The Devil is in the Spectrum

TRSP: Topologically Regularized Side-Path

Mitigating Representation Collapse in LLMs

ICML 2026 · Presentation

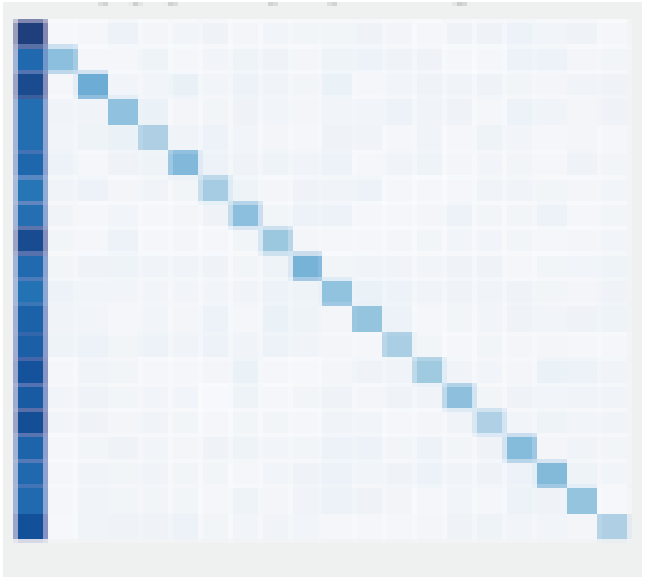
TRSP: Topologically Regularized Side-Path

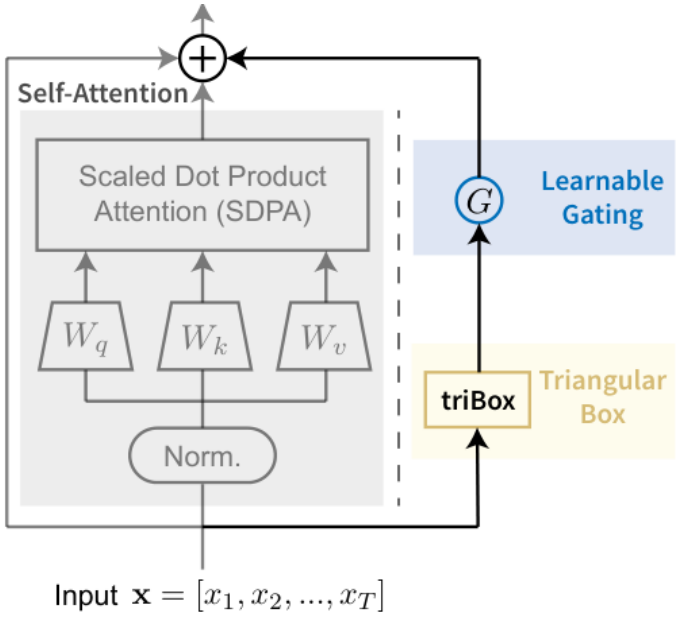
TL;DR

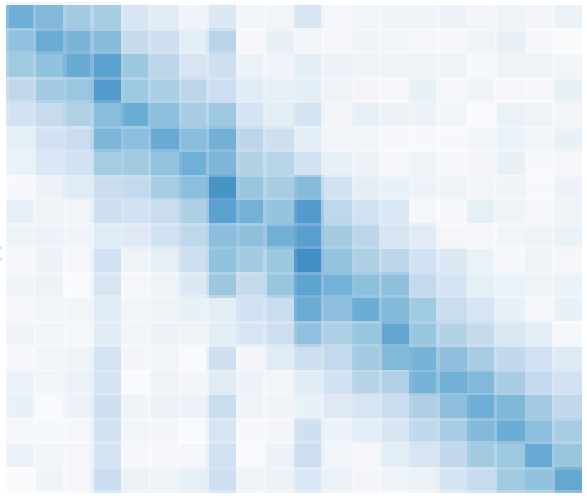
The attention matrix of large models often allocates excessive attention to the 0-th token, leading to information loss

We attach a specially designed side-path

The attention matrix returns to normal, and performance improves



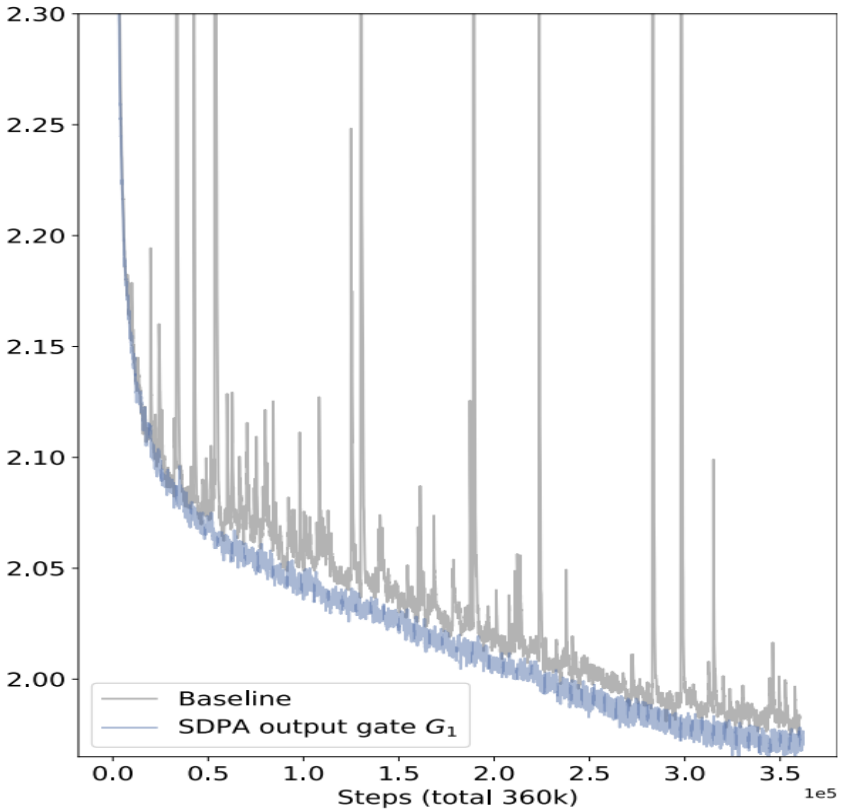




TRSP: Topologically Regularized Side-Path

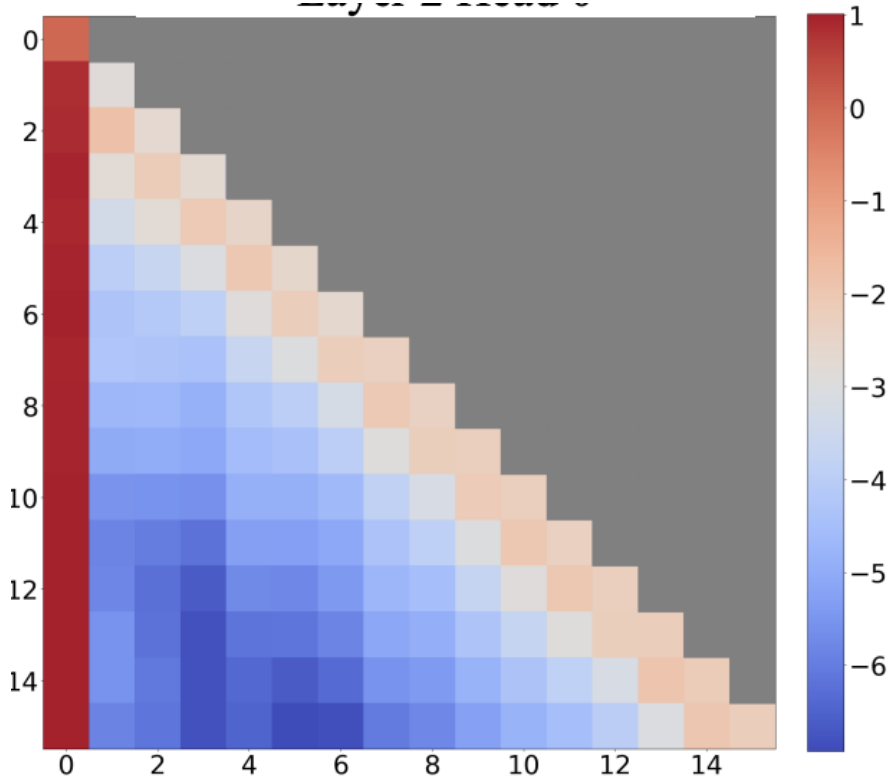
Background

We observe that large language models in long-context scenarios, experience loss spikes during training:



NIPS 2025

Large language models face representation collapse in long-context scenarios: token features degenerate into a low-dimensional, uninformative space.

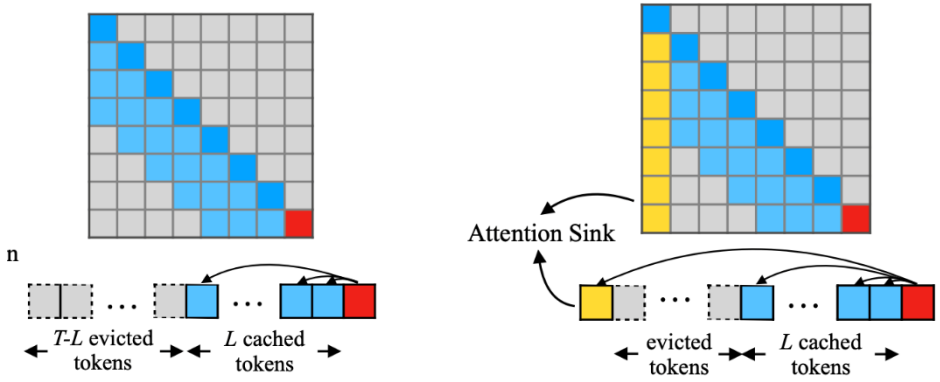


NIPS 2025

TRSP: Topologically Regularized Side-Path

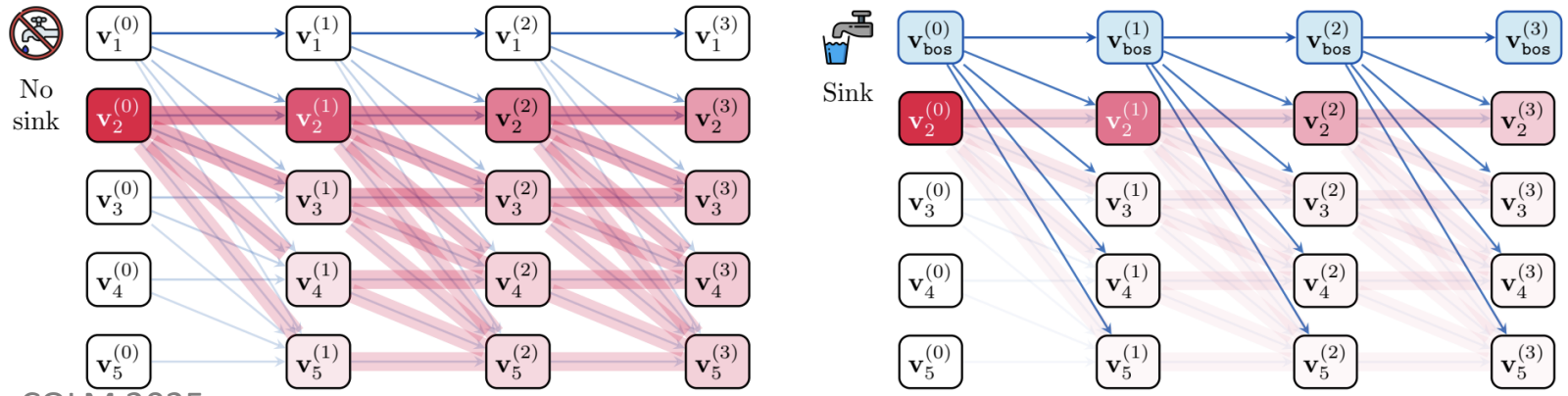
Background

Sliding Window Attention does not suffer from this issue, but yields worse performance; adding the sink token back actually improves it:



ICLR 2024

Forcibly removing the attention sink leads to the over-mixing problem:



COLM 2025

TRSP: Topologically Regularized Side-Path

Background & Math

Analyzing the cause of attention sink formation:

Define Loss as L, Output is $O_i = \sum_{j=1}^{N} p_j * V_j$， pj is attention score， Vj is attention Value， gradient $g = -\frac{\partial L}{\partial O_i}$

For any pj, $\frac{\partial L}{\partial p_j} = \left(\frac{\partial L}{\partial O}\right)^{\top} \frac{\partial O}{\partial p_j}$， notice $\frac{\partial O}{\partial p_j} = V_j$,so we have：

$$\frac{\partial L}{\partial p_j} = \left(\frac{\partial L}{\partial O}\right)^{\top} V_j = -g^{\top} V_j$$

Note that backpropagation is a local minimization of the first-order Taylor expansion. Taking a first-order Taylor expansion of L with respect to the attention weights pj, we have：

$$L(p) \approx L(p_{current}) + \sum_{j=1}^{N} \frac{\partial L}{\partial p_j} p_j$$

Substitute the partial derivatives， we have $\min\{L(p)\} = \min_p \sum_{j=1}^{N} (-g^{\top} V_j) p_j + L(p_{current})$， so the aim is $\max_p \sum_{j=1}^{N} (g^{\top} V_j) p_j$

Now let $c_j = g^{\top} V_j$， so our aim is：

$$\text{Maximize} \quad \sum_{j=1}^{N} c_j p_j$$

$$\text{Subject to} \quad \sum_{j=1}^{N} p_j = 1$$

$$p_j \geq 0 \quad \forall j$$

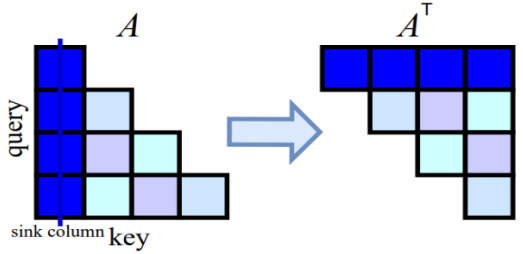
Note that this is a simplex with n-1 degrees of freedom， $p \sim \Delta^{n-1}$, its extrema must occur at the vertices, we have

$$p_{\mathrm{im}} \to 1, p_{ij} \to 0, j \neq \mathrm{m}$$

Now $p_{ij}/p_{im} = exp(q_i * k_j)/exp(q_i * k_m) = exp\left(q_i * \left(k_j - \mathrm{k_m}\right)\right) \to 0$, means

$$q_i * \left(k_j - k_{\mathrm{m}}\right) \to -\infty$$

$||q_i||$ or $||k_i - k_m||$tends towards infinity, forming a sink.



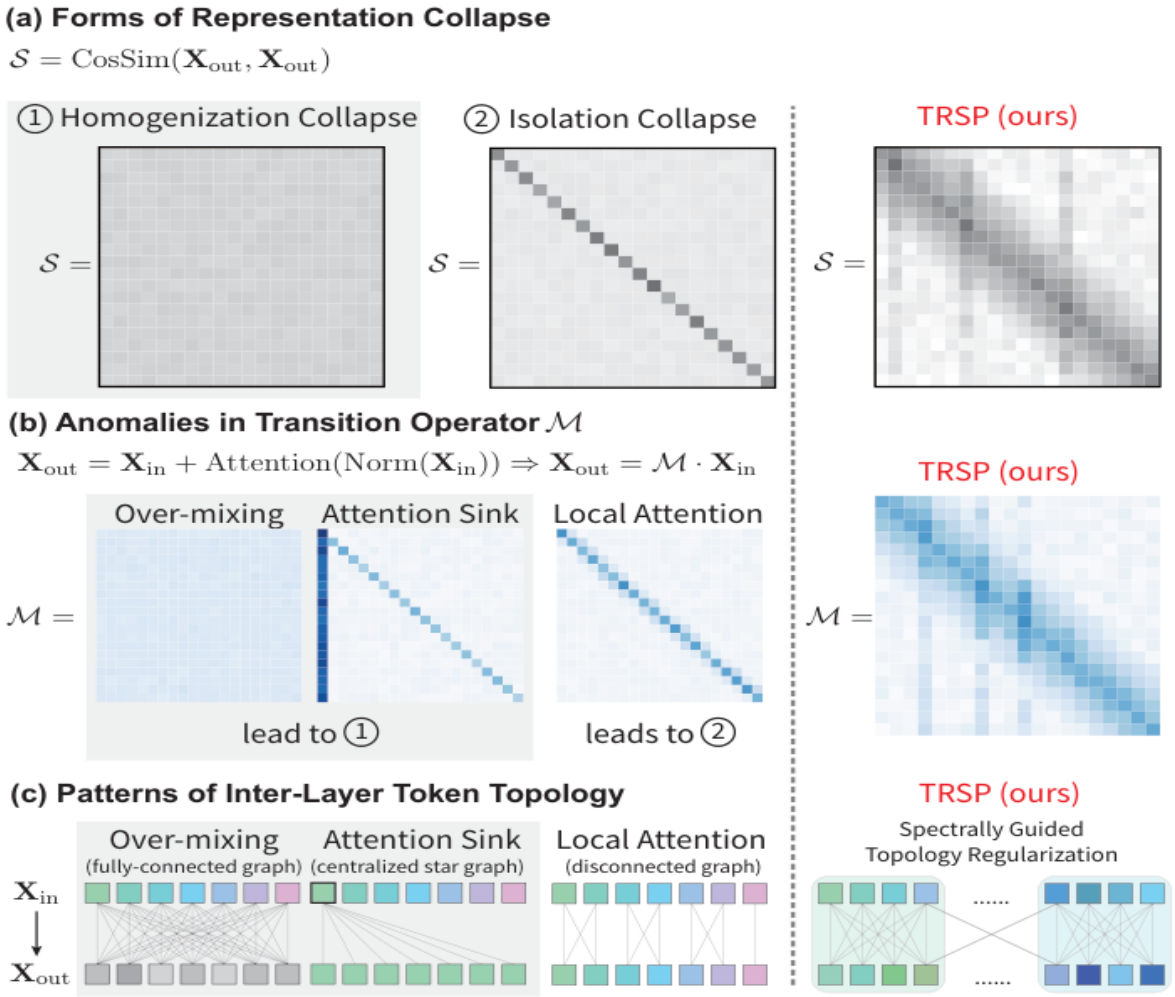
Finally, we note that due to the causal mask, the 0-th column is naturally the largest under random initialization, so all excess attention is inherently allocated to the 0-th token

TRSP: Topologically Regularized Side-Path

Motivation

We find that existing solutions tend to drift toward two pathological extremes — "homogenization collapse" (over-mixing and attention sinking leading to rank deficit) and "isolation collapse" (localized attention causing contextual disconnection).

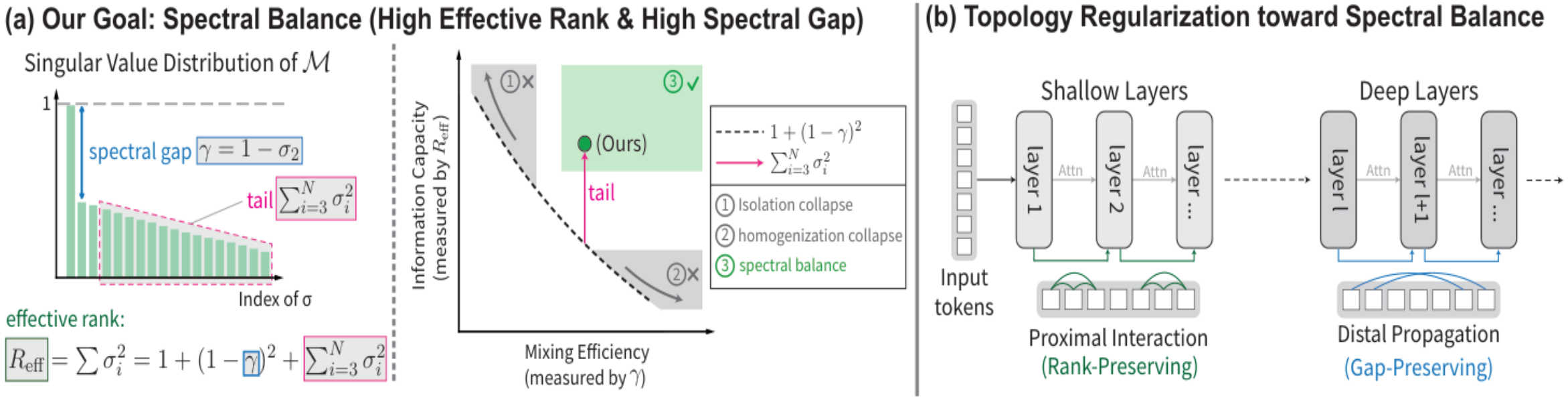
Further analysis reveals that these two extremes correspond to two types of problems: "lack of information" and "lack of information interaction"



TRSP: Topologically Regularized Side-Path

Intuition

Therefore, starting from linear algebra, we conduct a singular value spectrum analysis of the transition operator M:
The issue of information loss—information capacity (effective rank R_{eff}).
The issue of information interaction—mixing efficiency (spectral gap γ).
This leads to the following approach: regularizing the topological structure—preserving rank through shallow, short-range interactions, and preserving the spectral gap through deep, long-range propagation.



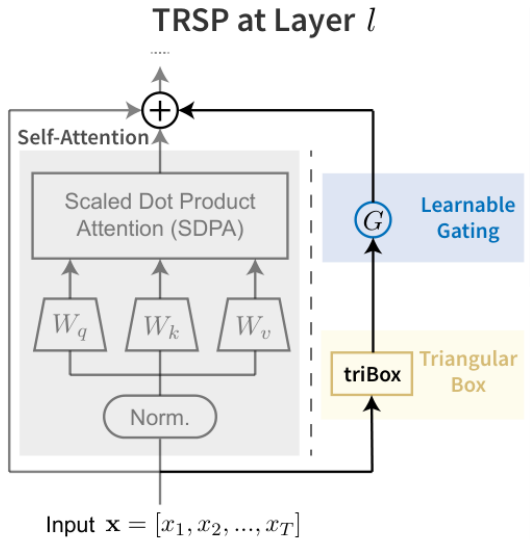
TRSP: Topologically Regularized Side-Path

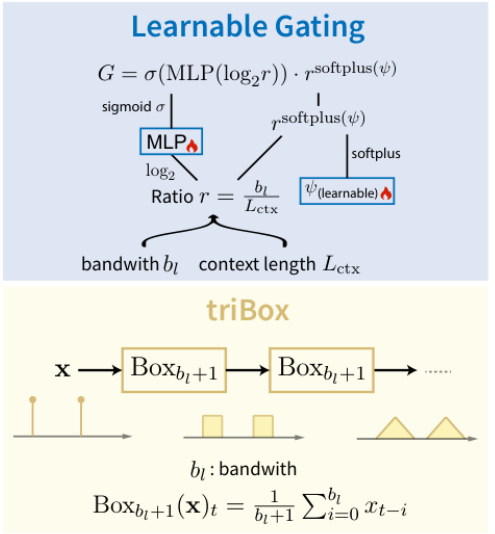
Framework

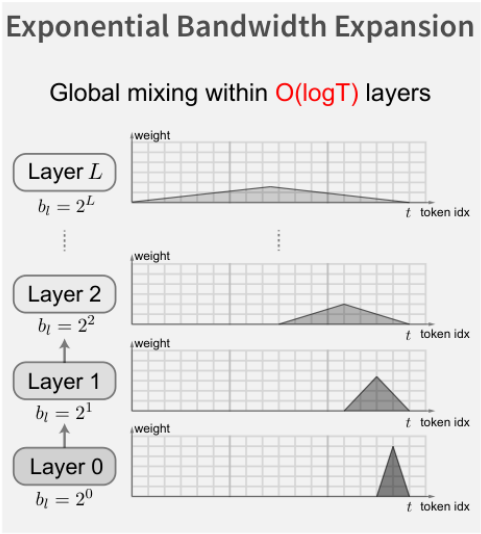
1. triBox — a parameter-free triangular box filter (a cascaded causal box filter based on O(T) prefix sums), which introduces a topologically regularized mixing operation in parallel alongside the attention mechanism.

2. Layered bandwidth—the window size $b_\ell = \min(2^\ell, L_{\text{ctx}})$ grows exponentially with depth, thereby constructing a Cayley graph topology with a constant spectral gap.

3. Long-context gating—involves only about 50 learnable parameters; the gating value $\gamma_\ell$depends solely on $r = b_\ell/T$, thereby ensuring that spectral properties are asymptotically preserved as $T \to \infty$.







TRSP: Topologically Regularized Side-Path

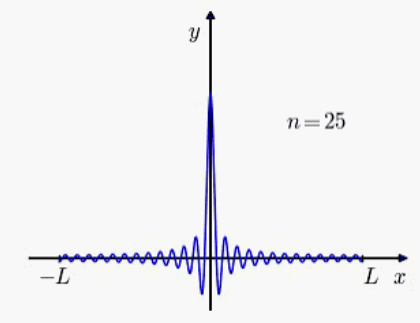
Framework

1. triBox guarantees effective rank

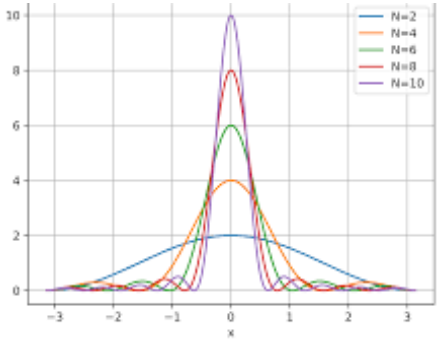
Effective Rank $R_{eff} = \Sigma\, \sigma_i^2 / \sigma_1^2$

Therefore, the large mass of the sink token is distributed to nearby tokens via TriBox on the side-path

First-order mean Box (Dirichlet kernel): Second-order mean TriBox = box∘box (Fejér kernel):



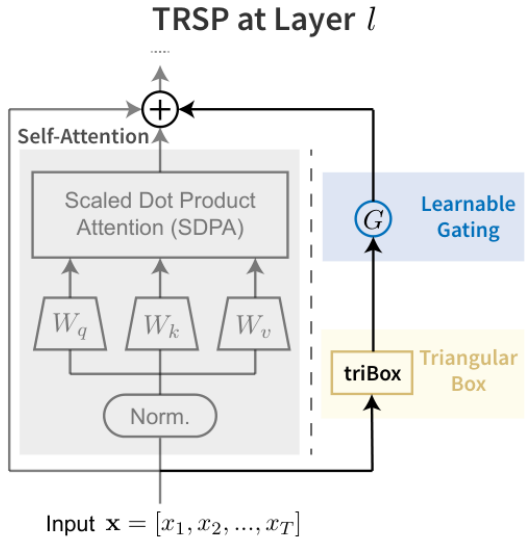
Contains both positive and negative values, which may cancel out or form new sink tokens



Non-negative everywhere, no cancellation

So we use triBox:

Accumulation only on the side-path preserves information, allowing non-sink information needed by the next layer to bypass the sink token, without affecting the original attention score computation

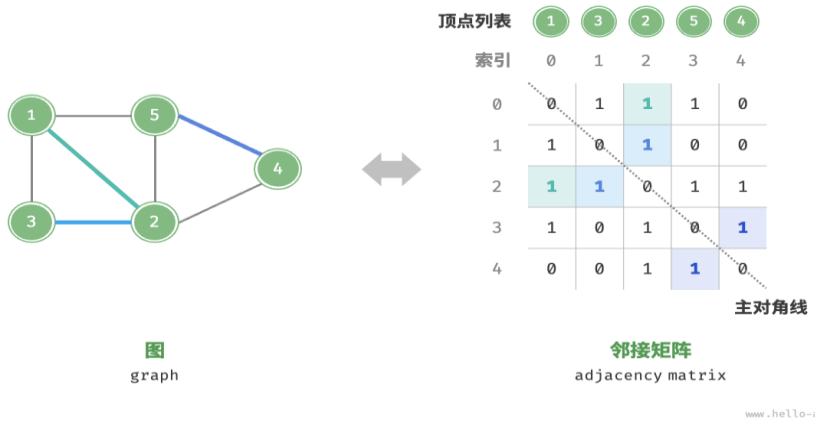


TRSP: Topologically Regularized Side-Path

Framework

2. Layer-wise bandwidth guarantees spectral gap

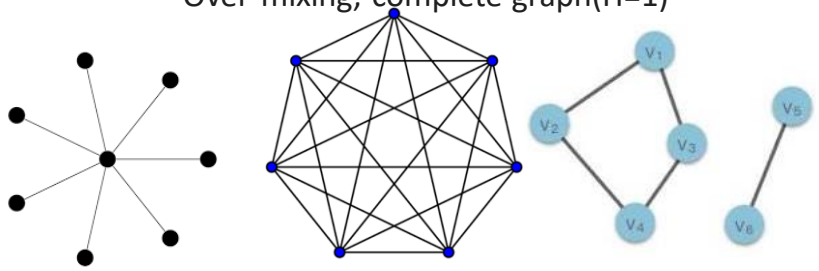
A priori: The diameter of the graph (H, the distance between the two most distant nodes in terms of the shortest path) is inversely proportional to the spectral gap (γ) of the adjacency matrix:$H * \gamma = C(\text{Constant})$



Binary lifting edge connections: For any pair of nodes $(x, y)$with distance d, represent $d$in binary $D_{(2)}$; clearly, at most $\log_2 D$bits are 1, while the rest are 0. By adding an edge corresponding to each binary bit, the nodes (x, y)become mutually reachable via at most $\log_2 D$ edges; in this case, the graph diameter is $H = \log_2 n$.

Under homogenization collapse, the graph diameter is 1 (over-mixing) or 2 (attention sink), spectral gap → 1; Under isolation collapse, some nodes become unreachable, diameter → ∞, spectral gap → 0; We construct doubling edges: the l-th layer connects nodes at distance 2^l, ensuring diameter = log n, spectral gap → 1/log n

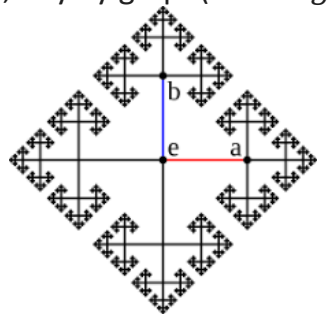
Over-mixing, complete graph(H=1)



Attention sink, star graph(H=2)

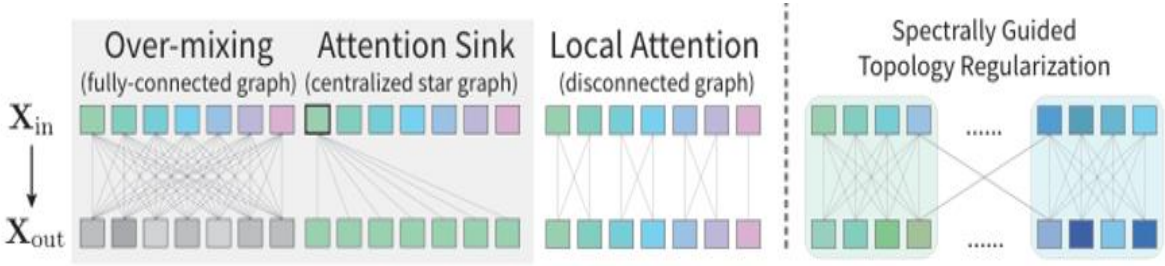
Isolation, disconnected graph(H= ∞)

Ours, Cayley graph($H = log_2 n$)



Assuming the depth is L=logn, so γ tends to:

$$\lim_{n\to\infty}\left(1+\frac{C}{\ln(n)}\right)^{\ln(n)} = e^{C} > 0$$



TRSP: Topologically Regularized Side-Path

Result

TRSP boosts general capability AND extrapolates to 8× the training context.

General Capability (Llama-3.2-1B)

MMLU: 35.1 → 37.8 (+2.7)
HellaSwag: 25.9 → 29.3 (+3.3)
Stable Rank: 1.40 → 1.93 (+38%)
Extra params: only ≈ 50 (O(1) cost)

Long-Context (RULER, post-train)

Train @ 4K, eval at 8K / 16K / 32K
8K : 65.7 (LoRA 63.8, Gated 45.1)
16K : 62.8 (LoRA 60.5, Gated 45.1)
32K : 60.4 (LoRA 57.5, Gated 40.3)

NoLiMa Extrapolation (pre-train @ 1K, 109M)

Length : 1K 2K 4K 8K
Transformer : 1.000 0.992 0.804 0.238
Gated Attn : 1.000 0.984 0.793 0.336
Diff Trans. : 1.000 1.000 0.902 0.539
TRSP (Ours) : 1.000 1.000 0.988 0.832 ← +29 pts.

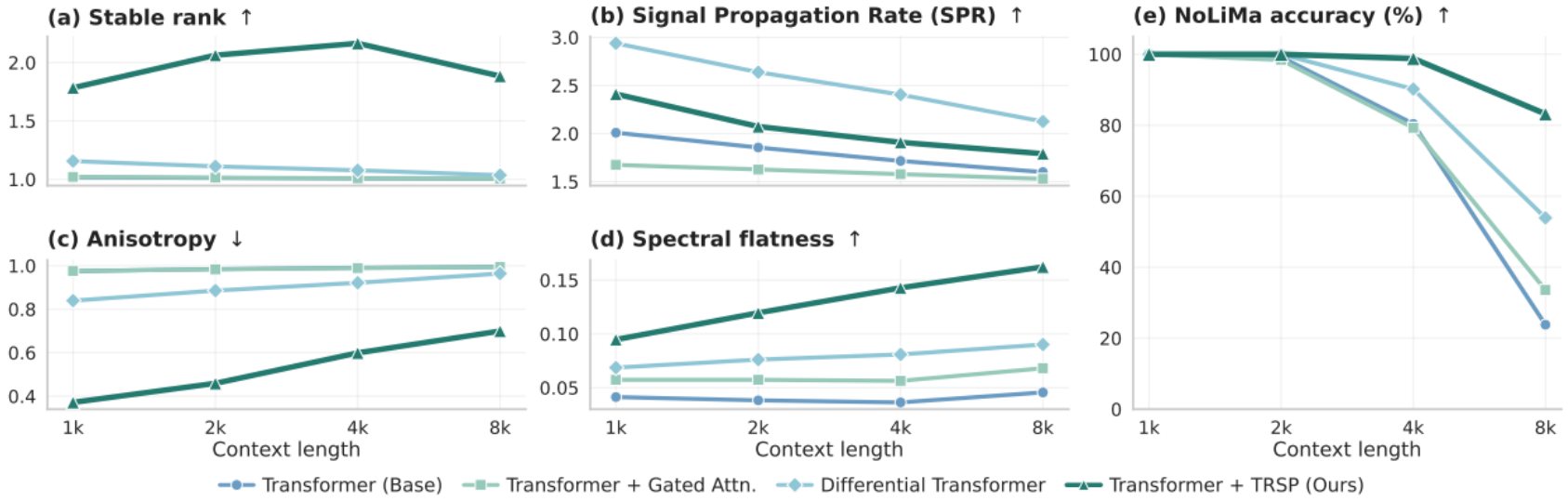
Method	Extra Params	MMLU Acc ↑	HellaSwag Acc ↑	HellaSwag PPL ↓	Spectral Diagnostics (Avg.) SPR ↑	Rank ↑	Anisotropy ↓	Flatness ↑
Llama-3.2-1B (Raw)	-	35.09	25.94	**1.67**	**1.60**	1.40	0.21	0.42
LoRA	5.6M	36.01	27.97	3.58	1.51	1.36	0.22	0.42
LoRA + Gated Attention (Qiu et al., 2026)	5.6M + 67.2M	23.19	24.79	24.94	1.59	1.44	0.19	0.44
LoRA + TRSP (Ours)	5.6M + 50	**37.76**	**29.26**	3.05	1.59	**1.93**	**0.18**	**0.69**

Bench.	Setting	Method	Train	(Extra) Params	2×	4×	8×
RULER	Fine-tuning	Llama-3.2-1B (Raw)	4K	–	63.85	59.23	59.55
		LoRA	4K	5.6M	63.82	60.50	57.51
		LoRA + Gated Attention	4K	5.6M+67.2M	45.06	45.06	40.29
		LoRA + TRSP (Ours)	4K	5.6M+50	**65.68**	**62.78**	**60.38**
NoLiMa	From-scratch	Transformer (Base)	1K	109.8M	99.2	80.4	23.8
		Transformer + Gated Attention	1K	113.5M	98.4	79.3	33.6
		Differential Transformer	1K	109.8M	**100.0**	90.2	53.9
		Transformer + TRSP (Ours)	1K	109.8M	**100.0**	**98.8**	**83.2**

TRSP: Topologically Regularized Side-Path

Justification

Why does it work? — Spectral health directly translates into accuracy retention.



Spectral evidence (vs. context length)

- Stable Rank: TRSP stays high; baselines and Diff-Trans decay → richer feature space.
- SPR: TRSP keeps SPR > 1.6; baselines become over-damped.
- Anisotropy: TRSP stays low; baselines collapse into a narrow cone.
- Flatness: TRSP grows with T → operator stays well-conditioned.
- Conclusion: TRSP doesn't 'learn' to read long text — it structurally prevents the collapse of M.

Fig 5. Accuracy retention 1K → 8K (NoLiMa). TRSP retains 83% at 8× train length.

Thanks