end.		
It is a tall gray trash can. The trash can is under the left side of the counter, to the left of the door when you enter.		

Table 5. Qualitative Results (Continued)

Description	Chat-Scene [2]	QuatRoPE (Ours)
Stand in front of the free-standing board in the room. Looking down the side of the table closest to you, it is the second chair down the row.		
Case 1: The monitor is next to the leftmost window. The monitor is black and rectangular. Case 2: The monitor is on the silver table. The monitor is the closest to the window.		
The bookshelf is between another bookshelf and a red wall. The bookshelf is brown and rectangular.		
The Ottoman is in the back, middle of the room. There is an identical ottoman to the right of it.