SELF-ATTENTION AS CONNECTION PROPAGATION: AN EXACT GEOMETRIC OPERATOR REPRESENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-attention is a ubiquitous primitive in modern sequence models, yet its operator-level geometry is only partially understood. We view a token sequence as a *vector field over the token-position graph* and identify attention as a *connection walk*: messages are aggregated by a nonnegative walk matrix while being *transported* along each edge by a learned linear map. Within this framework, we derive an exact representation of *single-head attention (SHA)* through a connection propagation step with *constant* transport, and of *multi-head attention (MHA)* through a *single edge-dependent connection walk* whose effective transport is an attention-gated mixture of headwise transports. We further clarify the conditions under which the corresponding operator reduces to a *random-walk connection Laplacian*, highlighting the roles of stochasticity, reversibility, and metric-compatible transports. Empirically, we examine trained Transformers across scales and structures (encoder/decoder). Layerwise diagnostics reveal low-drift middle-layer routing regimes and pronounced diagonal structure in averaged transport Gram matrices, while a separate selected-edge spectral analysis reveals concentrated transport energy with unequal directional gains. Overall, the paper provides a precise connection-walk formalism that links self-attention to classical geometric operators, along with a set of operator-level tools for analyzing transformer models from a geometric perspective.

1 INTRODUCTION

Self-attention aggregates value vectors using input-dependent weights, enabling Transformers to capture long-range token dependencies (Vaswani et al., 2017). We give this computation an operator-level geometric interpretation: token representations form a vector field over the token-position graph (the token graph), and multi-head attention (MHA) admits an exact connection-walk representation. The blocks $T_{ij} = A_{ij}O_{ij}$ combine scalar routing weights A_{ij} with edgewise feature maps O_{ij}. This formulation distinguishes token routing, feature transport, and their joint action (Figure 1).

1.1 RELATED WORK

Graph operator view. Graph Transformer Networks learn graph structures through soft edge-type selection and meta-path composition (Yun et al., 2019). Spectral graph transformers combine Laplacian positional information with fully connected attention (Kreuzer et al., 2021), while analyses of attention masks and LayerNorm examine their effects on rank collapse (Wu et al., 2024).

Tensor operator view. Elhage et al. (2021) separate query–key routing from value–output maps and express headwise attention using Kronecker products. Tensor approaches extend attention to multi-way inputs through matricization (Babiloni et al., 2020) and use tensor decompositions to reduce parameter complexity (Ma et al., 2019). Factorizing queries, keys, and values into contextual low-rank components can also reduce the attention cache (Zhang et al., 2025).

PDE and dynamical-systems view. Continuous-depth Transformer formulations treat block parameters as functions of depth and analyze the resulting non-autonomous neural ODE (Tong et al., 2025). Other formulations interpret Transformer updates as forward-Euler steps and introduce trajectory-level regularization to stabilize the evolution (Kan et al., 2025).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

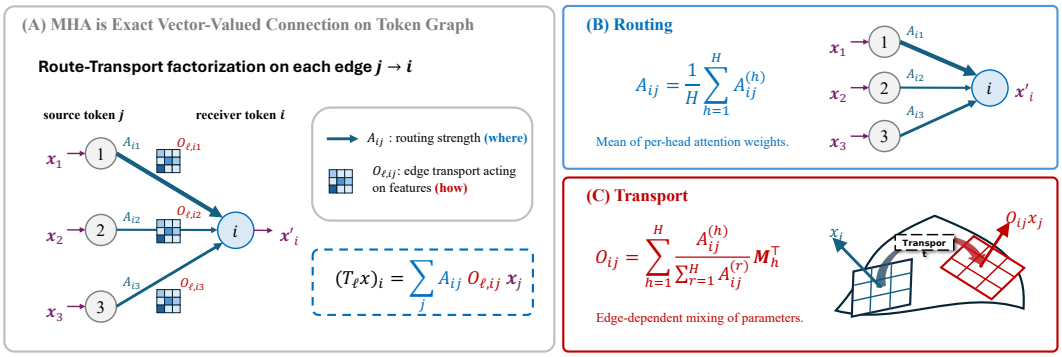


Figure 1: **Multi-head attention as a connection walk.** (A) Each message combines scalar routing A_{ij} with a feature map O_{ij}. (B) Routing averages the per-head attention weights. (C) The transport is a conditional mixture of the headwise value–output maps $M_h^\top$. The diagram shows the normalized walk $T(A, O)$; native attention is $H\mathcal{T}(A, O)X + \mathbf{1}b$, retaining head-count scaling and the affine term (Section 4.2). The feature maps are general linear maps and need not be orthogonal.

Energy-functional view. Modern Hopfield networks connect attention-like retrieval to energy minimization and associative-memory fixed points (Ramsauer et al., 2021). This connection clarifies retrieval dynamics and motivates nonlinear attention variants (Farooq, 2025).

Kernel and approximation view. Kernel interpretations support both approximation and expressivity analyses. Performer approximates softmax attention with positive random features for linear-time computation (Choromanski et al., 2021); broader efficient-attention methods are reviewed by Tay et al. (2023). Other analyses characterize attention through non-Mercer kernels on Banach spaces, with representer-style and universal-approximation results (Wright & Gonzalez, 2021), or study learning explicit Transformer kernels (Chowdhury et al., 2021).

Geometric and gauge-equivariant view. Gauge-equivariant networks formalize local frame changes and constrain how operators act on features (Cohen et al., 2019; Bronstein et al., 2021). Vector diffusion maps and graph connection Laplacians describe diffusion of vector fields with edge transports (Singer & Wu, 2012; Bandeira et al., 2013). Sheaf neural networks and neural sheaf diffusion equip graphs with vector spaces and linear maps, then define diffusion through sheaf or connection Laplacians (Bodnar et al., 2022; Barbero et al., 2022).

Our formulation identifies standard MHA as a *single edge-dependent connection-walk operator*, coupling query-key routing to conditional mixtures of headwise value-output maps. We derive exact connection-walk representations of single-head attention (SHA) and MHA (§4) and identify conditions under which the induced operator reduces to a random-walk connection Laplacian (§5.4). The resulting analysis establishes a worst-case amplification constraint for causal routing, a value-group bound on transport rank, and bounds on the joint action and approximation error (§5). We then turn these operators into diagnostics and report their behavior across trained Transformers from 124M to 8B parameters (§6): layerwise diagnostics reveal structured routing and diagonal patterns in averaged transport Gram matrices, while a separate selected-edge spectral analysis reveals concentrated transport energy with unequal directional gains.

2 Preliminaries: Attention and Operator Conventions

An attention matrix specifies which tokens communicate. The value–output maps specify what is communicated. The connection-walk framework below combines these structures into a single operator and separates their roles in communication.

Let $X \in \mathbb{R}^{n\times d}$ contain token features as rows, with positions $i, j \in \{0, \dots, n-1\}$. The entry $A_{ij}^{(h)}$ is the weight from source j to destination i in head h. We work at evaluation time, with attention dropout disabled and at least one admissible source per row. Thus $A^{(h)}$ is nonnegative and row-stochastic. Causal masking additionally gives $A_{ij}^{(h)} = 0$ for $j > i$. Fix the input that constructs these weights.

For a test field V of the same shape, the affine-free attention action is

$$\mathcal{F}_X(V) = \sum_{h=1}^{H} A^{(h)} V M_h, \qquad M_h = W_{V,h} W_{O,h} \in \mathbb{R}^{d\times d}, \tag{1}$$

where $W_{O,h}$ is the output-projection slice for head h. Here H counts query heads; shared value groups are treated separately below. The native affine output is $\mathcal{F}_X(X) + \mathbf{1}b$, where $b = \sum_h b_{V,h} W_{O,h} + b_O$ when these biases are present. Query/key biases belong to the construction of $A^{(h)}$.

We use $\mathrm{vec}^r(V) = \mathrm{vec}(V^\top)$ to stack column token vectors. A row-action map M_h therefore appears as $M_h^\top$ in the stacked operator:

$$T_{\mathrm{native}}[X] = \sum_h A^{(h)} \otimes M_h^\top, \qquad \mathrm{vec}^r(\mathcal{F}_X(V)) = T_{\mathrm{native}}[X]\mathrm{vec}^r(V). \tag{2}$$

This is a linear operator on V for each constructing input X; it is not the Jacobian of attention with respect to X. Appendix B.1 gives the vectorization derivation and bias bookkeeping.

3 Connection Walks and Connection Laplacians

A scalar walk aggregates values from neighboring vertices. A connection walk also transforms the feature vector carried along each edge. This distinction lets us describe token selection and feature transformation in one graph operator.

3.1 Vector fields and edge transport

Let $G = (\mathcal{V}, E)$ be a directed graph with n vertices. A vector field assigns a column vector $v_i \in \mathbb{R}^d$ to each vertex i; equivalently, its row-array representation has $V_i = v_i^\top$. Let A be nonnegative and row-stochastic, with $A_{ij} > 0$ only when $(j \to i) \in E$. Row i is a distribution over sources used to update destination i. Assign an edge map $O_{ij} : \mathbb{R}^d \to \mathbb{R}^d$ taking source features into destination coordinates. We allow arbitrary linear maps, including singular maps. The term *connection walk* refers to this algebraic propagation rule; metric connections impose additional constraints.

Definition 1 (Connection walk)**.** The connection-walk operator $T(A, O) \in \mathbb{R}^{nd\times nd}$ has blocks

$$[T(A,O)]_{ij} = A_{ij} O_{ij}, \qquad (T(A,O)v)_i = \sum_j A_{ij} O_{ij} v_j. \tag{3}$$

Its corresponding row-array action is

$$(\mathcal{T}(A,O)V)_i = \sum_j A_{ij} V_j O_{ij}^\top, \qquad \mathrm{vec}^r(\mathcal{T}(A,O)V) = T(A,O)\mathrm{vec}^r(V). \tag{4}$$

When $O_{ij} = I_d$ on every active edge, this reduces to ordinary averaging of vector features. When all active edges share a map O_0, the block operator is $A \otimes O_0$. General edge-dependent transport retains the block product $T_{ij} = A_{ij} O_{ij}$, but need not be a single Kronecker product.

Repeated application transports features along graph paths: scalar route weights multiply, while feature maps compose in path order. Appendix B.2 gives the path expansion and the continuous-time propagator for a fixed connection.

3.2 The connection Laplacian and parallel fields

Define the random-walk Laplacian-sign operator and its generator by

$$L(A,O) = I - T(A,O), \qquad \mathcal{G} = T(A,O) - I = -L(A,O). \tag{5}$$

Then $(Lv)_i$ is the discrepancy between v_i and its transported neighbor average. A field is *parallel* along active edges when $v_i = O_{ij} v_j$ for every $A_{ij} > 0$. Row stochasticity immediately gives $Tv = v$ for such a field, so parallel fields lie in $\ker L$. The converse and a nonnegative edge-energy interpretation hold in the metric regime established in Proposition 5.

These definitions apply to arbitrary linear edge maps. The remaining step is to identify attention's routing weights and transports; the next section gives the exact construction.

4 SELF-ATTENTION AS A CONNECTION WALK

The connection construction applies exactly to standard attention. A single head uses one map on all active edges; multiple heads induce edge-dependent maps through their relative attention weights.

4.1 SINGLE-HEAD ATTENTION: CONSTANT TRANSPORT

Proposition 1 (Single-head connection walk)**.** *For row-stochastic A and row-action head map $M = W_V W_O$, the affine-free single-head action $V \mapsto AVM$ is the connection walk with $O_{ij} = M^\top$ on every active edge. Its stacked operator is $T = A \otimes M^\top$.*

Proof. At each destination, $(AVM)_i = \sum_j A_{ij} V_j M$, which is Eq. (4) with $O_{ij}^\top = M$. □

4.2 MULTI-HEAD ATTENTION: CONDITIONAL FEATURE TRANSPORT

For the realized head weights in Eq. (1), define head-mean routing and conditional head weights by

$$A_{ij} = \frac{1}{H} \sum_h A_{ij}^{(h)}, \qquad \alpha_h(i,j) = \frac{A_{ij}^{(h)}}{\sum_r A_{ij}^{(r)}} \quad \text{when } A_{ij} > 0. \tag{6}$$

The column-action feature map at an active edge is

$$O_{ij} = \sum_h \alpha_h(i,j) M_h^\top. \tag{7}$$

Set $O_{ij} = 0$ on inactive edges; any completion gives the same attention action.

Theorem 1 (Exact multi-head connection walk)**.** *Under the evaluation-time assumptions of Section 2, A is row-stochastic and the affine-free native operator is*

$$T_{\text{native}}[X] = \sum_h A^{(h)} \otimes M_h^\top = H\,T(A,O). \tag{8}$$

Consequently, for every test field V, $\mathcal{F}_X(V) = H\mathcal{T}(A,O)V$, and the native attention output is

$$\text{Attn}(X) = H\mathcal{T}(A,O)X + \mathbf{1}b. \tag{9}$$

The identity holds for masked and grouped-query attention.

Proof. Averaging the head row sums gives $\sum_j A_{ij} = 1$. On an active edge,

$$H A_{ij} O_{ij} = \sum_h A_{ij}^{(h)} M_h^\top,$$

which is the (i,j) block of Eq. (2). On an inactive edge, nonnegativity implies every $A_{ij}^{(h)} = 0$, so both blocks vanish. Row stochasticity makes each value bias a token-independent output contribution, giving the affine term from Section 2. □

For the empirical displays, the corresponding row-action notation is $A_{\text{eff}} = A$ and $O_{ij}^{\text{eff}} = O_{ij}^\top$; Appendix B.1 gives the Gram-matrix correspondence.

The factor H follows from the head-mean convention and must be retained when reconstructing the native output. The construction is exact, without requiring invertible, orthogonal, or reciprocal feature maps. It separates routing from feature transformation while retaining their coupling: the same head weights determine both A_{ij} and the conditional mixture O_{ij}. The next section examines which properties each factor determines and which depend on their joint connection action.

5 GEOMETRY OF THE ATTENTION CONNECTION

Exact connection structure does not by itself determine amplification or metric diffusion. Routing can amplify shared source information, while the value pathway restricts which feature directions are transmitted. Their joint operator determines the gain on a field. The results below separate these constraints and give sufficient conditions for a metric connection Laplacian, using standard matrix and graph-operator arguments.

5.1 ROUTING: SELECTIVITY AND AMPLIFICATION

Row stochasticity fixes the spectral radius $\rho(A) = 1$, but does not fix the Euclidean operator norm. Source concentration and norm amplification measure different aspects of communication.

Proposition 2 (Routing gain and causal mixing)**.** *For a nonnegative row-stochastic $A \in \mathbb{R}^{n\times n}$,*

$$1 \le \|A\|_2 \le \sqrt{\|A\|_1} \le \sqrt{n}. \tag{10}$$

If A is also lower triangular, $\|A\|_2 \le 1$ holds if and only if $A = I$.

The lower bound follows from $A\mathbf{1} = \mathbf{1}$; the upper bound follows from $\|A\|_2^2 \le \|A\|_1\|A\|_\infty$ and $\|A\|_\infty = 1$. For the causal conclusion, non-expansion and $A\mathbf{1} = \mathbf{1}$ force column sums one; a nonnegative lower-triangular doubly stochastic matrix is the identity. Appendix B.3 gives the proof.

Identity routing and a pure first-position sink $\mathbf{1}e_0^\top$ both have maximally concentrated rows, yet their norms are 1 and $\sqrt{n}$. The difference is whether many destinations read the same source. Consequently, any nontrivial causal mixing has a worst-case Euclidean gain above one. This constraint concerns routing alone, not the full attention action. For a nonzero native field X, the realized routing gain $\|AX\|_F/\|X\|_F$ can be much smaller than $\|A\|_2$. Their depthwise changes can differ.

5.2 FEATURE TRANSPORT: A CONDITIONAL DICTIONARY WITH LIMITED CAPACITY

Routing identifies the selected sources, but leaves the transformation of their messages unspecified. That transformation is constrained by the value pathway: each active O_{ij} lies in the convex hull of the column-action head maps $\{M_h^\top\}_{h=1}^H$. Heads therefore provide a shared transport dictionary, while conditional head weights select its mixture separately at each edge. For fixed head maps, the affine dimension of this family is at most $H - 1$. This dictionary dimension is different from the rank of an individual map: even a single dictionary element can have high rank within its head capacity.

Proposition 3 (Value-group capacity)**.** *Suppose G value groups have shared maps $W_{V,g} \in \mathbb{R}^{d\times d_v}$, and head h uses group $g(h)$. Then every effective edge map satisfies*

$$\operatorname{rank}(O_{ij}) \le \min(d, Gd_v). \tag{11}$$

For ordinary MHA, $G = H$. For grouped-query attention, the bound depends on value groups rather than query-head count.

Indeed, the row-action map is $O_{ij}^\top = \sum_g W_{V,g}B_{g,ij}$ with $B_{g,ij} = \sum_{h:g(h)=g} \alpha_h(i,j)W_{O,h}$, so it factors through a Gd_v-dimensional value space. Appendix B.4 states the factorization explicitly. When $Gd_v \ge d$, this bound alone implies no rank reduction below the ambient width.

Architecture bounds available capacity; learned weights and conditional mixtures determine its use. If $O_{ij} = U\Sigma V^\top$, singular values specify available gains, while the right and left singular vectors specify input and output directions. Low effective rank and isotropy are separate properties: retaining most energy in a few directions does not make the gains equal within those directions. Nor do matching spectra imply matching maps or matching actions. For example, O and $-O$ have identical singular values and opposite nonzero actions. Appendix B.6 relates these distinctions to polar decomposition.

5.3 JOINT ACTION: AGGREGATION AND ORIENTATION

Routing gain and edge-map capacity describe the factors separately. Aggregation couples them: messages transformed into different directions can reinforce or cancel at the destination. Put $s_{ij} = \|O_{ij}\|_2$. Their combined gain obeys

Proposition 4 (Joint gain bound)**.** *For nonnegative A and arbitrary edge maps,*

$$\|T(A,O)\|_2 \le \|A \odot s\|_2. \tag{12}$$

When $O_{ij} = O_0$ on every active edge, $T = A \otimes O_0$ and $\|T\|_2 = \|A\|_2\|O_0\|_2$.

The proof compares each output-token norm with the scalar action of $A \odot s$ on the input-token norms (Appendix B.5). General MHA need not attain this bound: differently oriented messages can cancel or reinforce. The same blockwise argument bounds approximation errors using the matrix of error-block norms, as detailed in that appendix.

For a nonzero native input field, write $\widehat{X} = X/\|X\|_F$. Then

$$\|\operatorname{Attn}(X) - \mathbf{1}b\|_F = H\|X\|_F\,\|\mathcal{T}(A,O)\widehat{X}\|_F. \tag{13}$$

Head-count scale, input magnitude, and directional gain all enter the native update. Thus changes in A, spectra of O, and the action of T should be measured separately. The representation relates these measurements; it does not assign a causal explanation to correlations between their depth profiles.

5.4 From the attention walk to a metric connection Laplacian

The joint gain bound controls amplification without assuming that propagation is a diffusion. The following sufficient conditions make routing and feature transport compatible with a common metric and give $L = I - T$ that interpretation. For positive weights π_i summing to one, use $\langle v, w\rangle_\pi = \sum_i \pi_i v_i^\top w_i$ on column token fields. The classical metric connection regime requires additional structure (Singer & Wu, 2012; Bandeira et al., 2013).

Proposition 5 (Contraction and the metric connection Laplacian)**.** *If A is row-stochastic, $\pi^\top A = \pi^\top$, and $\|O_{ij}\|_2 \le 1$ on active edges, then $\|Tv\|_\pi \le \|v\|_\pi$. If, in addition, routing is reversible, $\pi_i A_{ij} = \pi_j A_{ji}$, and active transports are orthogonal and reciprocal, $O_{ji} = O_{ij}^\top$, then L is self-adjoint and positive semidefinite in this geometry, with*

$$\langle v, Lv\rangle_\pi = \frac{1}{2}\sum_{i,j} \pi_i A_{ij}\|v_i - O_{ij}v_j\|_2^2. \tag{14}$$

Reciprocity includes active self loops.

These statements concern normalized T, not the native head sum HT or a residual block. The first assertion follows from convexity of squared norm and stationarity. Reversibility and reciprocal orthogonal maps supply self-adjointness and the quadratic-form identity; Appendix B.7 proves both assertions. In this regime the kernel of L consists exactly of parallel fields on active edges, and the semigroup e^{-tL} is contractive in the weighted norm. The nonnegative spectrum of L gives the metric diffusion interpretation of the transported-walk propagator (Appendix B.2).

Generic attention satisfies neither metric compatibility nor reversibility. In particular, causal cross-token routing cannot be reversible under a strictly positive π: a positive lower-triangular off-diagonal weight has no reverse edge. With arbitrary feature maps, even row-stochastic routing does not guarantee contraction, as the common map $O_0 = cI$ with $c > 1$ illustrates. For the general generator, its self-adjoint part determines instantaneous quadratic growth or decay and need not be negative semidefinite; its skew-adjoint part has zero quadratic form. Transient growth is a property of their joint evolution, not an effect attributable to the skew part alone. The metric Laplacian is therefore a special regime of the exact representation, not a general description of trained attention.

5.5 Consequences for interpreting and approximating attention

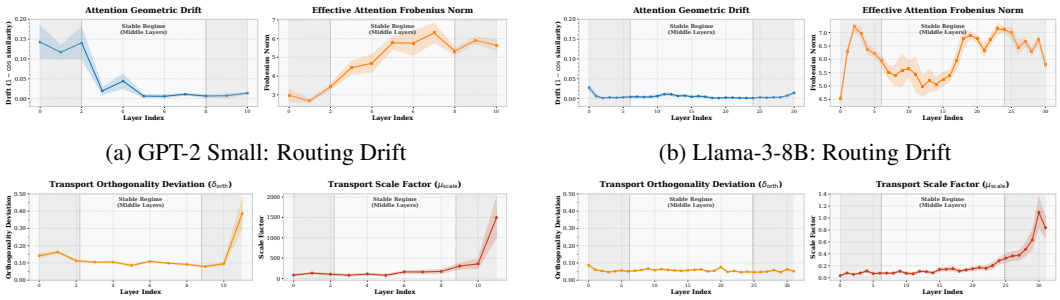
Attention gain depends jointly on routing, feature transport, and native head-count scaling. For causal routing, enforcing $\|A\|_2 \le 1$ would eliminate cross-token mixing (Proposition 2).

At fixed routing and with the native affine term preserved, discarding an edge without renormalization incurs a block error of size $A_{ij}\|O_{ij}\|_2$ in the normalized operator. A small routing weight can therefore carry a substantial message. Appendix B.5 bounds the resulting operator error; the realized error also depends on input feature directions, and prediction loss requires downstream evaluation.

The experiments examine routing similarity, source concentration, transport geometry, and effective capacity across depth. Context-length comparisons track routing and Gram diagnostics.

6 Experiments

The connection representation separates source selection through A from feature transmission through O. We examine their depthwise structure in trained Transformers, then ask how the observed patterns depend on context length. Routing curves and operator heatmaps establish the coordinate-level picture; singular values reveal how transport capacity is distributed across feature directions.



(a) GPT-2 Small: Routing Drift

(b) Llama-3-8B: Routing Drift

(c) GPT-2 Small: Gram Deviation

(d) Llama-3-8B: Gram Deviation

Figure 2: **Layerwise diagnostics of effective walks and transports.** Shaded regions mark the initial and final layers. **(a, b) Adjacent routing drift:** GPT-2 Small shows greater variation across depth, whereas Llama-3-8B exhibits a low-drift middle-layer regime ($d_A^{\ell\to\ell+1} \approx 0$). **(c, d) Transport Gram deviation:** Llama-3-8B has lower dimension-normalized deviation from a scalar Gram matrix ($\delta_{\text{orth}} \approx 0.05$ in middle layers) than GPT-2 Small. This diagnostic depends on feature dimension and does not by itself establish scaled isometry.

6.1 Experimental setup

The layerwise diagnostics cover GPT-2 (Small, Medium, Large, XL), Qwen2.5 (3B, 7B), Qwen3 (4B, 8B), Llama-3.2 (3B), and Llama-3 (8B). We sample 1,024 WikiText-2 sequences (Merity et al., 2017), use context 64 unless otherwise specified, and report average statistics. Figures 2 and 3 summarize the effective operators in Eqs. (6)–(7); Appendix D gives complete profiles.

A separate fp64 spectral study examines selected non-position-zero edges across the full depth of seven decoder checkpoints, also on WikiText-2 at context 64, with up to 32 contributing windows per checkpoint. Its population differs from the weighted all-pair Gram diagnostics. We use GPT-2 Small and Qwen3-1.7B as representative MHA/GQA cases; Appendix C gives the complete checkpoint list, edge-selection rules, coverage, and aggregation.

6.2 Routing structure across depth

We first ask how source selection changes between adjacent layers. The routing diagnostic is the cosine distance between effective attention matrices on valid support $\mathcal{M}$:

$$d_A^{\ell\to\ell+1} = 1 - \cos_\text{sim}\big(\text{vec}_{\mathcal{M}}(A_{\text{eff}}^{\ell+1}), \text{vec}_{\mathcal{M}}(A_{\text{eff}}^{\ell})\big).$$

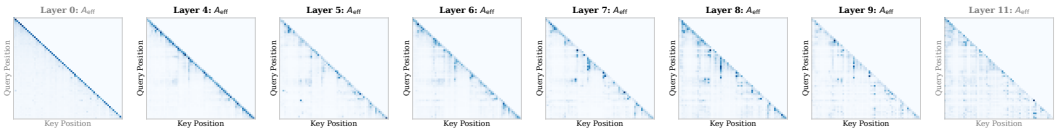
For decoder-only models, the support is causal lower-triangular. Smaller d_A means greater adjacent-layer similarity. The Frobenius norm $\|A_{\text{eff}}^{\ell}\|_F$ measures a different property: for fixed token count and row-stochastic routing, a larger norm indicates more concentrated source weights.

The upper panels of Figure 2 show a lower-drift middle-layer regime, more pronounced in Llama-3-8B than in GPT-2 Small. The accompanying Frobenius-norm curves distinguish similarity across depth from concentration within a layer. A routing pattern can remain similar between adjacent layers while assigning most of its mass to a restricted set of sources.

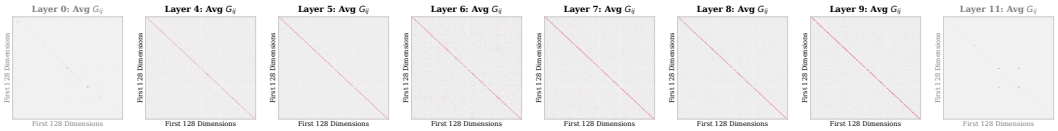
Figure 3a makes the routing structure visible for GPT-2 Small after removing the first-token attention sink. The displayed middle layers exhibit coherent token-interaction patterns. Together, the curves and heatmaps support structured routing with substantial adjacent-layer similarity in these regimes. They describe source selection; update similarity also depends on edge transports.

6.3 Transport geometry and effective capacity

We next ask how the selected messages are transformed. Proposition 3 bounds the available transport rank through the value groups, but leaves the distribution of gains within that capacity open. The Gram displays characterize coordinate-level structure; the singular-value analysis then resolves energy concentration and directional gains.

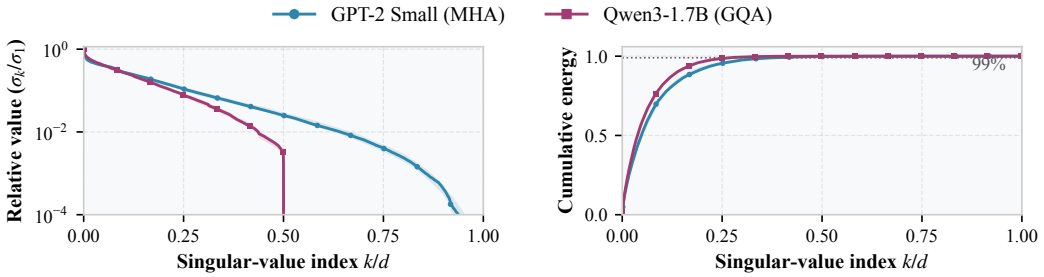


(a) Evolution of Effective Walk Weights A_{eff} (First token attention sink removed)



(b) Evolution of Transport Gram Matrix G_{ij} (Subset of 128 channels)

Figure 3: **Visualization of Learned Operators (GPT-2 Small). (a) Effective walk weights (A_{eff}):** With the attention sink removed, the displayed middle-layer attention maps show structured token interactions. **(b) Transport Gram structure (G_{ij}):** The displayed mean Gram matrix has pronounced diagonal structure. This does not establish orthogonality of individual transports: diagonal entries may be unequal, and averaging can suppress off-diagonal terms.



(a) Singular-value decay

(b) Cumulative transport energy

Figure 4: **Transport energy concentrates in a restricted set of directions.** Selected-edge spectra across depth for GPT-2 Small and Qwen3-1.7B, with 32 text windows each at context 64. Lines show pointwise window medians and bands show interquartile ranges after within-window edge/layer aggregation (Appendix C). (a) Singular values relative to the largest, displayed down to 10^{-4}. (b) Cumulative squared-singular-value energy, with the dotted line marking 99%.

Diagonal Gram structure. For the row-action maps used in the empirical displays, write $G_{ij}^{\ell} = O_{ij}^{\text{eff},\ell\top} O_{ij}^{\text{eff},\ell}$. Figure 3b shows pronounced diagonal structure in the averaged Gram matrix. The corresponding diagonal-to-off-diagonal magnitude ratios are approximately 15–20 in the reported middle-layer examples (Appendix E). This coordinate structure does not require uniform gains: even a diagonal Gram matrix can have unequal diagonal entries, and averaging can suppress off-diagonal terms from individual edges.

The lower panels of Figure 2 and Table 1 quantify Gram deviation across depth and checkpoints. For $\mu = \mathrm{Tr}(G)/d$, the reported diagnostic is $\delta_{\text{orth}} = \|G - \mu I_d\|_F/(\mu d)$, with weighted means over valid token pairs. Llama-3-8B has a lower middle-layer profile than GPT-2 Small, reaching values near 0.05; Table 1 extends the comparison to the other checkpoints. Its ordering is not uniformly monotone with model size. For a nonzero map, $\sqrt{d}\,\delta_{\text{orth}}$ is the relative Frobenius distance to μI_d, so these dimension-dependent values alone do not establish increasing proximity to scaled isometry across model widths.

Concentrated capacity with unequal gains. Singular values address the remaining geometric question directly. For $\sigma_1 \geq \cdots \geq \sigma_d$, the cumulative energy is $E(k) = (\sum_{a=1}^{k} \sigma_a^2)/(\sum_{a=1}^{d} \sigma_a^2)$. Define $r_{99} = \min\{k : E(k) \geq 0.99\}$; the fraction r_{99}/d measures the capacity carrying 99% of squared-singular-value energy, rather than exact algebraic rank.

Model	Relative Layer Depth		
	Initial (20%)	Middle (60%)	Final (20%)
GPT-2 Family			
Small (124M)	0.152	0.096	0.284
Medium (355M)	0.180	0.089	0.228
Large (774M)	0.102	0.066	0.206
XL (1.5B)	0.089	0.065	0.112
Qwen2.5 Family			
Base 3B	0.083	0.084	0.114
Base 7B	0.066	0.067	0.094
Qwen3 Family			
Base 4B	0.079	0.057	0.053
Base 8B	0.064	**0.052**	**0.051**
Llama-3 Family			
Base 8B	**0.062**	0.057	0.052

Table 1: **Dimension-normalized transport Gram deviation (δ_{orth}) across model families.** Results cover nine checkpoints, excluding Llama-3.2-3B, across three depth bands. This dimension-dependent diagnostic does not establish scaled isometry of individual transports.

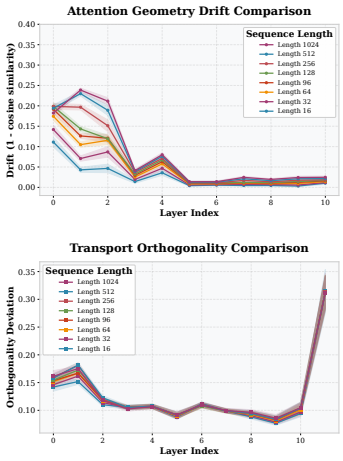


Figure 5: **Sensitivity to context length (GPT-2 Small).** We vary the input context length from 16 to 1,024 tokens. Longer contexts increase early-layer drift and change Gram deviation, while middle layers retain a similar low-drift and low-deviation profile in this tested range.

Figure 4 shows decaying singular values and concentrated cumulative energy for GPT-2 Small and Qwen3-1.7B. Their checkpoint median r_{99}/d values are 0.377 and 0.265, respectively. Most transport energy is therefore carried by a restricted fraction of feature directions in these selected-edge populations. The capacity theorem constrains the available directions; the spectra describe how their gains are distributed.

The retained singular values remain unequal. Even after restricting each edge to its r_{99} block, a fitted common gain leaves a substantial relative residual across all seven checkpoints (Appendix C). The combined picture is selective, anisotropic feature transmission: diagonal structure in averaged Gram matrices coexists with low effective energy rank and nonuniform gains on individual edges. It does not establish approximate scaled isometry or shared singular-vector directions across depth.

6.4 Dependence on context length

Figure 5 varies GPT-2 Small's context length from 16 to $1{,}024$ tokens. Longer contexts increase early-layer routing drift and change Gram deviation; middle layers retain lower drift and deviation, with a terminal rise in the latter. The separate spectral study uses context 64.

Appendices F and G extend the context comparisons; Appendix D.4 reports the BERT results.

Together, these diagnostics reveal structured source selection and feature transport with concentrated energy and unequal directional gains.

7 Conclusion

Multi-head attention admits an exact connection-walk representation: head-mean routing selects source tokens, and conditional mixtures of value–output maps transform their features. The representation yields a causal routing constraint, a value-group capacity bound, and conditions for a metric connection Laplacian. Empirical diagnostics reveal structured routing and diagonal averaged Gram patterns, while selected-edge spectra show concentrated transport energy with unequal directional gains. This framework separates source selection and feature transformation while retaining their coupled action on the input field, supporting geometric interpretation and operator approximation.

AI USE STATEMENT

We used generative AI assistants for writing and polishing, literature retrieval and discovery, and research ideation and execution. Assistance included revising prose and captions, finding and comparing related work, developing and checking mathematical arguments and empirical diagnostics, and implementing code for spectral analysis and visualization.

Verification included reviewing derivations against their stated assumptions, checking literature descriptions against primary sources, and comparing selected-edge spectral summaries and plots with numerical outputs. The authors take responsibility for the final text, mathematical claims, code, empirical results, and conclusions, including material developed with AI assistance.

ETHICS STATEMENT

This work studies the operator-level geometry of Transformer attention using publicly available pretrained model checkpoints and the public WikiText-2 dataset. The analyses use existing text data and involve no human-participant research. We do not identify societal consequences that warrant specific emphasis.

REPRODUCIBILITY STATEMENT

The connection-walk framework and exact attention representation are given in Sections 3–4. Their consequences and assumptions are stated in Section 5, with supporting derivations in Appendix B. Section 6 describes the empirical protocol and diagnostics. Appendix D contains complete layer-wise plots, Appendix E the transport statistics, and Appendix F the sequence-length comparisons. Appendix C details the selected-edge population, aggregation, and uncertainty estimates.

REFERENCES

Francesca Babiloni, Ioannis Marras, Gregory G. Slabaugh, and Stefanos Zafeiriou. TESA: tensor element self-attention via matricization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 13942–13951. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.01396. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Babiloni_TESA_Tensor_Element_Self-Attention_via_Matricization_CVPR_2020_paper.html.

Afonso S. Bandeira, Amit Singer, and Daniel A. Spielman. A cheeger inequality for the graph connection laplacian. *SIAM J. Matrix Anal. Appl.*, 34(4):1611–1630, 2013. doi: 10.1137/120875338. URL https://doi.org/10.1137/120875338.

Federico Barbero, Cristian Bodnar, Haitz Sáez de Ocáriz Borde, Michael M. Bronstein, Petar Velickovic, and Pietro Liò. Sheaf neural networks with connection laplacians. In Alexander Cloninger, Timothy Doster, Tegan Emerson, Manohar Kaul, Ira Ktena, Henry Kvinge, Nina Miolane, Bastian Rice, Sarah Tymochko, and Guy Wolf (eds.), *Topological, Algebraic and Geometric Learning Workshops 2022, 25-22 July 2022, Virtual*, volume 196 of *Proceedings of Machine Learning Research*, pp. 28–36. PMLR, 2022. URL https://proceedings.mlr.press/v196/barbero22a.html.

Cristian Bodnar, Francesco Di Giovanni, Benjamin Paul Chamberlain, Pietro Lió, and Michael M. Bronstein. Neural sheaf diffusion: A topological perspective on heterophily and oversmoothing in gnns. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/75c45fca2aa416ada062b26cc4fb7641-Abstract-Conference.html.

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Velickovic. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *CoRR*, abs/2104.13478, 2021. URL https://arxiv.org/abs/2104.13478.

Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamás Sarlós, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, David Benjamin Belanger, Lucy J. Colwell, and Adrian Weller. Rethinking attention with performers. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=Ua6zuk0WRH.

Sankalan Pal Chowdhury, Adamos Solomou, Avinava Dubey, and Mrinmaya Sachan. On learning the transformer kernel. *CoRR*, abs/2110.08323, 2021. URL https://arxiv.org/abs/2110.08323.

Taco Cohen, Maurice Weiler, Berkay Kicanaoglu, and Max Welling. Gauge equivariant convolutional networks and the icosahedral CNN. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2019. URL http://proceedings.mlr.press/v97/cohen19d.html.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/V1/N19-1423. URL https://doi.org/10.18653/v1/n19-1423.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Ahmed Farooq. A framework for non-linear attention via modern hopfield networks. *CoRR*, abs/2506.11043, 2025. doi: 10.48550/ARXIV.2506.11043. URL https://doi.org/10.48550/arXiv.2506.11043.

Kelvin Kan, Xingjian Li, and Stanley J. Osher. Ot-transformer: A continuous-time transformer architecture with optimal transport regularization. *CoRR*, abs/2501.18793, 2025. doi: 10.48550/ARXIV.2501.18793. URL https://doi.org/10.48550/arXiv.2501.18793.

Devin Kreuzer, Dominique Beaini, William L. Hamilton, Vincent Létourneau, and Prudencio Tossou. Rethinking graph transformers with spectral attention. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 21618–21629, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/b4fd1d2cb085390fbbadae65e07876a7-Abstract.html.

Xindian Ma, Peng Zhang, Shuai Zhang, Nan Duan, Yuexian Hou, Ming Zhou, and Dawei Song. A tensorized transformer for language modeling. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 2229–2239, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/dc960c46c38bd16e953d97cdeefdbc68-Abstract.html.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Byj72udxe.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners, 2019.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Lukas Gruber, Markus Holzleitner, Thomas Adler, David P. Kreil, Michael K. Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=tL89RnzIiCd.

Amit Singer and H-T Wu. Vector diffusion maps and the connection laplacian. *Communications on pure and applied mathematics*, 65(8):1067–1144, 2012. URL https://arxiv.org/abs/1102.0075.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey. *ACM Comput. Surv.*, 55(6):109:1–109:28, 2023. doi: 10.1145/3530811. URL https://doi.org/10.1145/3530811.

Llama Team. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024. doi: 10.48550/ARXIV.2407.21783. URL https://doi.org/10.48550/arXiv.2407.21783.

Qwen Team. Qwen3 technical report. *CoRR*, abs/2505.09388, 2025. doi: 10.48550/ARXIV.2505.09388. URL https://doi.org/10.48550/arXiv.2505.09388.

Anh Tong, Thanh Nguyen-Tang, Dongeun Lee, Duc Nguyen, Toan M. Tran, David Leo Wright Hall, Cheongwoong Kang, and Jaesik Choi. Neural ODE transformers: Analyzing internal dynamics and adaptive fine-tuning. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=XnDyddPcBT.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Matthew A. Wright and Joseph E. Gonzalez. Transformers are deep infinite-dimensional non-mercer binary kernel machines. *CoRR*, abs/2106.01506, 2021. URL https://arxiv.org/abs/2106.01506.

Xinyi Wu, Amir Ajorlou, Yifei Wang, Stefanie Jegelka, and Ali Jadbabaie. On the role of attention masks and layernorm in transformers. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/1ac3030fc57850b0fb11dfe9d4880ad7-Abstract-Conference.html.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *CoRR*, abs/2412.15115, 2024. doi: 10.48550/ARXIV.2412.15115. URL https://doi.org/10.48550/arXiv.2412.15115.

Seongjun Yun, Minbyul Jeong, Raehyun Kim, Jaewoo Kang, and Hyunwoo J. Kim. Graph transformer networks. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp.

11960–11970, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/9d63484abb477c97640154d40595a3bb-Abstract.html.

Yifan Zhang, Yifeng Liu, Huizhuo Yuan, Zhen Qin, Yang Yuan, Quanquan Gu, and Andrew Chi-Chih Yao. Tensor product attention is all you need. *CoRR*, abs/2501.06425, 2025. doi: 10.48550/ARXIV.2501.06425. URL https://doi.org/10.48550/arXiv.2501.06425.

A Discussion

The connection-walk representation puts token selection, feature transformation, and their coupled action in one operator. This geometric description remains exact when routing is directed or feature maps are singular: the edge maps need not preserve a metric. It relates the factors, while separate diagnostics need not determine their joint action on an input field.

The exact identity concerns a realized evaluation-time attention module with standard linear value and output pathways. Its constructing input fixes the head weights, producing a linear field operator rather than the attention Jacobian. Row-stochastic routing assumes disabled attention dropout and at least one admissible source per row. The native multi-head action also retains its head-count factor and affine term; properties of the normalized connection walk do not automatically transfer to the complete attention or residual update.

Writing $L = I - T$ does not by itself make attention a metric diffusion. Proposition 5 gives sufficient conditions for contraction and for a self-adjoint positive-semidefinite connection Laplacian. Generic directed or singular attention maps need not meet those conditions. The exact representation also makes no prediction of a depthwise limit or of improved prediction from imposing metric properties.

The empirical results describe the sampled text, fixed checkpoints, and distinct edge populations stated in Section 6. Aggregating edges or feature coordinates can conceal heterogeneity, and routing or spectral statistics alone cannot establish similarity of native actions. The framework therefore motivates comparisons of routing, transport, and matched-input action when assessing changes across layers or approximations. Claims about other architectures, modalities, or downstream performance require their own measurements and validation.

B Representation Details and Proofs

B.1 Vectorization, head maps, and affine terms

For row features $V_i \in \mathbb{R}^{1\times d}$, row-consistent stacking is

$$\mathrm{vec}^r(V) = \mathrm{vec}(V^\top) = [V_0^\top; \ldots; V_{n-1}^\top] \in \mathbb{R}^{nd}. \tag{15}$$

The standard identity $\mathrm{vec}(UBC) = (C^\top \otimes U)\,\mathrm{vec}(B)$ gives

$$\mathrm{vec}^r(AVM) = (A \otimes M^\top)\mathrm{vec}^r(V), \tag{16}$$

since $(AVM)^\top = M^\top V^\top A^\top$.

For the row-action notation used in the empirical displays, $A_{\mathrm{eff}} = A$ and $O_{ij}^{\mathrm{eff}} = O_{ij}^\top$. Consequently, $O_{ij}^{\mathrm{eff}\,\top} O_{ij}^{\mathrm{eff}} = O_{ij} O_{ij}^\top$ acts in output feature coordinates, while $O_{ij}^\top O_{ij}$ acts in input feature coordinates. The two square Gram matrices have the same eigenvalues, but their coordinate patterns differ. The transpose conversion does not change the source/destination indices or the connection action in Eq. (4). For standard dot-product attention, query and key projections construct $A^{(h)} = \mathrm{softmax}(Q_h K_h^\top/\tau_h + \mathcal{M}_h)$ row-wise. Positional transforms of queries and keys can change these weights without changing the head-sum argument. The result assumes the value/output pathway has the linear form specified in Eq. (1); extra edge-dependent value operations require their own effective maps.

With affine values, a head contributes

$$A^{(h)}(VW_{V,h} + \mathbf{1}b_{V,h})W_{O,h} = A^{(h)}VM_h + \mathbf{1}b_{V,h}W_{O,h}.$$

Adding output bias yields the broadcast row $b = \sum_h b_{V,h}W_{O,h} + b_O$. Because $A_{ij} = 0$ implies all nonnegative head weights at that edge vanish, inactive-edge completion never changes the affine-free operator. Choosing the zero map makes the capacity bound apply to all edges as well. The native head sum is HT, whereas all normalized-connection bounds concern T. Contraction of T is not a contraction certificate for HT or for a residual block containing it.

B.2 TRANSPORTED PATHS AND THE CONNECTION SEMIGROUP

For an integer $k \geq 1$, the block $[T^k]_{ij}$ sums the transformations along all length-k walks from source j to destination i. With $i_0 = j$ and $i_k = i$,

$$[T^k]_{ij} = \sum_{i_1,\ldots,i_{k-1}} \left(\prod_{r=1}^{k} A_{i_r i_{r-1}} \right) O_{i_k i_{k-1}} \cdots O_{i_1 i_0}. \tag{17}$$

Block multiplication proves this identity by induction. The scalar factors weight each path, while the ordered product composes feature transformations. Even when path weights agree, different map orientations can produce different aggregate actions. The powers here repeat one fixed connection; Transformer layers generally induce different input-dependent connections.

For a fixed connection, the continuous-time propagator satisfies

$$e^{-tL} = e^{-t} \sum_{k=0}^{\infty} \frac{t^k}{k!} T^k, \qquad t \geq 0. \tag{18}$$

This expresses the propagator as a sum over transported walks. The algebraic identity requires no reversibility. Section 5.4 gives sufficient conditions for contraction and a metric diffusion interpretation. No continuous-depth limit of a trained Transformer is assumed.

B.3 PROOF OF THE CAUSAL ROUTING CONSTRAINT

The norm bounds in Proposition 2 follow from row stochasticity as stated in the main text. To prove the causal assertion, suppose $\|A\|_2 \leq 1$ and let $q = \mathbf{1}/\sqrt{n}$. Then $Aq = q$ and $B = I - A^\top A$ is positive semidefinite, with $q^\top B q = 0$. Thus $Bq = 0$, so $A^\top q = A^\top A q = q$: the column sums are also one. For a nonnegative lower-triangular matrix, row zero now forces $A_{00} = 1$; column zero then forces $A_{i0} = 0$ for $i > 0$. Removing row and column zero leaves the same problem in dimension $n - 1$. Induction gives $A = I$. The converse is immediate. This establishes a worst-case statement; it does not supply a lower gain bound on every field or on the coupled operator.

B.4 PROOF OF THE VALUE-GROUP CAPACITY BOUND

For an active edge, collect the shared value maps and mixed output slices:

$$\mathcal{V} = [W_{V,1} \ \cdots \ W_{V,G}] \in \mathbb{R}^{d \times G d_v}, \qquad \mathcal{B}_{ij} = \begin{bmatrix} B_{1,ij} \\ \vdots \\ B_{G,ij} \end{bmatrix} \in \mathbb{R}^{G d_v \times d}.$$

Equation (7) gives $O_{ij}^\top = \mathcal{V}\mathcal{B}_{ij}$, so its rank is at most $\min(d, G d_v)$. The zero completion satisfies the same bound. With unequal group widths the bound becomes $\min(d, \sum_g d_{v,g})$. A normalized ceiling G/H additionally assumes $d = H d_v$; it is not a general identity for arbitrary architectures.

The bound describes exact algebraic capacity. Numerical rank and energy rank are different summaries: numerical rank uses singular values above a declared relative threshold, while r_{99} is the smallest rank retaining 99% of squared singular-value mass. Neither rank is fixed by this bound.

B.5 PROOF OF THE JOINT GAIN BOUND AND APPROXIMATION CONSEQUENCE

Write a stacked field as $v = (v_0, \ldots, v_{n-1})$ and set $u_j = \|v_j\|_2$. The block triangle inequality gives

$$\|(Tv)_i\|_2 \leq \sum_j A_{ij} \|O_{ij}\|_2 \|v_j\|_2 = [(A \odot s)u]_i.$$

Since the right side is nonnegative, summing squares yields $\|Tv\|_2 \leq \|(A \odot s)u\|_2 \leq \|A \odot s\|_2 \|u\|_2$, and $\|u\|_2 = \|v\|_2$. For a common edge map O_0, the singular values of $A \otimes O_0$ are pairwise products, giving the equality in Proposition 4.

More generally, let $E_{ij} = A_{ij}O_{ij} - \widetilde{A}_{ij}\widetilde{O}_{ij}$ be the block error of an approximation, and $e_{ij} = \|E_{ij}\|_2$. The same proof gives $\|T - \widetilde{T}\|_2 \le \|e\|_2$. At fixed routing, put $B_{ij} = e_{ij} = A_{ij}\|O_{ij} - \widetilde{O}_{ij}\|_2$. For any row field V, preserving the native affine term then gives

$$H\|\mathcal{T}(A, O)V - \mathcal{T}(A, \widetilde{O})V\|_F \le H\|B\|_2\|V\|_F. \tag{19}$$

If edges are simply discarded without renormalizing the retained weights, $e_{ij} = A_{ij}\|O_{ij}\|_2$ on discarded edges and zero otherwise. These are frozen-operator error bounds; a prediction-loss guarantee requires additional information about the downstream model and data.

B.6 Polar decomposition and active anisotropy

For a compact singular value decomposition $O = U_r\Sigma_r V_r^\top$, write $O = QP$ with $P = V_r\Sigma_r V_r^\top \succeq 0$ and $Q = U_r V_r^\top$. The canonical Q is an isometry on the support of P and zero on its orthogonal complement. For invertible square O, Q is orthogonal. This separates directional alignment from positive feature scaling at one edge. It does not by itself make aggregation a diffusion, ensure reciprocal edge maps, or make the whole operator contractive.

For a fixed active rank r, the best scaled partial isometry on the singular subspaces has scalar gain $\bar{\sigma}_r = r^{-1}\sum_{k=1}^r \sigma_k$. Its relative Frobenius defect on the active block is

$$\delta_r = \frac{\left(\sum_{k=1}^r (\sigma_k - \bar{\sigma}_r)^2\right)^{1/2}}{\left(\sum_{k=1}^r \sigma_k^2\right)^{1/2}}.$$

It vanishes exactly when the retained gains are equal. Effective-rank reduction therefore does not imply small δ_r. Full-width Gram errors normalized by ambient dimension can dilute the defect of a low-rank map; the active-block comparison avoids treating that dilution as evidence of isometry.

B.7 Proof of the conditional metric regime

For stationary positive π and contractive active maps, Jensen's inequality gives

$$\begin{aligned}\|Tv\|_\pi^2 &\le \sum_{i,j} \pi_i A_{ij}\|O_{ij}v_j\|_2^2 \\ &\le \sum_j \Big(\sum_i \pi_i A_{ij}\Big)\|v_j\|_2^2 = \|v\|_\pi^2.\end{aligned}$$

This part needs neither reversibility nor reciprocal transports. The weighted adjoint has blocks $T_{ij}^{*\pi} = (\pi_j/\pi_i)A_{ji}O_{ji}^\top$. Detailed balance and reciprocal maps give $T^{*\pi} = T$, including diagonal blocks when reciprocity is imposed on active self loops.

With orthogonal maps, expansion of the edge discrepancies gives

$$\begin{aligned}\frac{1}{2}\sum_{i,j}\pi_i A_{ij}\|v_i - O_{ij}v_j\|_2^2 &= \frac{1}{2}\|v\|_\pi^2 + \frac{1}{2}\|v\|_\pi^2 - \sum_{i,j}\pi_i A_{ij}v_i^\top O_{ij}v_j \\ &= \langle v, (I - T)v\rangle_\pi.\end{aligned}$$

Row sums supply the first squared-norm term and stationarity supplies the second. The identity itself uses orthogonality and stationarity; the stronger reversible, reciprocal assumptions additionally supply self-adjointness. Hence L is PSD under the stated conditions, and its kernel consists of parallel fields on all active edges. This kernel can be trivial if edge transports are inconsistent. Self-adjointness supplies a weighted orthonormal eigenbasis with nonnegative Laplacian eigenvalues λ_a. In this basis, e^{-tL} multiplies each coefficient by $e^{-t\lambda_a} \le 1$ for $t \ge 0$, proving contraction. Outside this regime the sum of squared discrepancies is still nonnegative, but need not equal the quadratic form of $I - T$.

C Transport Spectra and Effective Capacity

Checkpoints and edge selection. The fp64 study covers GPT-2 Small, Medium, Large, and XL; Qwen3-1.7B and 4B; and Llama-3.2-3B. It uses WikiText-2 validation windows at context 64,

excluding position-zero sources and destinations. The protocol considers 64 windows for the threshold sweep and the first 32 for spectra. We form a shared shortlist of up to eight edge positions across spectral windows, favoring positions that repeatedly enter the top-ten candidates across adjacent layers. Candidates are scored by the smaller routing weight at the two layers. A shortlisted position contributes only when it is among the candidates for that window and layer pair. Summaries use only the left-endpoint spectrum of each eligible pair. The retained factors, products, and singular values are computed in float64, with reconstruction residuals below 10^{-6} in every checkpoint.

Active-subspace geometry. Energy concentration does not establish uniform gains within the retained subspace. On each edge's fixed r_{99} block, define the relative distance to the nearest scalar singular-value profile as

$$\eta_{99} = \min_{s \geq 0} \frac{\|\Sigma_{r_{99}} - sI_{r_{99}}\|_F}{\|\Sigma_{r_{99}}\|_F}, \tag{20}$$

where $\Sigma_r = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$. Zero means equal gains on the fixed block. Checkpoint medians range from 0.514 to 0.628, while median r_{99}/d ranges from 0.217 to 0.427. Thus restricting attention to the energy-carrying subspace does not remove the observed anisotropy.

Aggregation and interpretation. For each edge we compute the energy-rank fraction r_{99}/d defined in Section 6.3 and the active defect η_{99} in Eq. (20). We take an edge median within each window–layer, then a layer median within each window, then the median across contributing windows. Table 2 reports these checkpoint medians and percentile intervals from 10,000 window bootstrap draws. Windows are the resampling unit; edges and layers are not treated as independent replicates. GPT-2 Medium contributes 13 eligible windows and requires particular caution when comparing its interval with other checkpoints.

Checkpoint	Windows	r_{99}/d	Active defect η_{99}
GPT-2 Small	32	0.377 [0.367, 0.391]	0.560 [0.548, 0.569]
GPT-2 Medium	13	0.349 [0.309, 0.492]	0.587 [0.566, 0.614]
GPT-2 Large	32	0.427 [0.383, 0.462]	0.590 [0.578, 0.600]
GPT-2 XL	32	0.409 [0.388, 0.437]	0.628 [0.620, 0.644]
Qwen3-1.7B	32	0.265 [0.260, 0.273]	0.533 [0.530, 0.535]
Llama-3.2-3B	32	0.217 [0.215, 0.222]	0.514 [0.510, 0.523]
Qwen3-4B	32	0.245 [0.237, 0.251]	0.520 [0.513, 0.525]

Table 2: **Restricted effective capacity with anisotropic active transport.** Checkpoint medians and 95% window-bootstrap intervals in the selected-edge fp64 population. Zero active defect means equal singular values on the fixed r_{99} block. These statistics use a separate population and normalization from the Gram diagnostic in Table 1.

For Figure 4, normalize each edge's singular values by its largest singular value and its cumulative squared energy by total squared energy. Apply the same edge–layer–window median hierarchy pointwise at each index k/d. Bands are the interquartile range of the window-level curves; they describe variation, not uncertainty on a mean. A crossing of the aggregated energy curve is not substituted for the median of edgewise r_{99}. The left plot truncates its visible ordinate at 10^{-4} for readability; the energy curves and all statistics retain the complete spectra.

The singular values are invariant to the row/column transpose convention. Normalization removes an overall scalar gain, so the plots describe spectral shape rather than absolute operator amplification. A diagonal coordinate Gram matrix need not have equal diagonal entries; a diagonal averaged Gram matrix also need not describe the geometry of each contributing edge. The active-block defect directly tests unequal singular values after fixing the energy rank. This distinguishes restricted capacity from a scaled partial isometry and does not assume that singular-vector directions persist across layers.

D Complete Evolution Plots

We provide a detailed layerwise breakdown for each model in two side-by-side plots:

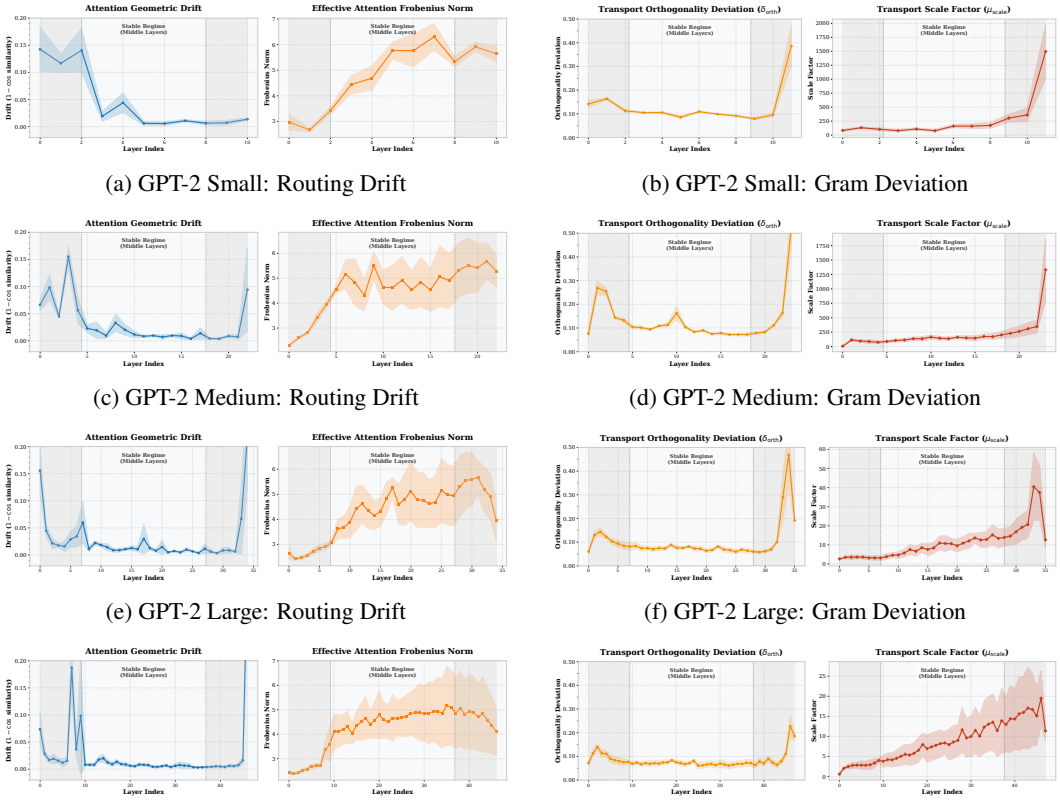
- **Left:** Adjacent Routing Drift (d_A) and Frobenius Norm ($\|A_{\text{eff}}\|_F$).

- **Right:** Transport Gram Deviation (δ_{orth}) and Mean Squared Gain (μ_{scale}).

This layout compares adjacent routing similarity and source concentration with transport Gram deviation and mean squared gain. The retained plots label δ_{orth} as orthogonality deviation; it is the dimension-normalized Gram diagnostic defined in Section 6.3, not a certificate of scaled isometry.

D.1 GPT-2 Series (from 124M to 1.5B)

The GPT-2 family (Radford et al., 2019) (Figure 6) compares checkpoints of different sizes within one model family. **Observation:** GPT-2 Small (Row 1) shows greater layerwise variation, while GPT-2 XL (Row 4) has a smoother profile and lower reported Gram deviation. Table 1 gives the depth-band averages. This is a comparison of trained checkpoints with different feature dimensions, not a controlled scaling experiment or evidence of convergence to isometry.



(a) GPT-2 Small: Routing Drift

(b) GPT-2 Small: Gram Deviation

(c) GPT-2 Medium: Routing Drift

(d) GPT-2 Medium: Gram Deviation

(e) GPT-2 Large: Routing Drift

(f) GPT-2 Large: Gram Deviation

(g) GPT-2 XL: Routing Drift

(h) GPT-2 XL: Gram Deviation

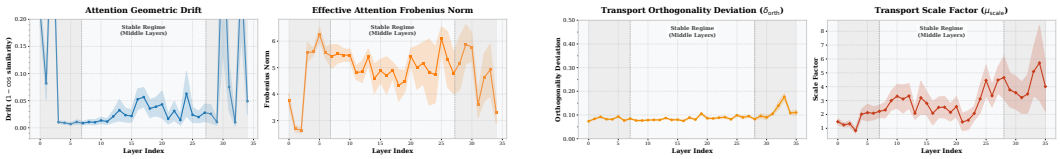
Figure 6: **GPT-2 Series Evolution.** Routing drift and concentration (left) and Gram deviation and mean squared gain (right) across four model sizes. The larger variants show smoother profiles and lower reported Gram deviation; the diagnostic normalization depends on feature dimension.

D.2 Qwen Series (2.5 & 3)

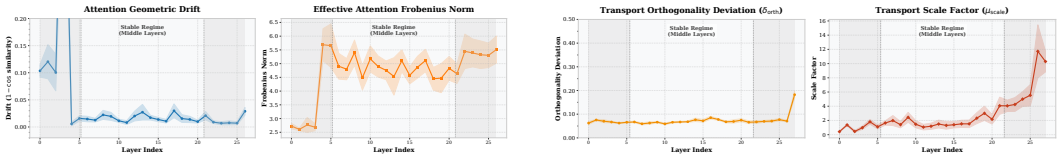
Figure 7 shows the results for the Qwen2.5 (Yang et al., 2024) and Qwen3 (Team, 2025) families. **Observation:** The displayed profiles include low-drift middle-layer regimes. Qwen3-8B has low reported Gram deviation; this coordinate diagnostic alone does not establish uniform transport gains.

D.3 Llama Series (3.2 & 3)

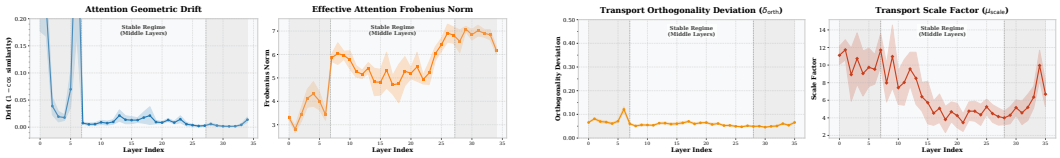
Figure 8 displays results for Llama-3.2-3B (Team, 2024) and Llama-3-8B (Team, 2024). **Observation:** Llama-3-8B has among the lowest reported Gram deviations in our evaluation, with δ_{orth} consistently



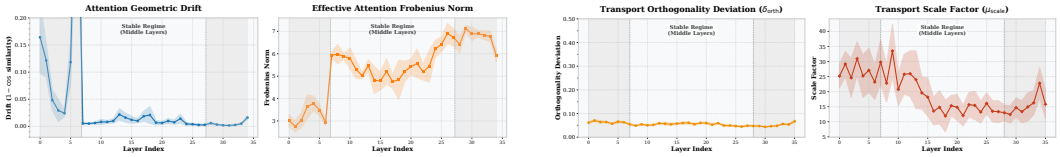
(a) Qwen2.5-3B: Routing Drift (b) Qwen2.5-3B: Gram Deviation



(c) Qwen2.5-7B: Routing Drift (d) Qwen2.5-7B: Gram Deviation



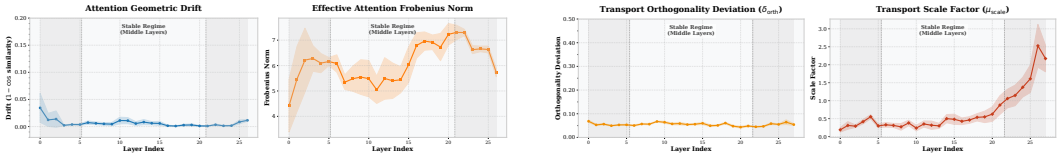
(e) Qwen3-4B: Routing Drift (f) Qwen3-4B: Gram Deviation



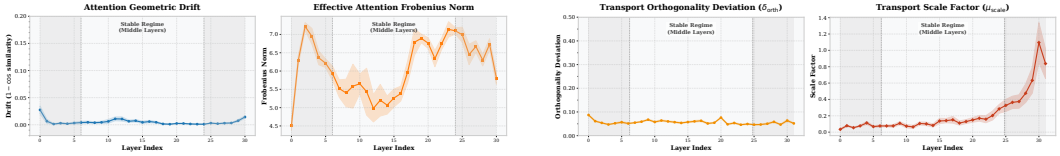
(g) Qwen3-8B: Routing Drift (h) Qwen3-8B: Gram Deviation

Figure 7: **Qwen Series Evolution.** Qwen2.5 and Qwen3 show low middle-layer routing drift.

below 0.06 for the majority of layers. Middle layers show low adjacent routing drift (left) and low Gram deviation (right).



(a) Llama-3.2-3B: Routing Drift (b) Llama-3.2-3B: Gram Deviation



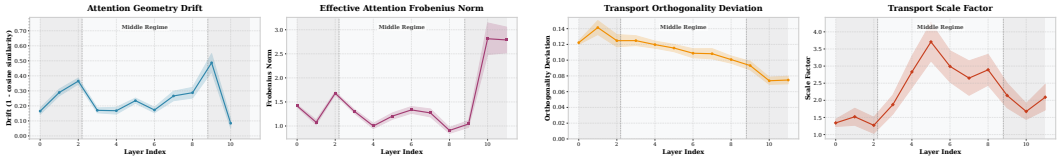
(c) Llama-3-8B: Routing Drift (d) Llama-3-8B: Gram Deviation

Figure 8: **Llama Series Evolution.** Llama models show low drift and low transport deviation under the proposed diagnostics.

D.4 Encoder-Only BERT

Figure 9 reports the same operator diagnostics for BERT-base-uncased (Devlin et al., 2019). Unless otherwise stated, the setup uses WikiText-2 test samples and a length-64 truncation. For BERT, we remove `[CLS]`, `[SEP]`, and `[PAD]` positions before computing token-graph diagnostics, and we

use the full valid bidirectional mask rather than a causal lower-triangular support. **Observation:** BERT shows high adjacent-layer routing similarity overall, with a different depth profile from the decoder-only GPT-2 probe. Its reported Gram deviation decreases in later layers. These observations extend the routing and coordinate-Gram diagnostics to this bidirectional checkpoint; the separate selected-edge spectral study does not include BERT.



(a) Effective-walk drift and $\|A_{\text{eff}}\|_F$.

(b) Transport Gram deviation and mean squared gain.

Figure 9: BERT-base-uncased diagnostics with the full valid bidirectional non-special-token mask.

E STATISTICAL ANALYSIS OF TRANSPORT MATRICES

E.1 DIAGONAL GRAM STRUCTURE

We quantify coordinate-level diagonal structure by examining the local Gram matrix $G_{ij}^{\ell} = O_{ij}^{\mathrm{eff},\ell\top} O_{ij}^{\mathrm{eff},\ell}$. We report the mean absolute magnitude of diagonal versus off-diagonal elements across three depth regimes: INITIAL (0–20%), MIDDLE (20–80%), and FINAL (80–100%).

As detailed in Table 3, the mean diagonal magnitudes exceed the mean off-diagonal magnitudes across the displayed model families. Focusing on the "Middle" depth band:

- For **GPT-2 Small**, the diagonal mean (133.8) is about **15x** the off-diagonal mean (9.0).
- For **Llama-3-8B**, the diagonal mean (0.1412) is over **20x** the off-diagonal mean (0.0068).

The Qwen series and larger GPT-2 variants also have middle-band diagonal-to-off-diagonal mean-magnitude ratios above $10\times$, as shown in Table 3. These are ratios of coordinate averages, not the row-sum condition defining strict matrix diagonal dominance.

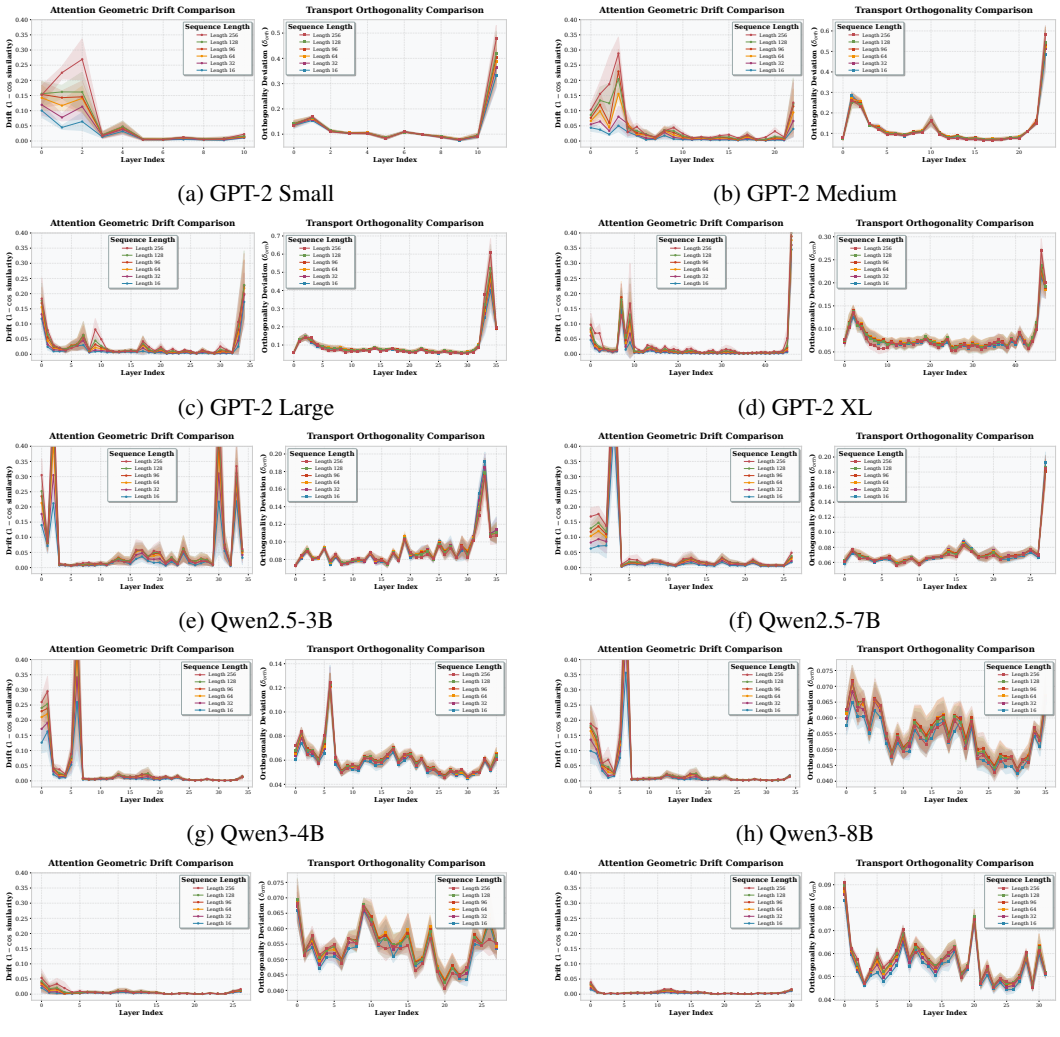
These coordinate-level statistics corroborate the diagonal structure in Figure 3. They do not establish proximity to a rotation or scaled isometry: diagonal gains may differ, and an aggregate Gram display need not represent any individual edge. Appendix C measures the effective rank and anisotropy of individual selected-edge spectra separately.

Table 3: Quantitative analysis of Transport Gram Matrix G_{ij}. We report the mean magnitude of diagonal vs. off-diagonal elements. The "Middle" regime (layers from 20% to 80% depth) has higher diagonal than off-diagonal means; this coordinate statistic does not establish equal singular values.

	INITIAL (20%)		MIDDLE (60%)		FINAL (20%)	
MODEL	DIAG	OFF-DIAG	DIAG	OFF-DIAG	DIAG	OFF-DIAG
GPT-2 (124M)	131.98	6.49	133.81	9.00	699.14	19.10
GPT-2 MEDIUM (355M)	81.15	4.20	155.00	10.64	496.23	15.47
GPT-2 LARGE (774M)	3.00	0.25	8.71	0.58	18.54	0.89
GPT-2 XL (1.5B)	2.21	0.18	5.93	0.40	12.35	1.16
QWEN2.5-3B	1.59	0.10	2.53	0.16	3.66	0.24
QWEN2.5-7B	1.10	0.06	1.76	0.09	7.51	0.46
QWEN3-4B	10.40	0.61	5.67	0.29	6.57	0.35
QWEN3-8B	26.73	1.45	16.23	0.76	14.97	0.69
LLAMA-3.2-3B	0.34	0.01	0.40	0.02	1.69	0.08
LLAMA-3-8B	0.08	0.004	**0.14**	**0.007**	0.61	0.03

F Ablation Study: Sequence Length

We examine how routing drift and transport Gram deviation vary with context length. The main text extends GPT-2 Small to 1,024 tokens in Figure 5; Figure 10 reports the broader model-family comparison for context lengths $L \in \{16, 32, 64, 96, 128, 256\}$.



(a) GPT-2 Small (b) GPT-2 Medium

(c) GPT-2 Large (d) GPT-2 XL

(e) Qwen2.5-3B (f) Qwen2.5-7B

(g) Qwen3-4B (h) Qwen3-8B

(i) Llama-3.2-3B (j) Llama-3-8B

Figure 10: **Impact of Sequence Length.** Comparing routing drift and Gram deviation across model families for context lengths $L = 16$ to $L = 256$. Across the tested architectures, deeper layers often recover lower drift despite increased initial drift for longer sequences.

G Extended Context-Length Diagnostic

The main-text ablation (Figure 5) probes context lengths up to 1,024 tokens for GPT-2 Small. Figure 11 shows the same curves at a larger size. The broad depth profiles persist across this range: longer contexts increase early-layer routing drift and change Gram deviation, while middle layers retain a similar low-drift, low-deviation profile. These are descriptive observations on a single model and probe, not a causal claim about long-context competence.

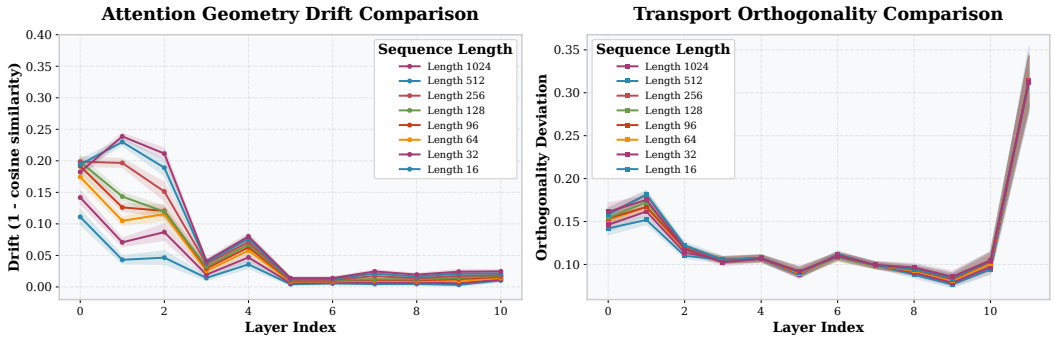


Figure 11: Extended context-length diagnostic for GPT-2 Small, probing context lengths up to 1,024 tokens. Longer contexts increase early-layer routing drift and change Gram deviation; the middle-layer low-drift, low-deviation profile remains visible in the tested range.

H ALGEBRAIC EXTENSIONS

H.1 CROSS-ATTENTION

For n_s source tokens and n_t target tokens, take $A^{(h)} \in \mathbb{R}^{n_t \times n_s}$ row-stochastic over sources. The same conditional head mixture gives rectangular blocks $T_{ij} = A_{ij}O_{ij}$ and the exact scaled head-sum representation. Source and target feature widths may also differ, with rectangular O_{ij}. The joint norm proof still applies. The square-walk spectral-radius statement and $I - T$ Laplacian require a common state space and do not transfer directly to this rectangular case. No cross-attention empirical claim is made here.

H.2 TENSOR-VALUED FEATURES

A finite-dimensional tensor space can replace $\mathbb{R}^d$ by a chosen vectorization. For matrix-valued features $V_j \in \mathbb{R}^{d_1 \times d_2}$, a separable edge action $V_j \mapsto B_{ij}V_jC_{ij}^\top$ becomes the linear map $C_{ij} \otimes B_{ij}$ under column vectorization. Routing and this feature map then form the same connection construction. This is an algebraic extension; spatial frame alignment, geometric locality, or diffusion on a grid requires additional modeling assumptions and empirical validation.