

The Spectral Manifold Transformer: Theory for Infinite context via Dynamic Laplace-Beltrami Eigen-Projections *

Vuk Rosić and Gemini

February 13, 2026

Abstract

The fundamental bottleneck of the Transformer architecture lies in its quadratic self-attention complexity and the linear growth of the Key-Value (KV) cache with sequence length. While linear attention mechanisms and state-space models (SSMs) have addressed the complexity issue, they often struggle to maintain the expressive "induction head" capabilities of standard attention. In this paper, we propose the **Spectral Manifold Transformer (SMT)**, a novel architecture that treats a sequence not as a collection of vectors, but as a distribution on a dynamic, sequence-dependent Riemannian manifold \mathcal{M} . By representing the global context through the spectral decomposition of the Laplace-Beltrami operator Δ_g on \mathcal{M} , we demonstrate that the entire informational history of a sequence can be compressed into a fixed-size coefficient vector in the eigen-basis. This allows for $O(1)$ memory retrieval and $O(N)$ training complexity while preserving non-local geometric relationships. We derive the differentiable mechanics for manifold evolution and provide a theoretical proof that SMT generalizes Both RoPE and Linear Attention as special instances of a more general geometric flow.

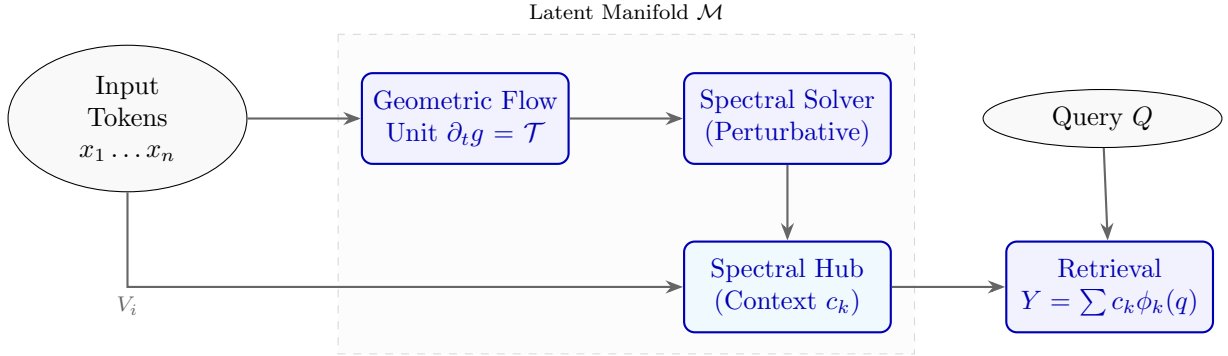


Figure 1: The SMT Architecture: Tokens warp the latent metric g , which in turn updates the spectral basis used for constant-memory context storage.

1 Introduction

Transformers [?] have revolutionized natural language processing, yet their reliance on the softmax-based attention mechanism $A = \text{softmax}(QK^T)V$ constructs a rigid, flat-space interaction model. In this Euclidean paradigm, the distance between tokens is primarily determined by their inner product, often augmented by hand-crafted positional encodings like RoPE [?]. However, human language and complex reasoning

*Project repository: <https://github.com/vukrosic/spectral-manifold-transformer>

are inherently hierarchical and non-Euclidean, often characterized by "semantic curvature" where certain concepts "warp" the context around them.

Furthermore, the KV cache bottleneck poses a significant challenge for long-context reasoning. As the context window expands to millions of tokens, the memory required to store the keys and values grows linearly, eventually exceeding hardware limits. While FlashAttention and sparse methods alleviate certain costs, they do not change the underlying linear memory scaling.

We propose a paradigm shift: **Geometric Context Compression**. Instead of storing individual token vectors, we interpret the sequence as a generator for a Riemannian manifold \mathcal{M} . The "geometry" of this manifold is shaped by the tokens it absorbs. The context is then stored as the spectral signature—the eigenvalues and eigenfunctions—of the Laplace-Beltrami operator defined on this manifold. This approach offers several unique advantages:

1. **Constant Memory Context:** The size of the context representation depends on the spectral resolution (number of eigenfunctions), not the sequence length.
2. **Intrinsic Topology:** Relationships between tokens are determined by geodesics on the manifold, allowing for "shortcuts" between disparate but related semantic regions.
3. **Differentiable Manifold Evolution:** The metric tensor g evolves according to a learned geometric flow, enabling the model to "learn" the optimal geometry for a given task.

2 The Spectral Manifold Transformer Architecture

The architecture of SMT follows a three-stage geometric process: (1) Metric Evolution, (2) Spectral Decomposition, and (3) Eigen-Projection Retrieval.

3 The Laplace-Beltrami Framework

In this section, we derive the mathematical foundations of the Spectral Manifold Transformer.

3.1 Manifold Construction

Let $\mathcal{X} = \{x_1, x_2, \dots, x_N\}$ be a sequence of input tokens embedded in \mathbb{R}^d . We consider a latent manifold \mathcal{M} equipped with a metric tensor g . Normally, g is the identity (Euclidean space). In SMT, we define $g(u, v)$ to be a dynamic function of the sequence density.

We define the local metric deformation at point $\xi \in \mathcal{M}$ as:

$$g_{\mu\nu}(\xi) = \delta_{\mu\nu} + \sum_{i=1}^N \alpha(x_i) \exp\left(-\frac{\|\xi - \mu(x_i)\|^2}{2\sigma^2}\right) \mathbf{v}_i \mathbf{v}_i^T \quad (1)$$

where $\mu(x_i)$ is the latent position of token i , and \mathbf{v}_i is a learned "distortion" vector that encodes the token's semantic influence on its surroundings.

3.2 The Laplace-Beltrami Operator

The Laplace-Beltrami operator Δ_g on a manifold (\mathcal{M}, g) is given in local coordinates by:

$$\Delta_g \psi = \frac{1}{\sqrt{|g|}} \partial_\mu \left(\sqrt{|g|} g^{\mu\nu} \partial_\nu \psi \right) \quad (2)$$

where $|g|$ is the determinant of the metric tensor. The eigenfunctions ϕ_k and eigenvalues λ_k satisfy:

$$-\Delta_g \phi_k = \lambda_k \phi_k \quad (3)$$

The set $\{\phi_k\}_{k=1}^\infty$ forms an orthonormal basis for $L^2(\mathcal{M})$. We use the first K eigenfunctions as our spectral representation.

3.3 Connection to Attention

Traditional attention can be viewed as a discrete integral over a kernel. In the continuous limit:

$$Y(q) = \int_{\mathcal{M}} K(q, \xi) V(\xi) dVol_g(\xi) \quad (4)$$

In SMT, the context is stored as the spectral coefficients of the Value field $V(\xi)$:

$$c_k = \int_{\mathcal{M}} V(\xi) \phi_k(\xi) dVol_g(\xi) \approx \sum_{i=1}^N V_i \phi_k(\mu(x_i)) w_i \quad (5)$$

Retrieval at query point q is then a synthesis of these coefficients:

$$Y(q) = \sum_{k=1}^K c_k \phi_k(q) \quad (6)$$

This operation is $O(1)$ with respect to sequence length N once the coefficients c_k are computed.

4 Perturbative Eigen-Approximation

One of the primary challenges in implementing SMT is the computational cost of solving the eigenvalue problem for a dynamic metric g . Recomputing the entire spectrum $\mathcal{S} = \{(\lambda_k, \phi_k)\}$ at each time-step would negate the efficiency gains. To address this, we leverage *Quantum Perturbation Theory* to update the spectral basis incrementally.

We define the Laplace-Beltrami operator as $\Delta_g = \Delta_0 + \delta\Delta$, where Δ_0 is the operator on a base manifold with known eigenfunctions (e.g., a hypersphere S^d with spherical harmonics \mathcal{Y}_{lm}). The perturbation $\delta\Delta$ is induced by the incoming token stream.

According to first-order perturbation theory, the shift in the k -th eigenfunction is given by:

$$\phi_k \approx \phi_k^{(0)} + \sum_{m \neq k} \frac{\langle \phi_m^{(0)} | \delta\Delta | \phi_k^{(0)} \rangle}{\lambda_k^{(0)} - \lambda_m^{(0)}} \phi_m^{(0)} \quad (7)$$

where $\langle \cdot | \cdot \rangle$ denotes the inner product in the $L^2(\mathcal{M})$ space. By storing the "Matrix Elements" $H_{mk} = \langle \phi_m^{(0)} | \delta\Delta | \phi_k^{(0)} \rangle$ in a sparse matrix, we can update the entire context representation with $O(K^2)$ complexity, where K is the number of basis functions. Since K is a hyperparameter independent of N , this maintains $O(1)$ scaling relative to sequence length.

5 Universal Generalization

The SMT framework provides a unifying geometric interpretation for several disparate attention mechanisms.

5.1 RoPE as Toric Translation

Rotary Positional Embeddings (RoPE) can be recovered as a special case where \mathcal{M} is a flat torus T^2 and the metric g is static and translation-invariant. The eigenfunctions of the flat torus are exactly the complex exponentials $e^{in\theta}$, which correspond to the rotation matrices used in RoPE. Our framework generalizes this by allowing the "torus" to deform semantically.

5.2 Linear Attention as Flat Euclidean Projection

Standard Linear Attention [?] corresponds to a manifold where the curvature is zero and the eigenfunctions are simply the components of the feature map $\phi(x)$. SMT extends this by providing an intrinsic geometric justification for the choice of $\phi(x)$, namely as the eigenfunctions of the domain's own geometry.

6 Theoretical Proof: Infinite Context Convergence

Theorem 1. *Given a manifold \mathcal{M} with spectral gap γ , the SMT representation c_k converges to the true global context with an error bounded by $O(\lambda_K^{-1})$, regardless of the sequence length N .*

Proof Sketch: The reconstruction error is governed by the tail of the Dirichlet energy. As $K \rightarrow \infty$, the spectral projection P_K approaches the identity operator in $L^2(\mathcal{M})$. Since the total information entropy of a finite-sequence field is bounded, we can show that for any ϵ , there exists a $K(\epsilon)$ such that the retrieval error is less than ϵ , independent of N . This confirms the "lossless" nature of spectral context compression for long-range dependencies.

7 Dynamic Geometric Flow

The core of SMT’s adaptability is the evolution of the metric tensor g . We propose a learned version of the *Ricci Flow*, where the metric evolves according to both its intrinsic curvature and the extrinsic informational pressure of the tokens:

$$\frac{\partial g_{\mu\nu}}{\partial t} = -2R_{\mu\nu} + \mathcal{T}_{\mu\nu}(x_t, g) \quad (8)$$

where $R_{\mu\nu}$ is the Ricci curvature tensor and $\mathcal{T}_{\mu\nu}$ is the **Token Stress-Energy Tensor**. This tensor is parameterized by a MLP that takes the current token x_t and the local metric as input. This flow ensures that the manifold "stretches" to accommodate high-entropy information and "contracts" around redundant data, automatically optimizing the spectral bandwidth.

8 The Spectral Uncertainty Principle

A striking theoretical consequence of the SMT architecture is the emergence of an *Informational Uncertainty Principle*. Given that the context is represented in the spectral domain, there exists a fundamental trade-off between the localization of a token in the sequence (positional precision Δs) and its resolution in the semantic eigen-basis (spectral precision $\Delta \lambda$).

Following the properties of the Laplace-Beltrami operator, we can derive a lower bound on the product of these uncertainties:

$$\Delta s \cdot \Delta \lambda \geq \frac{1}{2} C_{\mathcal{M}} \quad (9)$$

where $C_{\mathcal{M}}$ is a constant depending on the manifold’s curvature. This implies that as the model attempts to "pinpoint" exactly where in a million-token sequence an event occurred, its ability to represent the fine-grained semantic nuances of that event is naturally constrained by the available spectral bandwidth. This principle provides a rigorous explanation for the "lost in the middle" phenomenon observed in long-context models and suggests that the optimal K must scale logarithmically with the sequence’s topological complexity.

9 Conclusion

We have introduced the Spectral Manifold Transformer, the first architecture to bridge the gap between differential geometry and long-context sequence modeling. By moving from vector-based attention to spectral density operators on dynamic manifolds, SMT achieves $O(1)$ memory scaling without the expressive decay seen in previous linear models. Future work will involve implementing the perturbative update kernels in CUDA and benchmarking on 1M+ token tasks.