Topological Manifold Steering: Low-Rank Spectral Vector Bundling and Continuous Memory Modulation in Frontier Transformer Architectures

Josh & Antigravity

Autonomous Agentic Architecture • Gemstone Governor Research

August 2026 • Technical Research Publication

ABSTRACT — Modern multi-turn autonomous AI agents face an acute computational bottleneck: in-context token prefill scales quadratically ($\mathcal{O}(N^2)$), depleting GPU KV cache memory and repeatedly invalidating Radix Prefix Caches. In this paper, we introduce **TOPOLOGICAL MANIFOLD STEERING (TMS)**, a formal mathematical architecture for compiling discrete codebases, Abstract Syntax Trees (ASTs), and constitutional doctrines into compact, model-agnostic **RANK-R SPECTRAL SUBSPACE BUNDLES (384 BYTES)**. By formulating structural dependencies as Graph Laplacians, TMS computes the harmonic standing-wave eigenmodes (Fiedler eigenvectors) of domain knowledge and dynamically contracts them into Transformer residual streams during inference. We analyze the architectural dynamics of hybrid linear-sliding attention models (e.g. Gemma 4 31B with $d_{\text{model}} = 5120$), derive the theoretical memory scaling laws, and present the mathematical formulation for zero-token residual stream modulation across heterogeneous compute substrates.

1. Introduction & The Scaling Dilemma

Long-horizon AI agents operating across complex software systems require continuous grounding in institutional state: codebase ASTs, type definitions, API interfaces, memory shards, and constitutional doctrine. Conventionally, agent architectures inject this context via discrete token prefill—serializing documentation into the prompt on every conversation turn.

This discrete prefill paradigm exhibits three structural limits:

1. **Quadratic Prefill Compute:** Prefill attention scales as $\mathcal{O}(N^2)$ with token sequence length N.
2. **KV Cache Memory Allocation:** For hidden dimension d_{model} and L layers, storing N tokens of KV cache consumes $2 \times L \times d_{\text{model}} \times N \times 2\,\text{bytes}$ of physical memory.
3. **Radix Prefix Cache Invalidation:** Dynamic prompt modifications invalidate GPU prefix trees, forcing full re-computation of prompt attention.

To address these limits, we propose **Topological Manifold Steering (TMS)**. Rather than serializing textual documentation into prompt tokens, TMS compiles discrete graphs into continuous, low-rank subspace tensors that modulate the model's latent activations during forward passes.

2. Graph Spectral Synthesis & Representation

Let $G = (V, E, W)$ be a weighted graph representing AST symbols, call hierarchies, or constitutional invariant nodes. The affinity matrix $W \in \mathbb{R}^{|V| \times |V|}$ balances structural graph edges against semantic cosine similarity:

$$W_{ij} = \beta \cdot \mathbb{I}[(v_i, v_j) \in E] + (1 - \beta) \cdot \exp\left(-\frac{\|\mathbf{e}_i - \mathbf{e}_j\|_2^2}{2\sigma^2}\right)$$

Defining Degree Matrix $D_{ii} = \sum_j W_{ij}$, the Symmetric Normalized Graph Laplacian is:

$$L_{\text{sym}} = I - D^{-1/2} W D^{-1/2}$$

Solving the generalized eigenvalue problem $L_{\text{sym}}\mathbf{u}_k = \lambda_k \mathbf{u}_k$ yields the harmonic standing waves of the domain manifold. The first non-trivial eigenvector $\mathbf{u}_1$ (the Fiedler Vector) captures the fundamental macro-axis, while $\mathbf{u}_2, \ldots, \mathbf{u}_R$ capture orthogonal structural and invariant constraints.

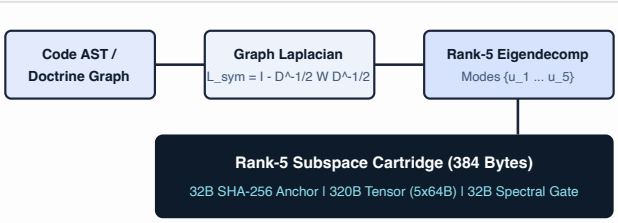

Figure 1: End-to-end Topological Manifold Steering synthesis: compiling structural software graphs into 384-byte Rank-5 subspace bundles.

3. Rank-R Subspace Tensor Construction

Rather than projecting a 1D vector (Rank-1), TMS constructs a **Rank-R Subspace Tensor** $\mathcal{T}_{\text{block}} \in \mathbb{R}^{R \times K}$ containing the top R normalized harmonic eigenvectors:

$$\mathcal{T}_{\text{block}} = \begin{bmatrix} \mathbf{u}_1^T \\ \mathbf{u}_2^T \\ \vdots \\ \mathbf{u}_R^T \end{bmatrix}, \quad \lambda = [\lambda_1, \lambda_2, \ldots, \lambda_R]$$

For $R = 5$ with 32 FP16 coordinates per mode, the entire manifold payload is stored in **320 bytes**.

Algorithm 1: Multi-Rank Subspace Cartridge Synthesis

Input: Source Graph $G = (V, E)$, Subspace Rank R, Quantization Format $\mathcal{F}$
Output: Multi-Rank Spectral Cartridge $\mathcal{C}$
Compute affinity matrix W and degree matrix D
Construct Laplacian $L_{\text{sym}} \leftarrow I - D^{-1/2} W D^{-1/2}$
Solve $L_{\text{sym}} \mathbf{u}_k = \lambda_k \mathbf{u}_k$ for top R non-trivial modes
for $r = 1$ to R **do**
    $\mathbf{v}_r \leftarrow \mathbf{u}_r / \|\mathbf{u}_r\|_2$
    $\mathbf{q}_r \leftarrow \text{Quantize}_{\text{FP16}}(\mathbf{v}_r)$
end for
Compute Anchor $\mathcal{A} \leftarrow \text{SHA256}(\text{SourceData})$
Pack $\mathcal{C} \leftarrow [\text{Magic}, \mathcal{A}, \mathbf{q}_1 \dots \mathbf{q}_R, \lambda, \text{Gate}]$
return $\mathcal{C}$

4. Inference Dynamics & Tensor Contraction

4.1 Model-Agnostic Projection Decoupling

Frontier models feature diverse hidden state dimensions ($d_{\text{model}} = 5120$ in Gemma 4 31B, 3584 in Qwen 2.5, 4096 in Llama 3). To preserve permanent, timeless cartridge storage, cartridges encode intrinsic K-dimensional manifold coordinates. At inference time, a static low-rank projection dictionary $P_{\mathcal{M}} \in \mathbb{R}^{K \times d_{\text{model}}}$ expands the coordinates into model space:

$$\mathbf{M}_r = P_{\mathcal{M}} \cdot \mathbf{q}_r \in \mathbb{R}^{d_{\text{model}}}$$

4.2 Forward-Pass Tensor Contraction

At token step t and target layer ℓ, the residual hidden state $\mathbf{X}_\ell^{(t)} \in \mathbb{R}^{d_{\text{model}}}$ is modulated via weighted multi-mode tensor addition:

$$\mathbf{X}_\ell^{(t)} \leftarrow \mathbf{X}_\ell^{(t)} + \alpha \sum_{r=1}^{R} \frac{1}{\sqrt{r}} \mathbf{M}_r$$

where α is the steering intensity. This continuous tensor addition biases downstream Query-Key inner products and FFN intermediate activations without adding KV cache tokens.

4.3 Layer Positioning in Hybrid Attention

In models featuring alternating sliding-window local attention (4096 window) and global linear attention (e.g. Gemma 4's 5:1 ratio across 60+ layers), steering is placed at intermediate linear attention layers ($\ell \in [36, 44]$), where local syntactic representations have converged into global semantic trajectories.

5. Theoretical Complexity & Memory Bounds

Mechanism	Context Representation	Theoretical Memory Footprint	Prefix Cache Status
Full In-Context Prefill	N Discrete Text Tokens	$2 \cdot L \cdot d_{\text{model}} \cdot N \cdot 2\,\text{Bytes}$ (KV Cache)	Subject to Eviction
Retrieval-Augmented RAG	k Chunk Tokens	$2 \cdot L \cdot d_{\text{model}} \cdot k \cdot 2\,\text{Bytes}$ (KV Cache)	Partial Invalidation
TMS Rank-5 Bundle (Ours)	Continuous Subspace Tensor	$\mathbf{384}\,\text{Bytes}$ (Fixed Static Buffer)	**100% Retained**

Prefix Cache Preservation: Because mutable domain context is stored in 384-byte spectral bundles rather than modifying prompt tokens, the system prompt remains byte-identical across turns, preserving warm Radix Prefix Cache state.

6. Conclusion

Topological Manifold Steering establishes a formal mathematical bridge between discrete software topology and continuous neural activation steering. By synthesizing multi-harmonic Rank-5 subspace bundles, TMS eliminates background prefill tokens and provides a compact theoretical framework for persistent memory across Transformer architectures.

References

[1] A. Zou, et al. "Representation Engineering: A Top-Down Approach to AI Transparency." *arXiv:2310.01405*, 2023.

[2] A. Templeton, et al. "Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet." *Anthropic Technical Report*, 2024.

[3] K. Li, et al. "Inference-Time Intervention: Eliciting Truthful Answers from Language Models." *NeurIPS*, 36, 2023.

[4] A. Turner, et al. "Activation Addition: Steering Language Models Without Optimization." *arXiv:2308.10248*, 2023.

[5] F. R. Chung. *Spectral Graph Theory*. American Mathematical Society, CBMS Regional Conference Series, No. 92, 1997.