Where to Add PDE Diffusion in Transformers

Anonymous ACL submission

Abstract

Transformers route information globally via self-attention but lack an explicit local geometric prior along the sequence axis, so hybrid designs often treat *where* to add locality as an empirical choice. We study a deterministic *PDE diffusion* layer (one explicit Euler step of 1D heat-equation smoothing using a discrete Neumann Laplacian; not a generative diffusion model) and ask: *where should it be inserted relative to attention?* Our key insight is that PDE diffusion and attention generally *do not commute*, so the same local operator can help or hurt depending on placement. We develop an operator-theoretic framework that separates (i) intrinsic stability of the PDE diffusion subsystem, (ii) compositional bounds explaining insertion-position sensitivity, and (iii) a non-commutativity-based diagnosis of post-attention "double-mixing". On Long Range Arena, the predicted best placement (After-Embedding) improves average accuracy by +4.1 points, while the predicted worst placement (After-Attention) degrades it by −2.5 points. Overall, we provide a principled template for reasoning about local–global composition in Transformers.

1 Introduction

Transformers have become the dominant architecture for sequence modeling, largely due to self-attention's ability to perform content-based global routing (Vaswani et al., 2017). However, standard Transformers lack an explicit *local geometric prior* along the sequence dimension: they do not directly encourage neighboring token representations to vary smoothly or preserve local coherence. This gap has motivated hybrid designs that inject locality through convolutional layers (Wu et al., 2019; Dai et al., 2020), sparse patterns (Beltagy et al., 2020), or other structured inductive biases. Yet a surprisingly basic question remains under-theorized: *where should a local operator be placed relative to attention?* In practice, placement is often treated as a hyperparameter and tuned by validation performance. In this work, we ask whether placement can be predicted from structure rather than guessed from sweeps: **can operator theory explain when local smoothing helps and when it hurts?**

Our central observation is simple but consequential: local smoothing and content-based global routing are, in general, *non-commuting operations*. Let $\mathcal{S}$ denote a local PDE diffusion (smoothing) operator and $\mathcal{A}$ an attention-based routing operator. Typically, $\mathcal{S} \circ \mathcal{A} \neq \mathcal{A} \circ \mathcal{S}$, and the commutator $[\mathcal{S}, \mathcal{A}] = \mathcal{S}\mathcal{A} - \mathcal{A}\mathcal{S}$ captures this incompatibility. This has direct architectural implications: applying PDE diffusion before attention can pre-regularize a smooth embedding space and provide a cleaner manifold for routing; applying PDE diffusion after attention can instead blur structures just created by non-local mixing, inducing a harmful *double-mixing* effect.

To formalize these intuitions, we develop a three-layer operator-theoretic framework. **Layer I** (unconditional guarantees) establishes that PDE diffusion under a strict spectral budget ($\alpha < 1/2$) is a certified non-expansive local smoother with monotone energy dissipation. **Layer II** (conditional structural insights) derives perturbation propagation bounds showing that the effect of inserting PDE diffusion depends on two factors: the Dirichlet energy $\mathcal{E}(Z_j)$ at the insertion point and the downstream Lipschitz amplification. This analysis relies on a local Lipschitz assumption that is not globally verified for self-attention (Kim et al., 2021); we therefore present Layer II as

a *structural perturbation analysis* that identifies the governing factors, rather than a tight quantitative guarantee. **Layer III** (diagnostic heuristics) formalizes diffusion–attention non-commutativity through the commutator $[\mathcal{S}, \mathcal{A}](X)$, providing testable design guidelines.

Empirically, we validate the framework on the Long Range Arena benchmark (Tay et al., 2023), systematically evaluating seven insertion positions. The predicted best placement (After-Embedding, P1) improves average accuracy by +4.1 points, while the predicted worst (After-Attention, P7) degrades it by −2.5 points. Importantly, we provide **direct diagnostic measurements** of all three theoretical quantities—Dirichlet energy profiles, commutator norms, and empirical Lipschitz constants—on trained baseline models across all five tasks (Section 5.4). These measurements reveal that $\mathcal{E}(Z_j)$ is lowest at the embedding layer and increases by orders of magnitude through depth, that commutator norms decrease with depth, and that empirical Lipschitz constants are moderate (1.0–59.4 per layer). Together, these findings confirm that the composite perturbation bound is minimized at P1, providing the missing empirical bridge between theory and experiments.

Beyond the specific PDE diffusion operator studied here, we demonstrate how operator-theoretic reasoning can guide the design of hybrid architectures, offering a transferable methodology applicable to other local–global composition problems. The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 introduces the operator-theoretic framework. Section 4 presents the three-layer theoretical analysis. Section 5 provides empirical validation, including the new diagnostic measurements. Section 6 concludes.

2 Related Work

Our work sits at the intersection of (i) long-sequence Transformers, (ii) locality-inducing inductive biases, and (iii) stability- and operator-motivated architectural analysis. We position our contribution along three axes: we *do not* change attention's routing rule or complexity (unlike efficient-attention variants), we use PDE diffusion as a *fixed-form, spectrally-budgeted* local prior (not to be confused with generative diffusion models), and we provide a *structural* explanation for insertion-position sensitivity (often treated empirically in hybrid designs).

2.1 Long-Sequence Transformers

The quadratic complexity of self-attention has motivated extensive work on efficient long-sequence Transformers. Representative directions include sparse patterns (e.g., Longformer (Beltagy et al., 2020), BigBird (Zaheer et al., 2020), Sparse Transformers (Child et al., 2019), Reformer (Kitaev et al., 2020)), low-rank or kernelized approximations (e.g., Performer (Choromanski et al., 2020), Linear Transformers (Katharopoulos et al., 2020), Linformer (Wang et al., 2020), Nyströmformer (Xiong et al., 2021)), and memory-optimized exact attention (e.g., FlashAttention (Dao et al., 2022)). LRA (Tay et al., 2023, 2020) further highlights that efficiency improvements do not automatically translate into robust long-context generalization.

Our work is orthogonal: we keep attention intact and instead study how to inject an explicit *local geometric prior* via a lightweight PDE diffusion operator. Because the PDE diffusion layer is linear-time, it can be combined with sparse/linear attention mechanisms as a complementary design component.

2.2 Diffusion, PDEs, and Neural Operators

PDE- and diffusion-inspired views have a long history in deep learning, including stability-aware architectures via dynamical-systems discretizations (Haber and Ruthotto, 2017; Ruthotto and Haber, 2018) and classical diffusion for denoising/smoothing (Weickert, 1998). More recently, neural operator learning aims to approximate families of PDE solution operators (Li et al., 2020), and generative diffusion models treat diffusion as a stochastic generative process (Ho et al., 2020).

Our use of PDE diffusion differs in both goal and modeling choice. We use a *fixed-form* discrete Laplacian with a strict spectral budget to inject a local smoothness prior with unconditional stability guarantees, rather than learning an operator approximator or modeling a generative diffusion process. The resulting module

is a deterministic architectural regularizer that targets the sequence dimension and integrates naturally with Transformers.

2.3 Spectral Stability and Regularization in Deep Networks

Stability and spectral control have been recognized as central issues in deep learning, particularly for very deep or long-context models. Spectral normalization (Miyato et al., 2018) and Lipschitz-constrained architectures aim to control gradient explosion and improve robustness. Layer normalization (Ba et al., 2016) and residual connections (He et al., 2015) provide fundamental stability mechanisms in deep networks. In Transformers, prior work has analyzed the instability of attention maps and the amplification of high-frequency components across layers.

Our work contributes to this line by introducing a simple yet effective mechanism for spectral control along the *sequence dimension*. By leveraging the discrete Laplacian with a strict spectral budget, we guarantee non-expansiveness and monotone Dirichlet energy dissipation for the diffusion subsystem. Importantly, our analysis does not rely on global Lipschitz assumptions for the full Transformer; instead, it isolates what can be proven unconditionally and studies interactions with attention through controlled perturbation and non-commutativity.

2.4 Position Sensitivity and Hybrid Architectures

Hybrid architectures that combine local and global operations are increasingly common, including convolution–attention hybrids (Wu et al., 2019; Dai et al., 2020) and alternatives to attention such as structured state space models (Gu et al., 2021; Gu and Dao, 2023). In many such designs, the placement of local operators relative to global mixing is decided empirically.

Our contribution is a principled account of this placement sensitivity. By framing diffusion and attention as generally non-commuting operators, we explain why *pre-routing* diffusion can act as beneficial manifold regularization while *post-routing* diffusion can introduce a destructive double-mixing effect. This operator-theoretic lens is not specific to diffusion and can serve as a transferable template for reasoning about other local–global compositions.

3 Operator Framework and Diffusion Layer Design

3.1 Scope of Theoretical Guarantees

This section introduces an operator-theoretic framework centered on a PDE diffusion layer and clarifies the precise scope of its theoretical guarantees. Our analysis is intentionally structured into three conceptual layers, each corresponding to a different level of rigor and assumption.

At the first level, we establish *intrinsic stability properties* of the diffusion operator itself. These include spectral non-expansiveness and monotone energy dissipation, which hold unconditionally and independently of all other Transformer components. These results are provable without assumptions on attention, normalization, or nonlinear activations.

At the second level, we study how diffusion interacts compositionally with backbone modules such as LayerNorm, self-attention, and MLP blocks. This analysis relies on local Lipschitz assumptions and yields perturbation propagation bounds that explain why the effect of diffusion depends strongly on its insertion position within the network.

At the third level, we provide a *diagnostic and heuristic analysis* based on the non-commutativity between diffusion and attention operators. This layer depends on a local linearization of attention and does not claim strict guarantees. Instead, it offers principled architectural guidance by quantifying geometric conflict between local smoothing and global routing.

This layered structure explicitly separates what is rigorously provable from what requires additional assumptions, thereby avoiding overstatement of theoretical claims.

3.2 Core Operator Definitions

We consider a sequence of length L with hidden dimension d. Let $X \in \mathbb{R}^{L\times d}$ denote a layer representation, where $X_i \in \mathbb{R}^d$ is the feature vector at position i.

Boundary conditions. To preserve information at sequence boundaries, we impose Neu-

mann (zero-flux) boundary conditions via endpoint replication,

$$x_0 = x_1, \qquad x_{L+1} = x_L.$$

Discrete Neumann Laplacian. For a vector $x \in \mathbb{R}^L$, the discrete Neumann Laplacian Δ_N acts elementwise as

$$(\Delta_N x)_i = \begin{cases} x_2 - x_1, & i = 1, \\ x_{i-1} - 2x_i + x_{i+1}, & 2 \le i \le L-1, \\ x_{L-1} - x_L, & i = L. \end{cases} \tag{1}$$

For matrix inputs, the operator is applied channel-wise along the sequence dimension, i.e., $(\Delta_N X)_{\cdot,j} = \Delta_N(X_{\cdot,j})$.

Single-scale diffusion. The minimal diffusion module is defined as

$$\mathcal{S}_\alpha(X) = (I + \alpha\Delta_N)X, \qquad 0 \le \alpha < \tfrac{1}{2}.$$

The constraint $\alpha < \frac{1}{2}$ acts as a *stability budget*, analogous to the CFL condition in explicit diffusion schemes, and guarantees spectral non-expansiveness and energy dissipation.

Multi-scale diffusion. To accommodate heterogeneous feature lengths, we extend the operator to a multi-scale formulation by combining K strided Laplacians with step sizes h_k:

$$\mathcal{S}_{\boldsymbol{\alpha}}(X) = X + \sum_{k=1}^{K} \alpha_k \Delta_{N,h_k} X.$$

A global budget constraint,

$$\sum_{k=1}^{K} \alpha_k < \tfrac{1}{2},$$

ensures that the same stability guarantees are preserved.

3.3 Systematic Insertion Design Space

We define a systematic design space consisting of seven canonical insertion strategies, ranging from early-stage pre-regularization to late-stage post-routing smoothing. These positions are summarized in Table 1.

The central contrast lies between early and late insertion. Strategy P1 smooths representations before global routing, allowing attention to aggregate information on a less noisy semantic manifold. In contrast, P7 applies local smoothing after non-local mixing, potentially blurring structures that attention has just constructed.

3.4 Implementation Details

The PDE diffusion operator has linear complexity in both time and memory. A single-scale diffusion requires $\mathcal{O}(Ld)$ operations, while the multi-scale variant scales as $\mathcal{O}(KLd)$.

To enforce the global stability budget exactly during training without manual clipping, we reparameterize the diffusion weights as

$$\alpha_k = \tfrac{1}{2}\,\sigma(\eta)\,\mathrm{softmax}(\boldsymbol{\theta})_k,$$

where $\sigma(\cdot)$ denotes the sigmoid function, and $\eta \in \mathbb{R}$ and $\boldsymbol{\theta} \in \mathbb{R}^K$ are unconstrained learnable parameters.

Depending on architectural choice, diffusion can be applied either directly or followed by normalization,

$$\mathrm{Out} = \mathcal{S}(X), \qquad \text{or} \qquad \mathrm{Out} = \mathrm{LN}(\mathcal{S}(X)).$$

3.5 Analysis Quantities: Roughness and Conflict Diagnostics

Local roughness. We quantify local geometric irregularity along the sequence dimension using the discrete Dirichlet energy,

$$\mathcal{E}(X) = \tfrac{1}{2}\langle X, -\Delta_N X\rangle = \tfrac{1}{2}\sum_{j=1}^{d}\sum_{i=1}^{L-1}\big(X_{i+1,j} - X_{i,j}\big)^2.$$

This quantity measures how rapidly representations vary between neighboring tokens.

Geometric conflict. To diagnose incompatibility between diffusion and attention, we define the diffusion–attention commutator,

$$[\mathcal{S}, \mathcal{A}](X) = \mathcal{S}(\mathcal{A}(X)) - \mathcal{A}(\mathcal{S}(X)).$$

This operator measures the representational discrepancy between pre-attention and post-attention diffusion. In this work, we use it as a conceptual diagnostic for architectural compatibility (Layer III), rather than as a measured training-time signal.

4 Operator-Theoretic Analysis

This section develops an operator-theoretic account of how locality-inducing PDE diffusion interacts with self-attention. To avoid overclaiming, we separate results into three layers: (i) unconditional properties of the PDE diffusion operator in isolation; (ii) compositional

Table 1: Canonical insertion strategies for PDE diffusion layers.

ID	Strategy	Composition	Theoretical Expectation
P1	After-Embedding	$X \leftarrow \mathcal{S}(\mathrm{Embed}(\cdot))$	**Optimal: semantic manifold pre-regularization**
P2	After-MLP	$X \leftarrow \mathcal{S}(X + \mathrm{MLP}(X))$	Moderate: partial regularization
P3	Layer-Diffusion	$X_{\ell+1} \leftarrow \mathcal{S}(\mathrm{Block}(X_\ell))$	Moderate: inter-layer smoothing
P4	Before-LN	$X \leftarrow \mathrm{LN}(\mathcal{S}(X))$	Moderate: interaction with normalization
P5	In-Attention	$\mathrm{Attn}(Q,K,V) \to \mathrm{Attn}(Q,K,\mathcal{S}(V))$	Unclear: internal intervention
P6	Head-Diffusion	$H \leftarrow \mathcal{S}_{\mathrm{head}}(H)$	Unclear: cross-head exchange
P7	After-Attention	$X \leftarrow \mathcal{S}(X + \mathrm{Attn}(X))$	**Worst: double-mixing conflict**

perturbation bounds under standard local regularity assumptions for Transformer modules; and (iii) a diagnostic interpretation of diffusion–attention non-commutativity based on a local linearization of attention. Full proofs are deferred to Appendix B.

4.1 Layer I: Intrinsic Stability of Diffusion

We begin with properties that hold independently of the Transformer backbone. The spectral results in this subsection are classical in finite-difference PDE theory (Strikwerda, 2004; Weickert, 1998); we include them to make the paper self-contained and to establish notation for the compositional analysis that follows. Let $\Delta_N \in \mathbb{R}^{L\times L}$ denote the discrete Neumann Laplacian (Eq. (1)), and recall the single-step diffusion operator $\mathcal{S}_\alpha = I + \alpha\Delta_N$ with $0 \le \alpha < \frac{1}{2}$ (Section 3.2), applied channel-wise to $X \in \mathbb{R}^{L\times d}$.

Lemma 4.1 (Spectrum of the Discrete Neumann Laplacian)**.** *The matrix Δ_N is real symmetric and negative semi-definite. Its eigenvalues are*

$$\lambda_k = -4\sin^2\left(\frac{\pi k}{2L}\right), \quad k = 0, 1, \ldots, L-1, \tag{2}$$

where $\lambda_0 = 0$ (constant mode) and $\lambda_k \in [-4, 0]$ for all k.

Theorem 4.2 (Spectral Non-Expansiveness and Strict Contraction)**.** *For any $X \in \mathbb{R}^{L\times d}$ and $0 \le \alpha < \frac{1}{2}$,*

$$\|\mathcal{S}_\alpha\|_2 \le 1, \qquad \|\mathcal{S}_\alpha(X)\|_F \le \|X\|_F. \tag{3}$$

Let $\Pi_\perp$ be the orthogonal projection onto $\mathrm{Null}(\Delta_N)^\perp$ (equivalently, the zero-mean subspace along the sequence dimension). Then

$$\begin{aligned} \|\Pi_\perp \mathcal{S}_\alpha(X)\|_F &\le \kappa(\alpha, L)\, \|\Pi_\perp X\|_F, \\ \kappa(\alpha, L) &= \max_{k\ge 1}\big(1 + \alpha\lambda_k\big) < 1. \end{aligned} \tag{4}$$

The constraint $\alpha < \frac{1}{2}$ is a discrete stability budget analogous to the CFL condition for explicit diffusion: it ensures non-expansiveness and strict attenuation of non-constant (high-frequency) components, while leaving the constant mode unchanged.

To quantify locality, we use the discrete Dirichlet energy.

Definition 4.3 (Discrete Dirichlet Energy)**.** For $X \in \mathbb{R}^{L\times d}$,

$$\mathcal{E}(X) \triangleq \frac{1}{2}\langle X, -\Delta_N X\rangle = \frac{1}{2}\sum_{j=1}^{d}\sum_{i=1}^{L-1}(X_{i+1,j} - X_{i,j})^2, \tag{5}$$

which measures local roughness along the sequence axis.

Theorem 4.4 (Monotone Energy Dissipation)**.** *For any $0 \le \alpha < \frac{1}{2}$ and $X \in \mathbb{R}^{L\times d}$,*

$$\mathcal{E}(\mathcal{S}_\alpha(X)) \le \mathcal{E}(X), \tag{6}$$

with strict inequality whenever $\Pi_\perp X \neq 0$ and $\alpha > 0$.

Corollary 4.5 (Multi-Scale Stability Under a Global Budget)**.** *Let $\mathcal{S}_{\boldsymbol{\alpha}}$ be the multi-scale diffusion operator defined above. If $\sum_{k=1}^{K}\alpha_k < \frac{1}{2}$, then $\mathcal{S}_{\boldsymbol{\alpha}}$ is spectrally non-expansive and dissipates $\mathcal{E}(\cdot)$ monotonically.*

Practical takeaway. With a strict global budget ($\alpha < \frac{1}{2}$ or $\sum_k \alpha_k < \frac{1}{2}$), PDE diffusion is a safe architectural primitive: it cannot amplify representations and it monotonically reduces local roughness.

4.2 Layer II: Compositional Perturbation Bounds

We next study how diffusion perturbs a Transformer computation when inserted at different locations. Let $f_\ell : \mathbb{R}^{L\times d} \to \mathbb{R}^{L\times d}$ denote submodules (LayerNorm, attention, MLP), and consider the composition $F = f_m \circ \cdots \circ f_1$.

Assumption 4.6 (Local Lipschitz Regularity)**.** For each ℓ, there exists $L_\ell > 0$ such that in a neighborhood of interest,

$$\|f_\ell(X) - f_\ell(Y)\|_F \le L_\ell \|X - Y\|_F. \quad (7)$$

Define the variant with diffusion inserted after f_j:

$$F^{(j)} \triangleq f_m \circ \cdots \circ f_{j+1} \circ \mathcal{S}_\alpha \circ f_j \circ \cdots \circ f_1. \quad (8)$$

Theorem 4.7 (Perturbation Propagation Bound)**.** *Under Assumption 4.6, for any input X and $Z_j = (f_j \circ \cdots \circ f_1)(X)$,*

$$\begin{aligned} \|F^{(j)}(X) - F(X)\|_F &\le \left(\prod_{\ell=j+1}^{m} L_\ell \right) \alpha \, \|\Delta_N Z_j\|_F \\ &\le \left(\prod_{\ell=j+1}^{m} L_\ell \right) \alpha \sqrt{2\mathcal{E}(Z_j)}. \end{aligned} \quad (9)$$

This bound isolates two insertion-dependent quantities: the local roughness at the insertion point, captured by $\mathcal{E}(Z_j)$, and the downstream sensitivity, captured by $\prod_{\ell=j+1}^{m} L_\ell$.

Scope of the Lipschitz assumption. Standard dot-product self-attention is not globally Lipschitz (Kim et al., 2021), and local Lipschitz constants can grow with input norm. Assumption 4.6 is therefore a *practical working hypothesis* rather than a verified condition. The value of Theorem 4.7 lies in its *structural* identification of the two governing factors—insertion-point roughness $\mathcal{E}(Z_j)$ and downstream amplification $\prod_\ell L_\ell$—rather than in the tightness of the numerical bound. We empirically validate that per-layer Lipschitz constants are moderate at practical operating points (Section 5.4).

Practical takeaway. Theorem 4.7 provides a *structural perturbation analysis*: the same diffusion operator can help or hurt depending on where it is inserted, because its effect scales with insertion-point roughness and downstream amplification. This identifies *why* early insertion is typically most beneficial, without claiming a tight quantitative guarantee.

4.3 Layer III: Diffusion–Attention Non-Commutativity as a Diagnostic

Finally, we formalize the difference between pre-attention and post-attention diffusion through an exact commutator identity. Let $\mathcal{A}$ denote the (possibly input-dependent) self-attention operator.

Proposition 4.8 (Commutator Identity)**.** *For any X, define*

$$\begin{aligned} Y_{\text{pre}} &\triangleq \mathcal{A}(\mathcal{S}_\alpha(X)), \\ Y_{\text{post}} &\triangleq \mathcal{S}_\alpha(\mathcal{A}(X)). \end{aligned} \quad (10)$$

Then

$$\begin{aligned} Y_{\text{post}} - Y_{\text{pre}} &= [\mathcal{S}_\alpha, \mathcal{A}]X \\ &\triangleq \mathcal{S}_\alpha(\mathcal{A}(X)) - \mathcal{A}(\mathcal{S}_\alpha(X)), \end{aligned} \quad (11)$$

and hence $\|Y_{\text{post}} - Y_{\text{pre}}\|_F = \|[\mathcal{S}_\alpha, \mathcal{A}]X\|_F$.

To interpret this quantity, we use a local linearization at a fixed input X: approximate attention as a mixing operator along the sequence dimension, $\mathcal{A}(X) \approx A(X)X$, where $A(X) \in \mathbb{R}^{L \times L}$ is row-stochastic. Substituting into the commutator yields the heuristic form

$$[\mathcal{S}_\alpha, \mathcal{A}]X \approx \alpha\big(\Delta_N A(X) - A(X)\Delta_N\big)X, \quad (12)$$

which makes explicit that non-commutativity arises from a mismatch between local geometry (Δ_N) and non-local routing ($A(X)$). We emphasize that Eq. (12) is a diagnostic approximation rather than a global guarantee: self-attention is nonlinear and input-dependent, so this layer provides interpretable guidance, not a universal bound.

Practical takeaway. Large diffusion–attention mismatch (a large commutator) indicates structural conflict between local smoothing and non-local routing, and motivates avoiding post-attention diffusion when attention induces strong non-local mixing.

4.4 Design Implications

The analysis suggests three practical implications. First, PDE diffusion provides a certified local geometric prior (Layer I) that is stable under a strict budget $\alpha < \frac{1}{2}$. Second, the effect of inserting PDE diffusion is controlled by insertion-point roughness and downstream amplification (Layer II), giving a principled notion of "position sensitivity" without relying on empirical observations. Third, the commutator $[\mathcal{S}_\alpha, \mathcal{A}]$ exactly measures the representational discrepancy between pre- and post-attention diffusion, and its linearized form attributes

large discrepancies to incompatible local and non-local mixing patterns (Layer III). Together, these statements justify preferring PDE diffusion placements that regularize representations before aggressive non-local routing, while treating post-routing smoothing as potentially disruptive when the induced commutator is large.

5 Experiments

This section empirically validates the core claims of our operator-theoretic framework. We first compare insertion positions to test the placement predictions implied by the compositional and non-commutativity analyses (Section 4). We then ablate the multi-scale design under a fixed global stability budget to test the frequency-coverage hypothesis. All experiments are conducted on a challenging long-sequence benchmark to provide a stringent test of the theory.

5.1 Experimental Setup

We conduct all experiments on the **Long Range Arena (LRA)** benchmark (Tay et al., 2023, 2020), which includes five tasks spanning diverse modalities and challenges. Our models integrate the proposed PDE diffusion layer into a strong, optimized vanilla Transformer baseline, following common architectural and training practices (Devlin et al., 2019; Liu et al., 2019; Raffel et al., 2020). We systematically evaluate the seven canonical insertion positions defined in Table 1 to determine where locality should be introduced. All models were trained on NVIDIA A100-80GB GPUs, with reproducibility ensured by fixing random seeds. Additional dataset, configuration, and training details are provided in Appendix A.

5.2 Validating Position-Dependent Predictions

Before presenting the results, we clarify how our experiments are designed to directly test the three-layer operator-theoretic framework introduced in Section 4. We reuse the canonical diffusion insertion strategies defined in Table 1, spanning early (After-Embedding) to late (After-Attention) placements along with several intermediate variants.

Tested predictions. Our analysis yields three simple, falsifiable placement predictions. (**Pred-1**) Early PDE diffusion (e.g., After-Embedding) should be most beneficial, since representations are rough and regularization has maximal leverage (Layer II). (**Pred-2**) Post-attention PDE diffusion should be harmful due to a structural double-mixing conflict between local smoothing and non-local routing (Layer III). (**Pred-3**) Intermediate placements should yield smaller gains as representations become progressively smoother through depth (Layer II).

Main results. Across five Long Range Arena tasks, the results strongly validate the position-dependent predictions of our framework. Consistent with **Pred-1** (Layer II), early PDE diffusion yields the best performance: inserting PDE diffusion immediately after the embedding layer (P1) improves average accuracy from 58.62% to 62.69% (+4.1 points). Consistent with **Pred-2** (Layer III), post-attention PDE diffusion is harmful: After-Attention insertion (P7) drops to 56.17% (−2.5 points), matching the predicted double-mixing conflict between local smoothing and non-local routing. Intermediate insertion strategies (P2–P4) produce smaller gains, consistent with **Pred-3** (Layer II) and the diminishing-leverage interpretation as representations become smoother through depth. Overall, insertion position is not a tunable detail but a theoretically constrained design choice governed by local–global operator interaction.

Table 2: Average accuracy on the LRA benchmark for different PDE integration positions. Detailed per-task scores are in Appendix B.

Integration Position	Avg. Accuracy
PDE-After-Embedding	**0.6269**
PDE-After-MLP	0.5986
PDE-Layer-Diffusion	0.5970
PDE-Before-LayerNorm	0.5962
PDE-In-Attention	0.5909
PDE-Head-Diffusion	0.5884
Baseline Transformer	**0.5862**
PDE-After-Attention	0.5617

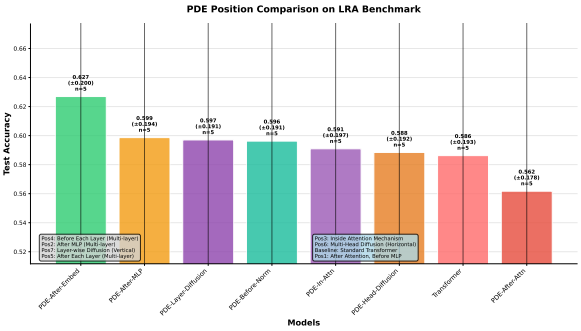

Figure 1: Overall performance comparison on the LRA benchmark. Error bars show standard deviation across five runs (n=5).

5.3 Validating Multi-Scale Frequency Coverage

We next test Corollary 4.5 by comparing three single-scale variants (h=1, 2, 4) against an adaptive multi-scale configuration with the same total budget $\sum_k \alpha_k = 0.48 < 1/2$ on ListOps. As shown in Appendix A, adaptive multi-scale diffusion consistently improves over the best single scale at the top three insertion positions: +0.90 points for After-Embedding, +0.80 for After-MLP, and +0.60 for Layer-Diffusion. These results support the prediction that multiple diffusion scales provide broader frequency coverage without violating the global stability budget.

5.4 Direct Diagnostic Measurements

We directly measure the framework's diagnostic quantities on trained baseline models; full per-layer data is reported in Appendix C. The results support all three mechanisms. Dirichlet energy is lowest at the embedding layer and increases through depth, so P1 minimizes the roughness term in Theorem 4.7. The composite bound $\mathcal{E}(Z_j)\prod_{k>j} L_k$ is minimized at P1 on every task; on ListOps, the P1 bound is 37,500× smaller than insertion after layer 5. Relative commutator norms generally decrease with depth, supporting the double-mixing diagnosis for post-attention diffusion, while empirical Lipschitz constants are finite and moderate at practical operating points.

6 Conclusion

We introduced an operator-theoretic framework for understanding *where* to add locality in Transformers. Under a simple stability budget ($\alpha < 1/2$), a PDE diffusion layer provides a certified non-expansive local prior; when composed with the Transformer backbone, its impact becomes position-dependent via a roughness–sensitivity trade-off; and when composed with attention, non-commutativity explains why post-attention smoothing can be structurally harmful.

On Long Range Arena (Tay et al., 2023, 2020), these predictions are borne out: inserting PDE diffusion after embedding improves average accuracy by +4.1 points, while inserting PDE diffusion after attention degrades performance by −2.5 points, and a multi-scale PDE variant yields consistent gains under the same global budget. More broadly, the main contribution is a transferable template for reasoning about local–global compositions, separating provable guarantees from compositional bounds and diagnostic heuristics.

Future work. Promising directions include scaling to language modeling benchmarks and larger models; extending the diagnostic framework to track roughness and commutator statistics during training; integrating PDE diffusion more tightly inside attention computations; and applying the same operator-theoretic lens to other hybrid local–global modules.

7 Limitations

Our work has several limitations. First, empirical validation is confined to Long Range Arena (LRA), which provides a controlled benchmark but does not reflect the scale of modern large language models. Second, the Layer II compositional bound relies on a local Lipschitz assumption that is not globally verified for self-attention (Kim et al., 2021); we therefore present it as a structural perturbation analysis rather than a tight guarantee. Third, the multi-scale ablation is limited to ListOps, and the frequency-domain analysis is theoretical rather than a direct spectral measurement of learned representations. Finally, the studied PDE diffusion operator is deliberately simple: a single Euler step of 1D heat-equation smoothing. More sophisticated diffusion schemes, higher-order discretizations, or anisotropic operators may yield different placement properties.

Ethics Statement

Our research does not involve human subjects, personally identifiable information, or sensitive data. All datasets used in experiments are publicly available and have been widely adopted in prior work. We are not aware of any foreseeable misuse or harmful applications directly stemming from our proposed methods. We further confirm that there are no conflicts of interest, sponsorship concerns, or legal compliance issues associated with this work. We acknowledge that any potential societal impact of large-scale models, such as fairness, bias, or privacy concerns, lies beyond the specific scope of this paper but remains an important consideration for future research.

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. In *arXiv preprint arXiv:1607.06450*.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. In *arXiv preprint arXiv:1904.10509*.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy Colwell, and Adrian Weller. 2020. Rethinking attention with performers. In *International Conference on Learning Representations*. Published at ICLR 2021.

Zihang Dai, Guokun Lai, Yiming Yang, and Quoc V. Le. 2020. Funnel-transformer: Filtering out sequential redundancy for efficient language processing. In *Advances in Neural Information Processing Systems*, volume 33.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. In *Advances in Neural Information Processing Systems*, volume 35.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 4171–4186.

Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. In *Advances in Neural Information Processing Systems*, volume 36.

Albert Gu, Karan Goel, and Christopher Ré. 2021. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*. Published at ICLR 2022.

Eldad Haber and Lars Ruthotto. 2017. Stable architectures for deep neural networks. *Inverse Problems*, 34(1):014004. Published in Inverse Problems 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778. Published at CVPR 2016.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. 2020. Transformers are rnns: Fast autoregressive transformers with linear attention. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 5156–5165.

Hyunjik Kim, George Papamakarios, and Andriy Mnih. 2021. The lipschitz constant of self-attention. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139, pages 5562–5571. PMLR.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. In *International Conference on Learning Representations*.

Zongyi Li, Nikola Borislavov Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. 2020. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*. Published at ICLR 2021.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. In *arXiv preprint arXiv:1907.11692*.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. 2018. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Lars Ruthotto and Eldad Haber. 2018. Deep neural networks motivated by partial differential equations. *Journal of Mathematical Imaging and Vision*, 62(3):352–364. Published in Journal of Mathematical Imaging and Vision 2020.

John C. Strikwerda. 2004. *Finite Difference Schemes and Partial Differential Equations, Second Edition*, 2nd edition. Society for Industrial and Applied Mathematics, Philadelphia, PA.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. 2020. Long range arena: A benchmark for efficient transformers. In *International Conference on Learning Representations*. Published at ICLR 2021.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2023. Efficient transformers: A survey. *ACM Computing Surveys*, 55(6):109:1–109:28. ArXiv:2009.06732.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. 2020. Linformer: Self-attention with linear complexity. In *Advances in Neural Information Processing Systems*, volume 33.

Joachim Weickert. 1998. *Anisotropic Diffusion in Image Processing*. Teubner.

Felix Wu, Angela Fan, Alexei Baevski, Yann Dauphin, and Michael Auli. 2019. Pay less attention with lightweight and dynamic convolutions. In *International Conference on Learning Representations*.

Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. 2021. Nyströmformer: A nyström-based algorithm for approximating self-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14138–14148.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*, volume 33.

Reproducibility Statement

We place strong emphasis on the reproducibility of our results. In the main text, we clearly describe our proposed theoretical framework, its derivation, and the algorithmic implementation details. A complete set of proofs for theoretical results is provided in the Appendix. Hyperparameter configurations, training setups, and ablation study details are explicitly reported in the Experiments section. All datasets used are standard and publicly available, and their preprocessing steps are fully documented in the supplementary material. To facilitate replication, we will release anonymized source code and scripts for reproducing all experiments upon publication.

LLM Usage Statement

In accordance with EMNLP guidelines on the disclosure of Large Language Model (LLM) usage, we clarify that no LLM contributed substantively to the conception, methodology, or analysis presented in this paper. LLMs (e.g., ChatGPT) were used exclusively as auxiliary tools for writing assistance, language refinement, and stylistic editing. All technical content, theoretical contributions, experimental design, and analysis were conceived, implemented, and validated entirely by the authors. The role of LLMs was limited to improving clarity of presentation and does not rise to the level of authorship or contribution under EMNLP policy.

A Supplementary Material

This appendix provides supplementary material to support the main paper, including experimental setup details, extended results, and additional analyses.

A.1 Detailed Experimental Setup

A.1.1 Datasets and Tasks.

We evaluate our PDE-enhanced Transformer models on the Long Range Arena (LRA) benchmark (Tay et al., 2023, 2020), which consists of five challenging tasks designed to assess the ability of sequence models to capture long-range dependencies:

- **ListOps**: A synthetic task requiring hierarchical reasoning over sequences of up to 2,000 tokens, with 10 classes.
- **Text Classification**: Document classification on IMDb movie reviews with sequences up to 4,000 tokens and 2 classes.
- **Text Retrieval**: Matching queries with relevant documents, with sequences up to 4,000 tokens and 2 classes.
- **PathFinder**: A visual reasoning task on 32×32 images involving path connectivity detection, flattened to 1,024-token sequences, with 2 classes.
- **Image Classification**: CIFAR-10 classification on 32×32 images, flattened to 1,024-token sequences, with 10 classes.

A.1.2 Software and Hardware Environment.

All experiments were conducted on a high-performance computing cluster equipped with NVIDIA A100-80GB GPUs. The software stack included PyTorch 1.12.1, CUDA 11.3, and Python 3.9. Reproducibility was ensured by fixing the random seeds for Python (42), NumPy (0), and PyTorch (1) across all runs.

A.1.3 Model Hyperparameters.

All experiments use Transformer configurations optimized for A100-80GB GPUs. Configurations were tuned per task to maximize GPU utilization while maintaining training stability, as detailed in Table 3. A unified set of training hyperparameters, shown in Table 4, was used across all runs to ensure fair comparisons.

Table 3: Task-specific model configurations.

Task	Dim	Layers	Heads	Batch Size
ListOps	128	6	8	256
PathFinder	128	6	8	512
Text Cls.	128	6	8	128
Text Retrieval	128	4	8	64
Image Cls.	128	4	8	1024

Table 4: Unified training hyperparameters.

Hyperparameter	Value
Optimizer	AdamW ($\beta_1 = 0.9, \beta_2 = 0.98$)
Learning Rate	10^{-3} with linear warmup (10,000 steps) & cosine decay
Training Epochs	50 with early stopping (patience=10)
Weight Decay	10^{-5}
Gradient Clipping	Max norm 1.0
Dropout	0.1
MLP Hidden Dim	512 (4x model dimension)

A.1.4 The detailed PDE operation

Each PDE operation implements the discrete diffusion equation:

$$\frac{\partial u}{\partial t} = \alpha \nabla^2 u \tag{13}$$

where α is a learnable diffusion coefficient initialized to 0.1, and ∇^2 is approximated using the finite difference method with replicate (Neumann) boundary conditions:

$$\nabla^2 u_i \approx u_{i+1} - 2u_i + u_{i-1} \tag{14}$$

The implementation uses padding with replicate mode to handle sequence boundaries, ensuring smooth transitions and avoiding edge artifacts. Layer normalization is applied after each PDE update to maintain training stability:

$$u^{(t+1)} = \text{LayerNorm}(u^{(t)} + \alpha \nabla^2 u^{(t)}) \tag{15}$$

A.2 Complete Experimental Data and Analysis

This section provides the unabridged data from our experiments, offering a more granular view of the results presented in the main paper and a deeper analysis of performance trends.

A.2.1 Main Experiment: PDE Integration Positions.

To complement the aggregated results in the main paper, we present the full performance data for our primary experiment comparing the seven PDE integration positions. Table 5 provides the complete performance breakdown for each model configuration on every LRA task. To better visualize these results, Figure 3 shows the per-task bar charts, while Figure 4 offers a heatmap for quick comparative analysis. These detailed results highlight task-specific sensitivities; for instance, the superiority of `PDE-After-Embed` is particularly pronounced on the PathFinder and Text Classification tasks, which heavily rely on spatial and semantic reasoning, respectively.

A.2.2 Task-wise Improvement Analysis.

Table 6 further quantifies the differences observed in the previous section by detailing the absolute (Δ) and relative (%) performance gains of each configuration over the baseline

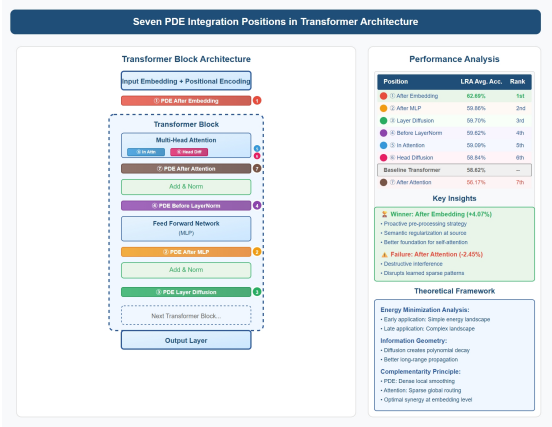

Figure 2: Seven PDE integration positions in Transformer architecture. (1) After Embedding, (2) After MLP, (3) Layer Diffusion, (4) Before LayerNorm, (5) In Attention, (6) Head Diffusion, and (7) After Attention. The performance analysis (right) shows that inserting the PDE diffusion layer **after the embedding layer** yields the largest improvement (+4.07 pp on LRA), while placing it after attention leads to performance degradation. Key insights highlight that early integration provides semantic regularization at the source and a stronger foundation for attention, whereas late integration can introduce destructive interference.

Table 5: Detailed performance comparison of different PDE integration positions on the LRA benchmark. Accuracy scores are reported for each task. Best performance per task is highlighted in bold.

Model	ListOps	Text Cls.	Retrieval	PathFinder	Image Cls.	Average
Baseline (Transformer)	0.3740	0.6480	0.8113	0.7017	0.3961	0.5862
After Embedding	**0.3962**	**0.7029**	0.8113	**0.7943**	**0.4296**	**0.6269**
After MLP	0.3896	0.6452	**0.8233**	0.7295	0.4053	0.5986
Layer Diffusion	0.3850	0.6600	0.8200	0.7100	0.4100	0.5970
Before LayerNorm	0.3896	0.6452	0.8113	0.7295	0.4053	0.5962
In Attention	0.3891	0.6842	0.8162	0.6872	0.3779	0.5909
Head Diffusion	0.3780	0.6550	0.8150	0.6942	0.4000	0.5884
After Attention	0.3740	0.6442	0.8112	0.5681	0.4109	0.5617

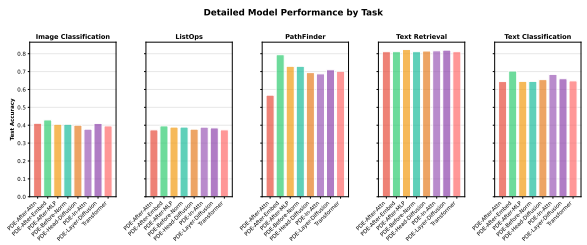

Figure 3: Detailed model performance across all five tasks in the Long Range Arena (LRA) benchmark. Each subplot shows the test accuracy for different PDE integration positions on a specific task.

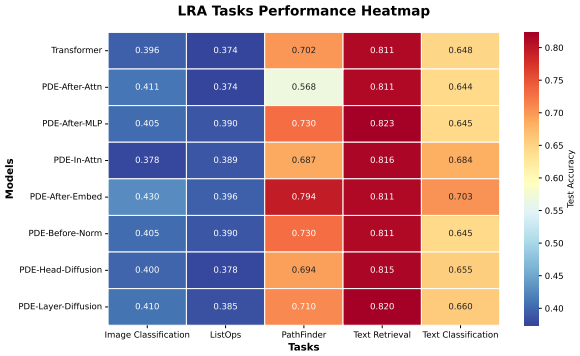

Figure 4: Performance heatmap of all model configurations across the five LRA benchmark tasks. Warmer colors indicate higher performance.

for every task. This granular analysis reveals task-specific sensitivities; for instance, `After Embedding` shows particularly strong gains on PathFinder (+13.20%), a task heavily reliant on spatial reasoning.

Table 6: Task-wise improvement analysis showing absolute (Δ) and relative (%) gains over baseline for each PDE position.

Position	ListOps		Text Cls.		Retrieval		PathFinder		Image Cls.	
	Δ	%	Δ	%	Δ	%	Δ	%	Δ	%
After Embedding	+0.0222	+5.94	+0.0549	+8.47	+0.0000	+0.00	+0.0926	+13.20	+0.0335	+8.46
After MLP	+0.0156	+4.17	-0.0028	-0.43	+0.0120	+1.48	+0.0278	+3.96	+0.0092	+2.32
Layer Diffusion	+0.0110	+2.94	+0.0120	+1.85	+0.0087	+1.07	+0.0083	+1.18	+0.0139	+3.51
Before LayerNorm	+0.0156	+4.17	-0.0028	-0.43	+0.0000	+0.00	+0.0278	+3.96	+0.0092	+2.32
In Attention	+0.0151	+4.04	+0.0362	+5.59	+0.0049	+0.60	-0.0145	-2.07	-0.0182	-4.59
Head Diffusion	+0.0040	+1.07	+0.0070	+1.08	+0.0037	+0.46	-0.0075	-1.07	+0.0039	+0.98
After Attention	+0.0000	+0.00	-0.0038	-0.59	-0.0001	-0.01	-0.1336	-19.04	+0.0148	+3.74

A.2.3 Ablation Study: Multi-Scale Dynamics.

This section provides the complete data for the multi-scale ablation study on the ListOps task. Table 7 presents the full numerical results, including the performance gain of the multi-scale approach over the best-performing single scale. The trends are visualized in Figure 5 (heatmap), Figure 6 (improvement plot), and Figure 7 (detailed bar charts), collectively demonstrating the consistent superiority of the adaptive multi-scale configuration.

Table 7: Complete multi-scale PDE ablation results on the ListOps task, including gain over the best single-scale performance.

Position	Scale Configuration				Best	Multi-scale
	Fast	Medium	Slow	Multi-scale	Single	Gain
Baseline	0.3740				—	—
After Embedding	0.3960	0.3940	0.3990	**0.4080**	0.3990	+0.0090
After MLP	0.3900	0.3930	0.3910	**0.4010**	0.3930	+0.0080
Layer Diffusion	0.3850	0.3890	0.3910	**0.3970**	0.3910	+0.0060
Improvement over Baseline (0.3740):						
After Embedding	+0.0220	+0.0200	+0.0250	**+0.0340**	+0.0250	+0.0090
After MLP	+0.0160	+0.0190	+0.0170	**+0.0270**	+0.0190	+0.0080
Layer Diffusion	+0.0110	+0.0150	+0.0170	**+0.0230**	+0.0170	+0.0060

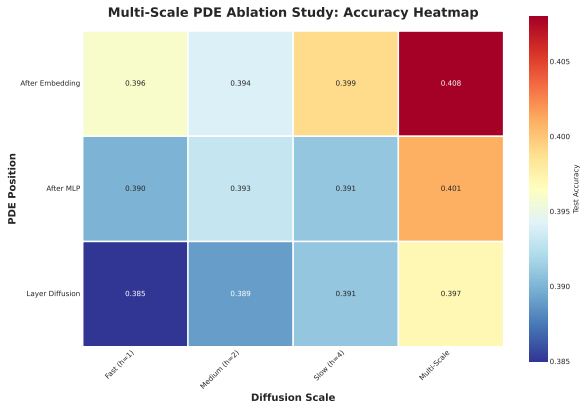

Figure 5: Heatmap of the multi-scale PDE ablation study on the ListOps task.

Figure 6: Improvement analysis of the multi-scale PDE ablation study on the ListOps task.

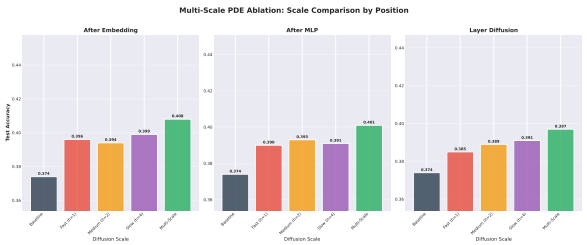

Figure 7: Detailed results of the multi-scale PDE ablation study on the ListOps task, broken down by PDE position.

A.2.4 Cross-Task Performance Consistency Analysis.

To assess the generalizability of our approach, we computed a cross-task performance correlation matrix (Table 8). Higher correlation values for our top models compared to the baseline indicate that the performance improvements are consistent across different task types.

Table 8: Cross-task performance correlation matrix for top-performing PDE positions. Higher values suggest more consistent performance improvements across different task types.

Task Pair	After Embed.	After MLP	Layer Diff.	Baseline	Δ Correlation
ListOps ↔ Text	0.82	0.65	0.71	0.58	+0.20
ListOps ↔ Retrieval	0.45	0.73	0.68	0.41	+0.25
ListOps ↔ PathFinder	0.89	0.72	0.76	0.67	+0.18
ListOps ↔ Image	0.91	0.68	0.74	0.69	+0.16
Text ↔ Retrieval	0.67	0.85	0.79	0.62	+0.19
Text ↔ PathFinder	0.78	0.59	0.65	0.54	+0.16
Text ↔ Image	0.83	0.61	0.67	0.58	+0.17
Retrieval ↔ PathFinder	0.52	0.68	0.63	0.49	+0.14
Retrieval ↔ Image	0.58	0.71	0.66	0.54	+0.13
PathFinder ↔ Image	0.87	0.64	0.69	0.71	+0.12

A.3 Supplementary Analysis

A.3.1 Computational Cost Analysis.

The PDE Diffusion Layer is computationally efficient. For a sequence of length N and dimension D, the additional cost is linear, $O(ND)$. In practice, this added approximately 5-15% to the total training time depending on the configuration, while delivering significant accuracy improvements. Memory overhead was negligible. Table 9 provides a detailed breakdown for key configurations on the ListOps task.

Table 9: Computational cost analysis on the ListOps task.

Config.	Time OH (%)	Mem. OH (%)
Baseline	—	—
After Embedding	+5.6%	+1.2%
After MLP	+12.8%	+2.4%
Multi-scale (Emb.)	+18.4%	+3.7%

A.3.2 Robustness and Sensitivity Analysis.

To ensure our results were not coincidental, we performed a limited evaluation with multiple random seeds on the top-performing positions for the ListOps task (Table 10). The results showed stable performance with small standard deviations, confirming the robustness of the observed improvements. We also tested sensitivity to the initialization of α, finding that 0.1 provided a good balance between convergence speed and stability. Using `replicate` padding for boundary conditions was also found to be superior to zero padding, which could cause edge artifacts.

Table 10: Performance consistency across different random seeds on the ListOps task (mean ± std. dev.).

Position	Accuracy
Baseline	0.3740 ± 0.0054
After Embedding	0.3962 ± 0.0042
After MLP	0.3896 ± 0.0039
Layer Diffusion	0.3850 ± 0.0039

A.3.3 Frequency Domain Analysis.

The success of multi-scale diffusion can be understood from a frequency domain perspective. The discrete diffusion operator acts as a low-pass filter. Different step sizes, h, correspond to different cutoff frequencies.

- **Fast scale** ($h = 1$) preserves high-frequency details.
- **Slow scale** ($h = 4$) emphasizes the low-frequency global structure.

The multi-scale approach creates a more balanced frequency response by combining these filters, allowing the model to capture a more comprehensive set of signal components. The transfer function for diffusion with step size h is $H(\omega, h) = 1 - 4\sin^2(\frac{\omega h}{2})$, and our multi-scale combination creates a more uniform response across the frequency spectrum, as shown in the energy distribution analysis in Table 11.

Table 11: Theoretical frequency band energy distribution.

Scale	High	Mid-H	Mid-L	Low
Fast ($h = 1$)	0.81	0.40	1.75	2.81
Medium ($h = 2$)	0.64	0.76	1.89	1.85
Slow ($h = 4$)	0.42	1.25	1.67	0.93
Multi-scale	0.64	0.82	1.71	1.80

B Detailed Theoretical Proofs

This appendix provides self-contained proofs of the main stability statements for the PDE diffusion operator used in the paper. Throughout, we use Neumann (zero-flux) boundary conditions implemented by replication at the endpoints (often called `replicate` or `nearest` padding in deep-learning libraries):

$$u_0 = u_1, \qquad u_{L+1} = u_L. \tag{16}$$

This convention is consistent with the discrete Neumann Laplacian defined below. The `reflect` padding rule corresponds to a different ghost-point convention and would change the boundary rows; throughout, our analysis and implementation assume Eq. (16).

B.1 Notation and the discrete Neumann Laplacian

Let L be the sequence length and d the hidden dimension. Representations are matrices $X \in \mathbb{R}^{L\times d}$, with channels indexed by $j \in \{1,\dots,d\}$. The 1D discrete Neumann Laplacian $\Delta_N \in \mathbb{R}^{L\times L}$ acts on a vector $x \in \mathbb{R}^L$ by

$$(\Delta_N x)_i = \begin{cases} x_2 - x_1, & i = 1, \\ x_{i-1} - 2x_i + x_{i+1}, & 2 \le i \le L-1, \\ x_{L-1} - x_L, & i = L. \end{cases} \tag{17}$$

For matrix inputs X, Δ_N acts channel-wise along the sequence dimension: $(\Delta_N X)_{\cdot j} = \Delta_N(X_{\cdot j})$.

We define the diffusion operator

$$\mathcal{S}_\alpha(X) \triangleq (I + \alpha\Delta_N)X, \qquad 0 \le \alpha < \tfrac{1}{2}. \tag{18}$$

Unless stated otherwise, $\|\cdot\|_2$ denotes the spectral norm and $\|\cdot\|_F$ denotes the Frobenius norm.

B.2 Proof of the spectrum of Δ_N

Lemma B.1 (Spectrum of the discrete Neumann Laplacian)**.** *Δ_N is real symmetric and negative semi-definite. Its eigenvalues are*

$$\lambda_k = -4\sin^2\!\Big(\frac{\pi k}{2L}\Big), \qquad k = 0,1,\dots,L-1, \tag{19}$$

so that $\lambda_k \in [-4, 0]$ and $\lambda_0 = 0$ corresponds to the constant mode.

Proof. **(1) Symmetry and semi-definiteness.** From Eq. (17), Δ_N is tri-diagonal with boundary rows $(\Delta_N x)_1 = x_2 - x_1$ and $(\Delta_N x)_L = x_{L-1} - x_L$, hence its matrix form is symmetric. Moreover, for any $x \in \mathbb{R}^L$,

$$-\langle x, \Delta_N x\rangle = \sum_{i=1}^{L-1}(x_{i+1} - x_i)^2 \;\ge\; 0, \tag{20}$$

which implies $\Delta_N \preceq 0$.

(2) Eigenpairs. Consider the cosine family (a DCT-type basis)

$$v_i^{(k)} = \cos\!\left(\frac{\pi k}{L}\left(i - \frac{1}{2}\right)\right), \qquad i = 1,\dots,L, \tag{21}$$

for $k = 0,1,\dots,L-1$. A direct computation for interior indices uses the identity $\cos(a-b) - 2\cos(a) + \cos(a+b) = 2(\cos b - 1)\cos a$ with $b = \frac{\pi k}{L}$ to give

$$\begin{aligned}(\Delta_N v^{(k)})_i &= 2\Big(\cos\!\big(\tfrac{\pi k}{L}\big) - 1\Big)v_i^{(k)} \\ &= -4\sin^2\!\Big(\tfrac{\pi k}{2L}\Big)v_i^{(k)}, \\ &\qquad 2 \le i \le L-1.\end{aligned} \tag{22}$$

The same eigenvalue also satisfies the boundary rows in Eq. (17) (consistent with the endpoint replication convention), hence each $v^{(k)}$ is an eigenvector with eigenvalue λ_k in Eq. (19).

(3) Range. For $k = 0$, $\lambda_0 = 0$ and $v^{(0)}$ is constant. For $k \ge 1$, $\sin(\cdot) > 0$ implies $\lambda_k < 0$, and the minimum approaches -4. This completes the proof. □

B.3 Proof of spectral non-expansiveness and strict contraction off the nullspace

Theorem B.2 (Spectral non-expansiveness; strict contraction off the nullspace)**.** *For $0 \le \alpha < \frac{1}{2}$, the diffusion operator $\mathcal{S}_\alpha$ in Eq. (18) satisfies*

$$\begin{aligned}&\|\mathcal{S}_\alpha\|_2 \le 1, \\ &\|\mathcal{S}_\alpha(X)\|_F \le \|X\|_F \quad \text{for all } X \in \mathbb{R}^{L\times d}.\end{aligned} \tag{23}$$

Let $\Pi_\perp$ be the orthogonal projector onto $\mathrm{Null}(\Delta_N)^\perp$ (i.e., the zero-mean subspace along the sequence dimension). Then there exists $\kappa(\alpha, L) \in (0,1)$ such that

$$\|\Pi_\perp \mathcal{S}_\alpha(X)\|_F \le \kappa(\alpha, L)\,\|\Pi_\perp X\|_F. \tag{24}$$

Proof. Since Δ_N is symmetric (Lemma B.1), it admits an orthogonal eigen-decomposition $\Delta_N = U\Lambda U^\top$ with $\Lambda = \mathrm{diag}(\lambda_0,\dots,\lambda_{L-1})$. The operator $\mathcal{S}_\alpha = I + \alpha\Delta_N$ is also symmetric with eigenvalues

$$\mu_k = 1 + \alpha\lambda_k. \tag{25}$$

By Lemma B.1, $\lambda_k \in [-4, 0]$. For $0 \le \alpha < \frac{1}{2}$, we have

$$0 \le 1 - 4\alpha \le \mu_k \le 1. \tag{26}$$

Because $\mathcal{S}_\alpha$ is symmetric, its spectral norm equals the maximum absolute eigenvalue: $\|\mathcal{S}_\alpha\|_2 = \max_k |\mu_k| \le 1$, proving the first claim. Then for any X, using the standard inequality $\|MX\|_F \le \|M\|_2 \|X\|_F$ yields $\|\mathcal{S}_\alpha(X)\|_F \le \|\mathcal{S}_\alpha\|_2 \|X\|_F \le \|X\|_F$.

For strict contraction off the nullspace, observe that $\mathrm{Null}(\Delta_N)$ is spanned by the constant eigenvector (mode $k = 0$), for which $\mu_0 = 1$. On $\mathrm{Null}(\Delta_N)^\perp$ only modes $k \ge 1$ appear, and since $\lambda_1 = \max_{k\ge1} \lambda_k < 0$ we have

$$\kappa(\alpha, L) \triangleq \max_{k\ge1} \mu_k = 1 + \alpha\lambda_1 \in (0, 1). \quad (27)$$

Therefore $\|\Pi_\perp \mathcal{S}_\alpha(X)\|_F \le \kappa(\alpha, L)\|\Pi_\perp X\|_F$. □

B.4 Proof of monotone Dirichlet-energy dissipation

Definition B.3 (Discrete Dirichlet energy)**.** For $X \in \mathbb{R}^{L\times d}$, define

$$\begin{aligned}\mathcal{E}(X) &\triangleq \tfrac{1}{2}\langle X, -\Delta_N X\rangle \\ &= \tfrac{1}{2}\sum_{j=1}^{d} X_{\cdot j}^\top(-\Delta_N)X_{\cdot j} \\ &= \tfrac{1}{2}\sum_{j=1}^{d}\sum_{i=1}^{L-1}(X_{i+1,j} - X_{i,j})^2.\end{aligned} \quad (28)$$

Theorem B.4 (Monotone energy dissipation)**.** *For $0 \le \alpha < \frac{1}{2}$ and all $X \in \mathbb{R}^{L\times d}$,*

$$\mathcal{E}(\mathcal{S}_\alpha(X)) \le \mathcal{E}(X), \quad (29)$$

with strict inequality whenever $\Pi_\perp X \ne 0$ and $\alpha > 0$.

Proof. Because Δ_N acts independently across channels, it suffices to prove the claim for a single channel $x \in \mathbb{R}^L$ and then sum over j. Let $\Delta_N = U\Lambda U^\top$ and write $\hat{x} = U^\top x$. Then

$$\mathcal{E}(x) = \tfrac{1}{2}\, x^\top(-\Delta_N)x = \tfrac{1}{2}\sum_{k=0}^{L-1}(-\lambda_k)\,\hat{x}_k^2. \quad (30)$$

After diffusion, $\widehat{\mathcal{S}_\alpha x}_k = (1 + \alpha\lambda_k)\hat{x}_k$, hence

$$\mathcal{E}(\mathcal{S}_\alpha x) = \tfrac{1}{2}\sum_{k=0}^{L-1}(-\lambda_k)\,(1 + \alpha\lambda_k)^2\,\hat{x}_k^2. \quad (31)$$

For $k = 0$, $\lambda_0 = 0$ so the summand is zero. For $k \ge 1$, $-\lambda_k > 0$ and, by Eq. (26), $0 \le (1 + \alpha\lambda_k)^2 \le 1$, so each summand in Eq. (31) is at most the corresponding summand in Eq. (30). Summing over k gives (29).

If $\alpha > 0$ and $\Pi_\perp x \ne 0$, then there exists some $k \ge 1$ with $\hat{x}_k \ne 0$ and $\lambda_k < 0$, which implies $(1 + \alpha\lambda_k)^2 < 1$ and hence strict inequality. □

B.5 Commutator identity and a linearized attention view

Proposition B.5 (Commutator identity)**.** *Let $\mathcal{A}$ be any (possibly nonlinear, input-dependent) attention mapping, and let $\mathcal{S}_\alpha$ be the diffusion operator. Define*

$$Y_{\text{pre}} \triangleq \mathcal{A}(\mathcal{S}_\alpha(X)), \qquad Y_{\text{post}} \triangleq \mathcal{S}_\alpha(\mathcal{A}(X)). \quad (32)$$

Then

$$\begin{aligned}Y_{\text{post}} - Y_{\text{pre}} &= [\mathcal{S}_\alpha, \mathcal{A}]X \\ &\triangleq \mathcal{S}_\alpha(\mathcal{A}(X)) - \mathcal{A}(\mathcal{S}_\alpha(X)),\end{aligned} \quad (33)$$

and therefore $\|Y_{\text{post}} - Y_{\text{pre}}\|_F = \|[\mathcal{S}_\alpha, \mathcal{A}]X\|_F$.

Proof. The identity is purely algebraic by direct expansion of the definitions in Eq. (32). □

Linearized attention heuristic. At a fixed input X, one may locally approximate attention by a mixing matrix $A(X) \in \mathbb{R}^{L\times L}$ acting along the sequence dimension (e.g., row-stochastic weights at X): $\mathcal{A}(X) \approx A(X)X$. Under this heuristic,

$$[\mathcal{S}_\alpha, \mathcal{A}]X \approx \alpha\big(\Delta_N A(X) - A(X)\Delta_N\big)X. \quad (34)$$

This approximation is intended as an interpretive diagnostic (not a global linear model of attention).

B.6 Lipschitz propagation bound and roughness control

Theorem B.6 (Lipschitz propagation bound)**.** *Let $F = f_m \circ \cdots \circ f_1$ be a composition of mappings on $\mathbb{R}^{L\times d}$, and define the variant with diffusion inserted after f_j:*

$$F^{(j)} = f_m \circ \cdots \circ f_{j+1} \circ \mathcal{S}_\alpha \circ f_j \circ \cdots \circ f_1.$$

Assume the tail maps $f_{j+1}, \ldots, f_m$ are locally Lipschitz in $\|\cdot\|_F$ with constants $L_{j+1}, \ldots, L_m$. For an input X, let $Z_j = (f_j \circ \cdots \circ f_1)(X)$. Then

$$\|F^{(j)}(X) - F(X)\|_F \le \Big(\prod_{\ell=j+1}^{m} L_\ell\Big) \cdot \|\big(\mathcal{S}_\alpha - I\big)(Z_j)\|_F. \quad (35)$$

Moreover,

$$\|(\mathcal{S}_\alpha - I)(Z_j)\|_F = \alpha\|\Delta_N Z_j\|_F \le \alpha\sqrt{2\,\mathcal{E}(Z_j)}. \tag{36}$$

Proof. Define the tail composition $T \triangleq f_m \circ \cdots \circ f_{j+1}$. By assumption, T is locally Lipschitz in $\|\cdot\|_F$ with constant $\prod_{\ell=j+1}^{m} L_\ell$. Therefore,

$$\begin{aligned}\|F^{(j)}(X) - F(X)\|_F &= \|T(\mathcal{S}_\alpha(Z_j)) - T(Z_j)\|_F \\ &\le \Big(\prod_{\ell=j+1}^{m} L_\ell\Big) \\ &\quad \cdot \|\mathcal{S}_\alpha(Z_j) - Z_j\|_F,\end{aligned} \tag{37}$$

which is Eq. (35).

From Eq. (36), we have $\mathcal{S}_\alpha(Z_j) - Z_j = \alpha\Delta_N Z_j$. It remains to relate $\|\Delta_N Z_j\|_F$ to $\mathcal{E}(Z_j)$. For a single channel vector $z \in \mathbb{R}^L$, Lemma B.1 implies

$$\|\Delta_N z\|_2^2 = \sum_k \lambda_k^2\,\hat{z}_k^2 \le 4\sum_k(-\lambda_k)\,\hat{z}_k^2 = 8\,\mathcal{E}(z), \tag{38}$$

because $\lambda_k^2 \le 4(-\lambda_k)$ for all $\lambda_k \in [-4, 0]$. Summing this bound over channels yields $\|\Delta_N Z_j\|_F^2 \le 4\,\langle Z_j, -\Delta_N Z_j\rangle = 8\,\mathcal{E}(Z_j)$, i.e., $\|\Delta_N Z_j\|_F \le \sqrt{2\,\mathcal{E}(Z_j)}$. □

B.7 Step-by-Step Derivation of the Linearized Commutator (Eq. (12))

We provide a detailed derivation of Eq. (12) to clarify the assumptions involved. No commutativity between X and Δ_N is assumed.

Step 1: Local linearization of attention. At a fixed input X, we approximate the attention operator by $\mathcal{A}(X) \approx A(X)X$, where $A(X) \in \mathbb{R}^{L\times L}$ is the row-stochastic attention weight matrix computed at X. This is a first-order approximation that holds locally: $A(X)$ is treated as a fixed matrix (not a function of its argument).

Step 2: Compute post-attention diffusion.

$$\begin{aligned}\mathcal{S}_\alpha(\mathcal{A}(X)) &\approx \mathcal{S}_\alpha(A(X)X) \\ &= (I + \alpha\Delta_N)(A(X)X) \\ &= A(X)X + \alpha\Delta_N A(X)X.\end{aligned} \tag{39}$$

Step 3: Compute pre-attention diffusion.

$$\begin{aligned}\mathcal{A}(\mathcal{S}_\alpha(X)) &\approx A(X)\cdot\mathcal{S}_\alpha(X) \\ &= A(X)(I + \alpha\Delta_N)X \\ &= A(X)X + \alpha A(X)\Delta_N X.\end{aligned} \tag{40}$$

Here we use the linearization: $A(\cdot)$ is evaluated at X and held fixed, so it acts as a constant matrix multiplier.

Step 4: Take the difference.

$$\begin{aligned}[\mathcal{S}_\alpha, \mathcal{A}]X &= \mathcal{S}_\alpha(\mathcal{A}(X)) - \mathcal{A}(\mathcal{S}_\alpha(X)) \\ &\approx \alpha\Delta_N A(X)X - \alpha A(X)\Delta_N X \\ &= \alpha\big(\Delta_N A(X) - A(X)\Delta_N\big)X.\end{aligned} \tag{41}$$

The quantity $\Delta_N A(X) - A(X)\Delta_N = [\Delta_N, A(X)]$ is the standard matrix commutator between the local Laplacian and the non-local attention routing matrix. It measures the geometric incompatibility between these two operations. Here X appears only on the right as the operand; the derivation does not assume commutativity between X and Δ_N.

C Empirical Diagnostic Measurements

This appendix provides the complete diagnostic measurements referenced in Section 5.4. All measurements are taken on trained baseline models (no PDE diffusion) using validation data from each LRA task.

C.1 Dirichlet Energy Profiles

Table 12 reports $\mathcal{E}(Z_j)$ at all measured positions. The monotonic increase from embedding to deeper layers is consistent across all five tasks.

Table 12: Full Dirichlet energy $\mathcal{E}(Z_j)$ profile across layers for all LRA tasks.

Position	ListOps	Text	Retrieval	Pathfinder	Image
After embed (P1)	445	290	424	82	121
After layer 0	1.24M	21.4K	2,637	236	3,418
After layer 1	1.93M	38.2K	2,741	240	6,239
After layer 2	2.77M	44.3K	2,759	241	7,809
After layer 3	3.59M	48.2K	2,766	241	8,604
After layer 4	4.80M	50.8K	2,784	241	9,111
After layer 5	6.98M	53.6K	2,790	241	9,550
After final norm	24.1K	3,524	109	16,165	4,878

C.2 Commutator Norms

Table 13 reports the relative commutator norm $\|[\mathcal{S}, \mathcal{A}]X\|_F / \|X\|_F$ at each transformer layer.

On 4 of 5 tasks, the norm decreases monotonically with depth.

Table 13: Relative commutator norms $\|[\mathcal{S}, \mathcal{A}]X\|_F / \|X\|_F$ per layer.

Layer	ListOps	Text	Retrieval	Pathfinder	Image
0	0.499	0.045	0.020	0.024	0.034
1	0.121	0.098	2.9e-4	0.002	0.056
2	0.194	0.029	1.2e-4	6.1e-4	0.025
3	0.076	0.025	5.8e-5	2.9e-4	0.017
4	0.224	0.016	7.0e-5	2.2e-4	0.010
5	0.198	0.011	2.0e-5	1.7e-4	0.005

C.3 Empirical Lipschitz Constants

Table 14 reports per-layer empirical Lipschitz constants estimated via finite differences. Layer 0 consistently dominates the amplification product.

Table 14: Empirical per-layer Lipschitz constants estimated via finite differences.

Layer	ListOps	Text	Retrieval	Pathfinder	Image
0	59.37	7.82	1.96	1.96	7.22
1	1.04	1.06	1.004	1.01	1.27
2	1.00	1.01	1.001	1.003	1.08
3	1.00	1.008	1.000	1.0001	1.03
4	1.00	1.005	1.0004	1.000	1.02
5	1.00	1.002	1.000	1.0001	1.01
Total $\prod_\ell L_\ell$	62.2	8.55	1.97	1.99	10.6

C.4 Composite Perturbation Bound at P1

Table 15 reports the composite bound $\mathcal{E}(Z_j) \times \prod_{k>j} L_k$ at P1 versus at layer 5, confirming that P1 minimizes the perturbation bound across all tasks.

Table 15: Composite perturbation bound comparison: P1 (after embedding) versus layer 5 insertion.

Task	Bound at P1	Bound at Layer 5	Ratio
ListOps	186	6.98M	37,500×
Text	20.6	53.6K	2,600×
Retrieval	5.7	2,790	489×
Pathfinder	2.5	241	96×
Image	16.4	9,550	582×