k-Maximum Inner Product Attention for Graph Transformers and the Expressive Power of GraphGPS

Jonas De Schouwer[1,2] Haitz Sáez de Ocáriz Borde[2] Xiaowen Dong[2]

[1]Stanford University [2]University of Oxford

ICLR 2026 GRaM Workshop

Geometry-Grounded Representation Learning & Generative Modeling

Introduction

Graph transformers overcome limitations of traditional GNNs (oversquashing, long-range dependencies), but their **quadratic complexity** limits scalability. We introduce **k-MIP attention**, dynamically selecting the k most relevant keys per query.

1. We introduce **k-MIP self-attention**, which demonstrates **linear memory scaling** and up to a **10× speedup** over full attention (PyTorch). This enables processing graphs with >500k nodes on a single A100 GPU.
2. **Integration into GraphGPS** with an **expressiveness upper bound** via the S-SEG-WL test.
3. **Universal approximation:** k-MIP transformers can approximate any full-attention transformer to arbitrary precision.
4. **Competitive results** on LRGB, City-Networks, ShapeNet-Part, and S3DIS benchmarks.

Method: k-MIP Attention

k-MIP attention restricts attention weights for each query to the k keys with the **highest inner product scores**:

k-MIP Self-Attention

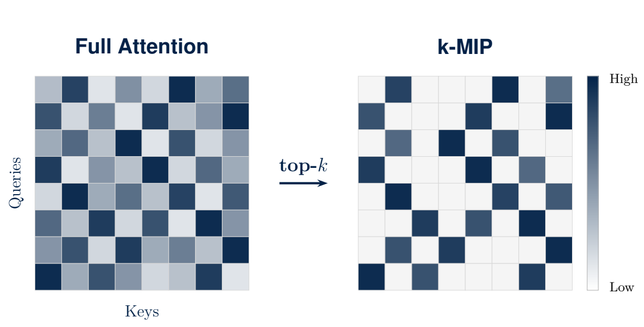
Given $\boldsymbol{X} \in \mathbb{R}^{N\times d}$, weight matrices $\boldsymbol{W}_Q^h, \boldsymbol{W}_K^h \in \mathbb{R}^{d\times d_K}$, $\boldsymbol{W}_V^h \in \mathbb{R}^{d\times d_V}$:

$$\boldsymbol{A}^h = \texttt{softmax}\left(\frac{1}{\sqrt{d_K}}\mathcal{T}_k\left(\boldsymbol{X}\boldsymbol{W}_Q^h(\boldsymbol{X}\boldsymbol{W}_K^h)^\top\right)\right) \tag{1}$$

$$\texttt{k-MIP}(\boldsymbol{X}) = \textstyle\sum_{h=1}^H \boldsymbol{A}^h\boldsymbol{X}\boldsymbol{W}_V^h\boldsymbol{W}_O^h \tag{2}$$

$\mathcal{T}_k$: retains k largest per row, rest set to $-\infty$.

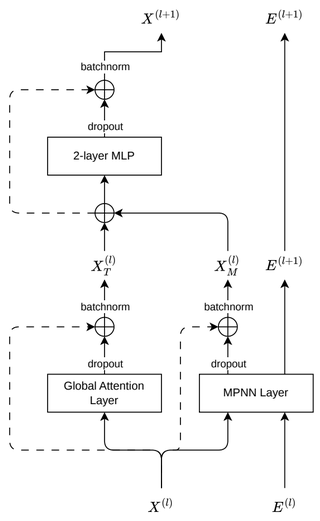
The intermediate matrix is stored as a **symbolic matrix** (computed lazily in GPU registers, never materialized), yielding **linear memory**. Top-k indices are reused in backpropagation, making the **backward pass virtually free**.



We integrate k-MIP attention into the **GraphGPS** framework [Rampášek et al., 2022] to create a graph transformer.

Each **GraphGPS** layer combines:

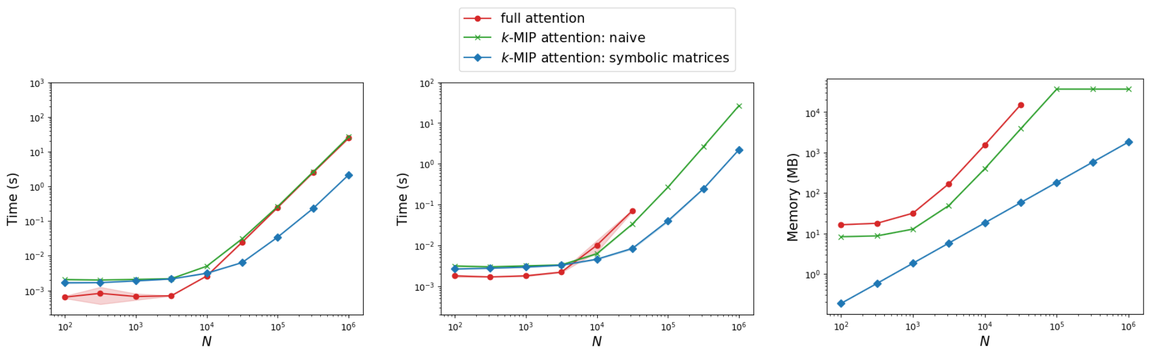
- **MPNN** (local message passing)
- **Global Attention** (k-MIP: any-to-any exchange)
- **2-layer MLP** (per-node transform)



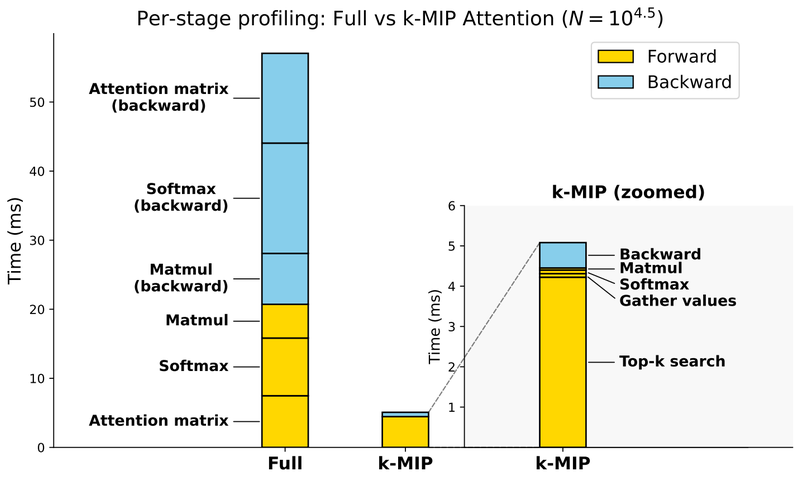
One GraphGPS layer
(based on [Rampášek et al., 2022])

Computational Efficiency

We compare runtime and memory of full attention vs. k-MIP attention, varying $N \in \{10^{i/2} \mid i = 4, \ldots, 12\}$ with $d_K = 10$, $k = 10$, on a single 40GB A100 GPU (mean ± std over 5 runs).



(a) Time (forward) (b) Time (forward+backward) (c) Peak memory (forward+backward)



(d) Per-stage runtime breakdown at $N = 10^{4.5}$

Results:

- **12.4**× inference (single fwd pass) speedup at $N = 10^6$; **8.5**× training (fwd+bwd) speedup at $N = 10^{4.5}$.
- Full attention OOM at $N \geq 10^5$; k-MIP scales to $N = 10^7$ via **linear memory**.
- Top-k search dominates the forward pass; **backward pass is nearly free** (top-k indices are reused).

Expressive Power of GraphGPS

We upper-bound GraphGPS's graph-distinguishing power via the S-SEG-WL test [Zhu et al., 2023].

Theorem 1

Let $\mathcal{A}$ be a GraphGPS instance enhancing node features with $\nu(\boldsymbol{A})$ and edge features with $\mu(\boldsymbol{A})$. Then $\mathcal{A}$ can only distinguish graphs also distinguishable by the S-SEG-WL test where $S = (f_A, f_R)$:

$$f_A(v, G) = \nu(\boldsymbol{A})_v \tag{3}$$

$$f_R(v, u, G) = \begin{cases} (0,\ \mathbb{1}_{u=v},\ \mathbf{0}_{d_e},\ \mathbf{0}_{d_{PE}}) & \text{if } (u,v) \notin E \\ (1,\ \mathbb{1}_{u=v},\ \boldsymbol{E}_{uv},\ \mu(\boldsymbol{A})_{uv}) & \text{if } (u,v) \in E \end{cases} \tag{4}$$

Color set $\mathcal{C} = \mathbb{R}^{d_{PE}} \cup \{0,1\}^2 \times \mathbb{R}^{d_e} \times \mathbb{R}^{d_{PE}} \cup \mathbb{R}^d \cup \mathbb{R}^{d_e}$.

Three Consequences:

1. Comparison to 1-WL. Without positional encodings (i.e. identical node and edge features), the S-SEG-WL test degenerates to the trivial form recovering 1-WL. GraphGPS is **upper bounded by 1-WL** in this setting.

2. More expressive encodings help. More discriminative encodings raise the expressiveness bound. With **Laplacian PE**, GraphGPS is **strictly more expressive than 1-WL**.

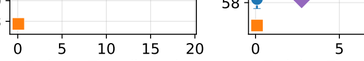
3. Origin of expressive power. Super-1-WL expressiveness claimed by previous graph transformers comes from **positional/structural encodings**, *not* the Transformer architecture. An expressive GNN with equivalent augmentations achieves the same distinguishing power.

Universal Approximation of Full-Attention Transformers

We prove k-MIP transformers can approximate any full-attention transformer to arbitrary precision. First, we define a unified framework.

Definition 1 (Generalized Transformer Block)

A *transformer block* $t_{\mathcal{A}}^{h,m,r} : \mathbb{R}^{N\times d} \to \mathbb{R}^{N\times d}$ implements:



$\mathcal{A}^{h,m}$: an attention layer, e.g. full attention or k-MIP attention (h heads, dim m). MLP^r: 2-layer ReLU MLP (width r).

Definition 2 (Class $\mathcal{T}_{\mathcal{A}}^{h,m,r}$)

$\mathcal{T}_{\mathcal{A}}^{h,m,r}$ is the class of transformers using attention $\mathcal{A}$, each a composition of blocks $t_{\mathcal{A}}^{h,m,r}$:

$$\mathcal{T}_{\mathcal{A}}^{h,m,r} := \left\{T_{\mathcal{A}} : \mathbb{R}^{N\times d} \to \mathbb{R}^{N\times d} \mid T_{\mathcal{A}} = t_{\mathcal{A}}^{h,m,r} \circ \cdots \circ t_{\mathcal{A}}^{h,m,r}\right\}$$

Theorem 2 (k-MIP Approximation Theorem)

For any $T_{full} \in \mathcal{T}_{full}^{h,m,r}$, any $\epsilon > 0$, any $p \in [1, \infty)$, and any compact $U \subseteq \mathbb{R}^{N\times d}$, there exists $T_{\text{k-MIP}} \in \mathcal{T}_{\text{k-MIP}}^{2,1,4}$ s.t.

$$\left(\int_U \|T_{\text{k-MIP}}(\boldsymbol{X}) - T_{full}(\boldsymbol{X})\|_p^p \, d\boldsymbol{X}\right)^{1/p} < \epsilon$$

Key insight: The sparsification of the k-MIP attention mechanism does not reduce the expressive power of the model class.

Note: This theorem does not guarantee that individual k-MIP layers can approximate full-attention layers, but only that the composition of many such layers can do so.

Experiments

We compare k-MIP against alternative scalable attention mechanisms, integrating each into GraphGPS for a fair comparison, on the City-Networks, LRGB, and ShapeNet-Part/S3DIS benchmarks. We also include MPNN baselines.

Long Range Graph Benchmark

	VOC-SP (F1↑)	COCO-SP (F1↑)	Pept-F (AP↑)	Pept-S (MAE↓)
GCN	20.78±.31	13.38±.07	**68.60±.50**	**.2460±.0007**
GINE	27.18±.54	21.25±.09	66.21±.67	.2473±.0017
GAT	27.15±.49	18.86±.11	67.87±.56	.2488±.0018
GatedGCN	38.80±.40	29.22±.18	67.65±.47	.2477±.0009
GPS+BigBird	38.75±.64	35.25±.27	64.83±.73	.2566±.0019
GPS+Performer	37.92±1.13	27.70±.27	66.06±.64	.2643±.0008
GPS+Transformer	**44.40±.65**	**38.84±.55**	65.34±.91	.2509±.0014
Exphormer	39.75±.37	34.55±.09	65.27±.43	.2481±.0007
GPS+k-MIP	39.69±.92	35.56±.45	66.27±.44	.2562±.0037

Large-Scale Point Cloud Datasets

	ShapeNet (F1↑)	S3DIS (mIoU↑)
GCN	60.18±.04	39.98±1.47
GINE	64.57±.35	44.16±.62
GAT	63.01±.17	44.24±1.14
GatedGCN	76.20±.32	63.71±1.28
GPS+BigBird	79.65±.98	67.92±.91
GPS+Performer	77.36±1.23	60.83±.56
GPS+Transformer	OOM	OOM
Exphormer	82.62±.31	**68.37±.23**
GPS+k-MIP	**82.68±.64**	67.99±1.51

*__First__, second, third best results highlighted.

Key findings:

- GPS+k-MIP scales to the London dataset with **569k nodes** on a single GPU. GPS+Transformer, Exphormer, and GPS+BigBird all return OOM on London.
- **MPNNs suffice for the City-Networks benchmark**, and outperform all graph transformers.
- On all benchmarks, GPS+k-MIP consistently ranks **among the top-performing scalable graph transformers**.

Key Takeaways

- k-MIP attention achieves **linear memory** and **10× speedup**, scaling to **>500k-node** graphs on a single GPU.
- k-MIP transformers are **universal approximators** of full-attention transformers.
- Graph transformer expressiveness comes from **positional/structural encodings**, not the attention architecture itself. This can be quantified via the S-SEG-WL test.

References

Ladislav Rampášek, Michael Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. Recipe for a general, powerful, scalable graph transformer. *Advances in Neural Information Processing Systems*, 35:14501–14515, 2022.

Wenhao Zhu, Tianyu Wen, Guojie Song, Liang Wang, and Bo Zheng. On structural expressive power of graph transformers. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3628–3637, 2023.