Rotary Position Encodings for Graphs

ICML 2026 Spotlight

Isaac Reid[*1,2] Arijit Sehanobish[*3] Cederik Höfs[*1]
Bruno Mlodozeniec[1,4] Leonhard Vulpius[1] Federico Barbero[5,2]
Adrian Weller[1,6] Krzysztof Choromanski[2]
Richard E. Turner[1,6] Petar Veličković[2,1]

[*]Core contributors. [1]Cambridge [2]Google DeepMind [3]Independent
[4]MPI for Intelligent Systems [5]Oxford [6]Alan Turing Institute

arXiv:2509.22259 github.com/cederikhoefs/Graph-RoPE

Position encodings: same problem, three domains

- Sentences: **order of words**
- Images: **layout of patches**
- Graphs: **node connectivity**

Transformers are permutation invariant.
Position encodings inject structure into attention.

Rotary Position Encodings

Rotate $\boldsymbol{q}_i, \boldsymbol{k}_i$ per token by position-dependent angles:

$$\{\boldsymbol{q}_i, \boldsymbol{k}_i\} \mapsto \bigoplus_{n=1}^{d/2} \begin{pmatrix} \cos\theta_n & -\sin\theta_n \\ \sin\theta_n & \cos\theta_n \end{pmatrix} \{\boldsymbol{q}_i, \boldsymbol{k}_i\}, \quad \theta_n = \boldsymbol{\omega}_n^\top \boldsymbol{r}_i$$

Logit depends only on the *relative* position:

$$\boldsymbol{q}_i^\top \boldsymbol{k}_j \;\to\; \boldsymbol{q}_i^\top \mathrm{RoPE}(\boldsymbol{r}_j - \boldsymbol{r}_i)\, \boldsymbol{k}_j$$

- Powers Llama, Gemma, modern ViTs

Can we apply RoPE to graphs?

Existing graph positional encodings

Two existing families of graph positional encodings to draw on:

- **Laplacian PE** (SAN, GraphGPS): *added* to each token

$$\boldsymbol{x}_i \to [\boldsymbol{x}_i,\ f(\{\lambda_k, \boldsymbol{\psi}_k[i]\}_k)]$$

  Spectral features carry rich structure; shown to help on long-range tasks

- **Structural-distance bias** (Graphormer, GraphiT): *added* to the logit

$$\boldsymbol{q}_i^\top \boldsymbol{k}_j + b(\mathrm{SPD}, R, \dots)$$

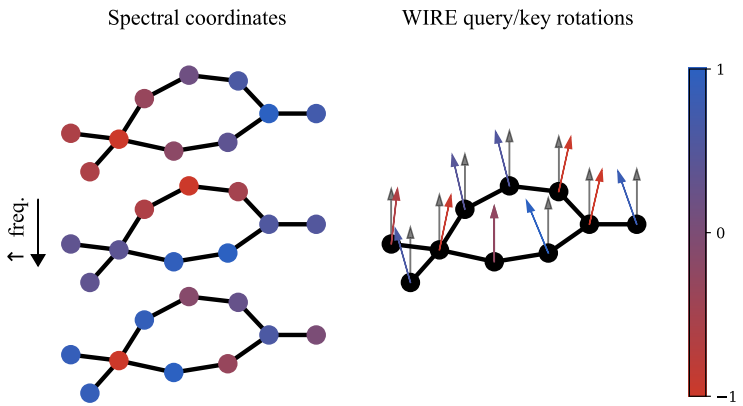
  Strong inductive bias, but instantiates $N \times N$ attention $\Rightarrow$ incompatible with linear attention

Idea: feed Laplacian features into RoPE $\Rightarrow$ structure as rotation, $\mathcal{O}(N)$ scaling.

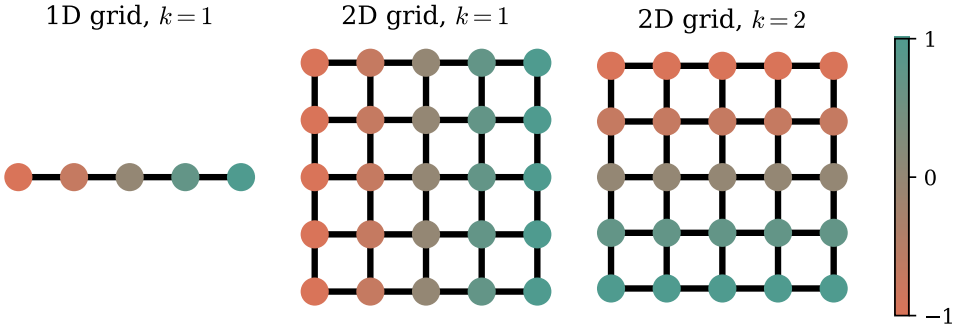
WIRE: Wave-Induced Rotary Encodings

Algorithm

1. Diagonalise Laplacian $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$
2. Spectral coordinates from lowest m eigenvectors: $\boldsymbol{r}_i = [\boldsymbol{u}_k[i]]_{k=0}^{m-1} \in \mathbb{R}^m$
3. Rotate $\boldsymbol{q}_i, \boldsymbol{k}_i$ by $\theta_n = \boldsymbol{\omega}_n^\top \boldsymbol{r}_i$; learnable $\{\boldsymbol{\omega}_n\}_{n=1}^{d/2} \subset \mathbb{R}^m$



Theory I: RoPE $\subset$ Wire on grids



- Path graph P_N: first nontrivial eigenvector

$$\boldsymbol{u}_1[i] = -\cos\left(\frac{\pi}{N}\left(i+\frac{1}{2}\right)\right)$$

  is monotone in token index
- $\boldsymbol{\omega}_n = [0, \omega_n, 0, \dots] \Rightarrow$ **classical RoPE** (LLMs)
- 2D grids factorise $\Rightarrow$ **ViT RoPE**

Theory II: WIRE depends on effective resistance

Normalised features $\boldsymbol{r}_i = [\boldsymbol{u}_k[i]/\sqrt{\lambda_k}]$, random $\boldsymbol{\omega}_n \sim \mathcal{N}(0, \omega \boldsymbol{I})$:

$$\mathbb{E}\Big[(\mathrm{RoPE}(\boldsymbol{r}_i)\boldsymbol{q}_i)^\top \mathrm{RoPE}(\boldsymbol{r}_j)\boldsymbol{k}_j\Big] = \boldsymbol{q}_i^\top \boldsymbol{k}_j \Big(1 - \tfrac{\omega^2}{2} R(i,j)\Big) + \mathcal{O}\Big(\omega^4\Big).$$

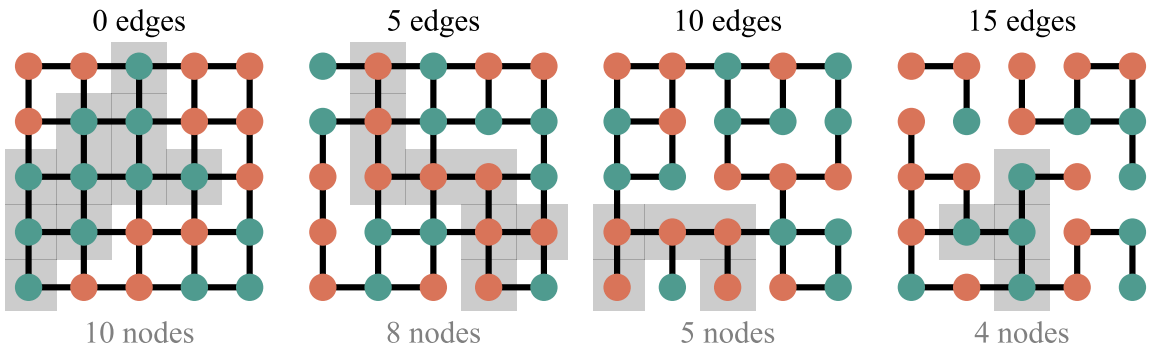
$R(i,j) = \boldsymbol{L}^\dagger_{ii} + \boldsymbol{L}^\dagger_{jj} - 2\boldsymbol{L}^\dagger_{ij}$: effective resistance, a metric on $\mathcal{N}$ that lower-bounds SPD.

Why this is remarkable

- Topological masking by structural distance
- **No $N \times N$ attention matrix required**
- Limit is **gauge invariant**

Synthetic: monochromatic subgraphs

Predict size of largest monochromatic connected subgraph on a coloured 5×5 grid with random edges deleted.



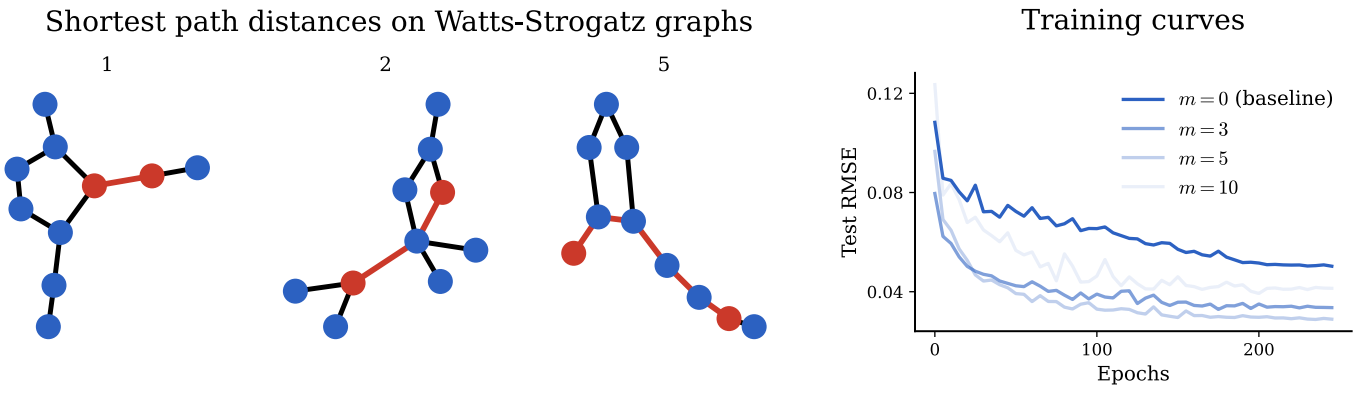
Edge deletion **interpolates grid** $\rightarrow$ **irregular topology**; stress-tests WIRE away from the ViT-RoPE limit.

	Test RMSE ↓ (deleted edges)			
m	0	5	10	15
0 (no WIRE)	0.060(1)	0.087(1)	0.081(1)	0.068(2)
3	**0.053(2)**	0.075(2)	0.072(3)	0.064(3)
5	0.057(2)	0.075(1)	0.070(2)	**0.056(4)**
10	0.055(2)	**0.068(5)**	**0.063(2)**	0.058(2)

Synthetic: shortest-path distance

Watts–Strogatz, $N = 10$, $k = 2$, $p = 0.6$; predict SPD between marked nodes.

Purely **topological** target; isolates WIRE's resistive-distance bias (Theorem II).



m	0 (base)	3	5	10
Test RMSE ↓	0.065(5)	0.048(6)	**0.038(6)**	0.045(4)

WIRE nearly **halves** the error.

Benchmarks: WIRE Performers close the gap

- Drop-in for GraphGPS + linear (Performer) attention
- Often matches $\mathcal{O}(N^2)$ softmax baselines

	Performer $\mathcal{O}(N)$		Transformer $\mathcal{O}(N^2)$
Dataset	Baseline	+ WIRE	Baseline
MNIST ↑	97.56(2)	**98.10(1)**	98.05(4)
CIFAR10 ↑	70.61(4)	71.15(3)	72.3(1)
PATTERN ↑	85.71(3)	**86.63(6)**	86.69(2)
CLUSTER ↑	76.90(3)	77.53(3)	78.02(6)
ogbg-molhiv ↑	0.776(2)	**0.785(2)**	0.788(1)
ogbg-molpcba ↑	0.238(3)	0.264(1)	0.291(3)
Peptides-struct ↓	0.2616(4)	0.2566(4)	0.2500(5)
MalNet-Tiny ↑	92.81(5)	**93.46(2)**	93.36(6)

MalNet-Tiny: WIRE Performer matches the quadratic Transformer; trains on a single T4 12 GB GPU.

Takeaway

WIRE: RoPE for graphs via Laplacian spectra

- Recovers standard RoPE on grids (LLMs, ViTs)
- Asymptotically downweights attention by **effective resistance**
- Stays $\mathcal{O}(N)$; fully compatible with linear attention
- $< 1\%$ of parameters; consistent gains across > 200 trained models

Paper: arXiv:2509.22259 Code: github.com/cederikhoefs/Graph-RoPE

Come find us at the poster!