DIFFUSION–ATTENTION CONNECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

We show that the diffusion map affinity matrix is the twisted Hadamard product of the self-attention matrix. Concretely, let the generalized feature-similarity matrix be $\mathcal{W} = M + A$ with $M = M^\dagger$ a Hermitian (real part, encoding geometry) and $A = -A^\dagger$ skew-Hermitian (imaginary part, encoding directionality). Softmax applied to the real logits from M yields a first-order, row/column-stochastic attention operator. The diffusion kernel then arises as the twisted Hadamard product (a Product-of-Experts identity), producing a symmetric second-order affinity whose spectrum matches diffusion maps. The skew part A contributes only phases; placing them outside the softmax yields a $U(1)$ gauge-equivariant "magnetic" variant without breaking stochasticity.

1 INTRODUCTION

Laplacian-eigenmap and diffusion-maps methods Belkin & Niyogi (2003); Coifman & Lafon (2006) and the attention mechanism Bahdanau et al. (2014); Vaswani et al. (2017) are two of the most widely used kernels in modern machine learning. Diffusion maps construct a Markov operator on samples and recover geometry via the spectrum of a (possibly normalized) graph Laplacian, enabling robust manifold learning and diffusion-based embeddings. Attention, by contrast, builds a row-stochastic kernel from pairwise scores via a softmax and powers today's Transformer architectures. Although typically applied in different domains, the two formalisms share a common backbone: (i) a *nonlinear kernelization* of pairwise scores (e.g., Gaussian or exponential), followed by (ii) a *row-wise normalization* that yields a random-walk interpretation. In statistical physics, this normalization corresponds to Gibbs–Boltzmann weights; in machine learning, it appears as the softmax. These parallels suggest a tighter connection than is usually made explicit. This connection we explore was motivated by recent linearization schemes in both theories in diffusion-maps Candanedo (2025) and in attention Tsai et al. (2019); Nauen et al. (2025).

We show that attention and diffusion maps are two faces of the same construction. Our starting point is a bilinear dissimilarity between samples, built from a feature→feature operator M estimated in a Koopman/EDMD fashion. From this formulation we obtain two directed *distances* $d^{\pm}_{ij}$ whose softmax normalizations coincide with the usual attention matrices (query→key and key→query). We then demonstrate that the symmetric (Hermitian) part of M governs the diffusion-map distance, while its skew part contributes only a complex phase that can be interpreted as a $U(1)$ connection. This leads naturally to the construction of an equivariant first-order graph operator based on attention, whose adjoint product recovers the magnetic Laplace–Beltrami operator used in diffusion maps. Finally, we provide a gauge-consistent magnetic variant by placing the complex phases outside the normalization, which preserves local $U(1)^N$ invariance.

Our contributions are as follows. Concretely, we (i) unify diffusion maps and attention through a common bilinear kernel; (ii) identify attention as a *first-order* equivariant difference whose composition yields the *second-order* diffusion operator; (iii) provide a principled recipe to learn M from data (Koopman-in-feature-space), separating geometry (symmetric part) from directionality (skew part); and (iv) introduce a simple magnetic/connection form compatible with both diffusion and attention. Notation-wise we use Einstein index and summation notation throughout.

2 SIMILARITY

To begin our discussion we wish to determine the similarity between two data samples. Samples in our dataset are characterized by a vector of numbers in some high dimensional space $v_X, w_X \in \mathbb{C}^D$. Over N samples, we obtain a high-dimensional point-cloud represented by the following *dataset* matrix $R_{iX} \in \mathbb{C}^{N\times D}$, with indices $i, j \in \{1, 2, \cdots, N\}$ and $X, Y \in \{1, 2, \cdots, D\}$. Our objective is to find similarities between our samples. This can be achieved by the generalized-distance matrix with a Bilinear Dissimilarity (BD) matrix (potentially asymmetric and complex), $\mathcal{W}_{XY} \in \mathbb{C}^{D\times D}$:

$$\mathcal{D}^2_{ij} = \mathbf{1}_i\mathcal{R}^2_j - R_{iX}\mathcal{W}_{XY}R^\dagger_{Yj} + \mathcal{R}^2_i\mathbf{1}_j - R_{jX}\mathcal{W}_{XY}R^\dagger_{Xi} \quad , \tag{1}$$

with $\mathcal{R}^2_i = R_{iX}\mathcal{W}_{XY}R^\dagger_{Yi} = (R_{iY}\mathcal{W}^\dagger_{YX}R^\dagger_{Xi})^* \in \mathbb{C}^N$, and $\mathbf{1}_i \in \mathbb{R}^N$ is a vector of entirely ones. In general $\mathcal{D}^2_{ij} \in \mathbb{C}^{N\times N}$, however it may be partitioned into real and imaginary parts by the Hermitian/Anti-Hermitian partition of the BD matrix: the Hermitian Mahalanobis bilinear form $M_{XY} = \frac{1}{2}\left(\mathcal{W}_{XY} + (\mathcal{W}^\dagger)_{XY}\right) \in \mathbb{C}^{D\times D}$ and the anti-Hermitian connection $A_{XY} = \frac{1}{2}\left(\mathcal{W}_{XY} - (\mathcal{W}^\dagger)_{XY}\right) \in \mathbb{C}^{D\times D}$, such that $\mathcal{W}_{XY} = M_{XY} + A_{XY}$:

$$\begin{aligned}\mathcal{D}^2_{ij} &= \Re\mathcal{D}^2_{ij} + i\Im\mathcal{D}^2_{ij}\\ &= D^2_{ij} + iB_{ij}\end{aligned}$$

The real-part obtained from the Hermitian term is:

$$D^2_{ij} = \underbrace{\mathbf{1}_iR^2_j - R_{iX}M_{XY}R^\dagger_{Yj}} + \underbrace{R^2_i\mathbf{1}_j - R_{jX}M_{XY}R^\dagger_{Yi}} \tag{2}$$

$$= d^{(-)}_{ij} + d^{(+)}_{ij} \tag{3}$$

with $R^2_i = R_{iX}M_{XY}R^\dagger_{Yi} \in \mathbb{R}^N$. While the purely imaginary-parts, i.e. iB_{ij} are obtained from the anti-Hermitian partition A_{XY}:

$$B_{ij} = \mathbf{1}_i\kappa_j + \kappa_i\mathbf{1}_j + i\underbrace{R_{iX}A_{XY}R^\dagger_{Yj}}_{a^{(+)}_{ij}} + i\underbrace{R_{jX}A_{XY}R^\dagger_{Yi}}_{a^{(-)}_{ij}}$$

with $\kappa_i = -iR_{iX}A_{XY}R^\dagger_{Yi} \in \mathbb{R}^N$, and $B_{ij} \in \mathbb{R}^{N\times N}$. Above B_{ij} is entirely real and symmetric. The pieces $a^{(\pm)}_{ij} = -(a^{(\pm)})^*_{ji} \in \mathbb{C}^{N\times N}$.

2.1 BILINEAR DISSIMILARITY MATRIX

We now discuss concrete choices for the bilinear dissimilarity (BD) matrix $\mathcal{W} \in \mathbb{C}^{D\times D}$ used in the generalized squared distance. A natural construction comes from Koopman/EDMD (Koopman (1931); Schmid (2010); Brunton & Kutz (2022)): given a sequence of feature vectors $\{R_{iX}\}_{i=1}^N$ (complex, not necessarily centered), define the zero–lag covariance and the τ–lag cross–covariance

$$C(0)_{XY} = \frac{1}{N}\sum_{i=1}^{N} R^\dagger_{Xi}R_{iY} \ \in \ \mathbb{C}^{D\times D}, \qquad C(\tau)_{XY} = \frac{1}{N-\tau}\sum_{i=1}^{N-\tau} R^\dagger_{Xi}R_{i+\tau,Y} \ \in \ \mathbb{C}^{D\times D}, \tag{4}$$

with $R^\dagger_i$ the conjugate transpose. We then set

$$\mathcal{W}_{XY} = C(0)^{-1}_{XZ}\,C(\tau)_{ZY}, \tag{5}$$

optionally using Tikhonov regularization $C(0)^{-1} \rightsquigarrow (C(0) + \lambda I_D)^{-1}$ or the whitened form $\mathcal{W} = C(0)^{-1/2}C(\tau)\,C(0)^{-1/2}$ for numerical stability. Here $C(0)$ is Hermitian PSD, whereas $C(\tau)$ and $\mathcal{W}$ are generally complex and asymmetric, which is desirable when modeling directed effects.

As an alternative, one can generate an asymmetric BD matrix from a random off–diagonal Wishart block. Draw $Q \in \mathbb{C}^{(2D)\times p}$ with i.i.d. entries (e.g., complex normal), form the Wishart matrix $W = QQ^\dagger \in \mathbb{C}^{(2D)\times(2D)}$, and partition

$$W = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22}\end{pmatrix}, \qquad W_{\sigma\sigma'} \in \mathbb{C}^{D\times D}.$$

Taking the off–diagonal block $\mathcal{W} := \frac{1}{p} W_{12}$ (or $W_{21} = W_{12}^*$) yields a complex, non-Hermitian BD matrix that induces an asymmetric similarity.

These examples recover familiar special cases: for $\mathcal{W}_{XY} = \delta_{XY}$ we obtain ordinary Euclidean squared distances (the law of cosines) and a trivial anti-Hermitian part $B_{XY} = 0_{XY}$; for learned Mahalanobis metrics one may take $\mathcal{W} = M \succeq 0$; and sample-dependent or anisotropic variants are obtained by letting $\mathcal{W}$ depend on (i, j) or by applying local scaling to $C(0)/C(\tau)$ before forming equation 5.

3 SOFTMAX

Let's define the following nonlinear-operator applied on a matrix $z_{ij} \in \mathbb{R}^{N\times N}$ (or the real part of a complex matrix), the i-axis softmax (the Boltzmann distribution of $-z_{ij}$ with $\beta = 1$, i.e. $\mathrm{softmax}_i(-z_{ij}) = \mathrm{Boltzmann}_i(z_{ij})$) is elementwise exponentiation followed by normalization over i, yielding an i-stochastic (column-stochastic) matrix. Analogously for the j-axis:

$$(\mathrm{softmax}_i(z_{ij}))_{ij} = \frac{e^{\odot(z_{ij})}}{\mathbf{1}_k\, e^{\odot(z_{kj})}} \;=\; \frac{e^{\odot(z_{ij})}}{\sum_k e^{\odot(z_{kj})}}, \tag{6}$$

$$(\mathrm{softmax}_j(z_{ij}))_{ij} = \frac{e^{\odot(z_{ij})}}{e^{\odot(z_{ik})}\,\mathbf{1}_k} \;=\; \frac{e^{\odot(z_{ij})}}{\sum_k e^{\odot(z_{ik})}}. \tag{7}$$

In particular, $\sum_i(\mathrm{softmax}_i(z_{ij}))_{ij} = \mathbf{1}_j$ for all j, and $\sum_j(\mathrm{softmax}_j(z_{ij}))_{ij} = \mathbf{1}_i$ for all i. Note in this definition only the real-part contributes to the normalization. Now let's consider the Product-of-Experts (PoE) identity of softmax. $z_{i|j} := (\mathrm{softmax}_i(z_{ij}))_{ij},\ s_{i|j} := (\mathrm{softmax}_i(s_{ij}))_{ij}$. Then

$$\begin{aligned}
(\mathrm{softmax}_i(z_{ij} + s_{ij}))_{ij} &= \mu_j\, z_{i|j}\, s_{i|j}, \quad \mu_j = \left(\sum_k z_{k|j} s_{k|j}\right)^{-1}. \\
&= \frac{(\mathrm{softmax}_i(z_{ij}))_{ij}\,(\mathrm{softmax}_i(s_{ij}))_{ij}}{\sum_k (\mathrm{softmax}_i(z_{kj}))_{kj}\,(\mathrm{softmax}_i(s_{kj}))_{kj}}. \\
(\mathrm{softmax}_j(z_{ij} + s_{ij}))_{ij} &= \tilde{\mu}_i\, \tilde{z}_{i|j}\, \tilde{s}_{i|j}, \quad \tilde{\mu}_i = \left(\sum_k \tilde{z}_{i|k}\, \tilde{s}_{i|k}\right)^{-1}.
\end{aligned}$$

$$(\mathrm{softmax}_i(z_{ij} + s_{ij}))_{ij} = \frac{z_{i|j}\, s_{i|j}}{\sum_k z_{k|j}\, s_{k|j}} \quad . \tag{8}$$

This was similarly done for $\tilde{z}_{i|j} := (\mathrm{softmax}_j(z_{ij}))_{ij},\ \tilde{s}_{i|j} := (\mathrm{softmax}_j(s_{ij}))_{ij}$. Now for the special case: $s_{ij} = 1_i\, c_j$ (analogously for $u_i \mathbf{1}_j$). Then

$$\begin{aligned}
s_{i|j} &= \frac{e^{c_j}}{\sum_k e^{c_j}} = \frac{1}{N} \quad \text{(uniform in } i\text{)}, \\
(\mathrm{softmax}_i(z_{ij} + s_{ij}))_{ij} &= \frac{z_{i|j}\,(1/N)}{\sum_k z_{k|j}\,(1/N)} \;=\; z_{i|j} \;=\; (\mathrm{softmax}_i z_{ij})_{ij}.
\end{aligned}$$

Hence softmax has a kind of shift-invariance:

$$\mathrm{softmax}_i\big(z_{ij} + \mathbf{1}_i\, c_j\big) = \mathrm{softmax}_i(z_{ij}) \quad , \tag{9}$$

$$\mathrm{softmax}_j\big(z_{ij} + u_j \mathbf{1}_j\big) = \mathrm{softmax}_j(z_{ij}) \quad , \tag{10}$$

and is a direct corollary of the product-of-experts identity.

4 SOFTMAX APPLICATION

Given our dataset is a point-cloud, we may define combinatorial-Laplacian as:

$$L_{ij} = D_i \delta_{ij} - \Omega_{ij} \quad . \tag{11}$$

With $\Omega_{ij} = \Omega_{ji} \in \mathbb{R}^{N\times N}$ the weighted adjacency-matrix (of the points), and $D_i = \sum_j \Omega_{ij}$ is called the degree vector. In order to create the diffusion-map random-walk-Laplacian matrix we define the following:

$$\Delta_{ij} = D_i^{\odot -1} L_{ij} = \delta_{ij} - D_i^{\odot -1}\Omega_{ij} = \delta_{ij} - P_{ij} \quad . \tag{12}$$

Above P_{ij} is the Markov transition-matrix (row-stochastic, i.e. over index j), because it shares a shifted spectrum to the random-walk Laplacian on the complete graph. Now our similarity measure in section 2 now may be used to define a weighted adjacency-matrix on the complete-graph, connecting all data-points with each other. However, in that section we had square-distances, i.e. dissimilar samples had a large value, to obtain similarities (dissimilar samples should be near zero) we use the Gaussian-Radial-Basis-Function, or the negative-exponentiation. This is the Diffusion-Map operator (with $\beta \in \mathbb{R}^+$ an inverse-temperature parameter playing the scale-parameter role, also termed ε, σ^2 in the literature):

$$P_{ij} = \text{softmax}_j\big(-\beta D_{ij}^2\big) \tag{13}$$

$$P_{ij} = \text{softmax}_j\big(-\beta\left(d_{ij}^- + d_{ij}^+\right)\big) \quad . \tag{14}$$

The diffusion-map-operator can be normalized-symmetrically, for numerics i.e. ease of diagonalization [1]. However, what about the case of just the $d_{ij}^{\pm}$? These yield the two attention-matrices:

$$\mathcal{A}_{ij}^- = \text{softmax}_i\big(-\beta d_{ij}^-\big) \tag{15}$$

$$\mathcal{A}_{ij}^+ = \text{softmax}_j\big(-\beta d_{ij}^+\big) \quad . \tag{16}$$

For the queries$\to$keys $(+)$ and keys$\to$queries $(-)$ versions. This can be shown for the most commonly used form $\mathcal{A}_{ij}^+$:

$$\begin{aligned}
\mathcal{A}_{ij} &= \text{softmax}_j\big(\beta Q_{iX} K_{Xj}^\top\big) = \text{softmax}_j\Big(\beta R_{iY} W_{YX}^{(Q)} W_{XZ}^{(K)} R_{Zj}^\dagger\Big) \\
&= \text{softmax}_j\Big(-\beta\left(-R_{iY} W_{YX}^{(Q)} W_{XZ}^{(K)} R_{Zj}^\dagger\right)\Big) \\
&= \text{softmax}_j\left(-\beta\left(\mathbf{1}_i\left(R_{iY} W_{YX}^{(Q)} W_{XZ}^{(K)} R_{Zj}^\dagger\right)_j - R_{iY} W_{YX}^{(Q)} W_{XZ}^{(K)} R_{Zj}^\dagger\right)\right) \\
&= \text{softmax}_j\big(-\beta d_{ij}^+\big) \quad .
\end{aligned}$$

And the connection to the diffusion-maps-operator P_{ij} is made using the PoE identity:

$$P_{ij} = \mu_i \mathcal{A}_{ij}^- \mathcal{A}_{ij}^+ \quad , \tag{17}$$

with $\mu_i = (\mathcal{A}_{kj}^- \mathcal{A}_{kj}^+)^{-1}$ being the usual row-normalization. Hence, succinctly we can have $P \propto \mathcal{A}^- \odot \mathcal{A}^+ = \mathcal{A} \odot \mathcal{A}^T = S$ (a twisted Hadamard product). This operator can be shown[2] to be symmetric[3].

5 Gauge Equivariance

We return to the skew–Hermitian similarity in §2. Let $a_{ij}^{(+)} = R_{iX} A_{XY} R_{Yj}^\dagger \in \mathbb{C}^{N\times N}$ arise from the anti-Hermitian part $A^\dagger = -A$ of the BD matrix; in particular $a_{ii}^{(+)}$ is purely imaginary, so $\Re\, a_{ii}^{(+)} = 0$. From the phase of $a_{ij}^{(+)}$ we define an edge connection (a $U(1)$ parallel transport)

$$U_{ij} \;=\; \exp\big(i\, q\, \arg a_{ij}^{(+)}\big), \qquad |U_{ij}| = 1, \;\; q \in \mathbb{R}, \tag{18}$$

[1] Symmetric normalization uses inverse-square-root $D_j^{-1/2}$ applied to both rows and columns, i.e. $S_{ij} = D_i^{-1/2}\Omega_{ij}D_j^{-1/2}$. A matrix normalized symmetrically shares the same spectrum as the row-normalized version.

[2] $S_{ji} = \mathcal{A}_{ji}\mathcal{A}_{ij}{}^* = (\mathcal{A}_{ij}(\mathcal{A}_{ji})^*)^* = (S_{ij})^*$.

[3] Assume real-logits $d^\pm \in \mathbb{R}^{N\times N}$ satisfy $d^+ = (d^-)^\top$ and use inverse-temperature β. Define bidirectional attentions $\mathcal{A}^- = \text{softmax}_i(-\beta\, d^-), \mathcal{A}^+ = \text{softmax}_j(-\beta\, d^+)$. Then,

$$\big(\text{softmax}_i(d^+)\big)_{ij} = \frac{e^{d_{ij}^+}}{\sum_k e^{d_{ik}^+}} = \frac{e^{d_{ji}^-}}{\sum_k e^{d_{ki}^-}} = \big(\text{softmax}_j(d^-)\big)_{ji},$$

hence $\mathcal{A}^+ = \big(\mathcal{A}^-\big)^T$.

with the convention $U_{ij} = 1$ if $a_{ij}^{(+)} = 0$. This construction gives the standard properties: unit modulus, $U_{ji} = U_{ij}^*$ so $U^\dagger = U^{-1}$, and $U(1)$ gauge covariance under node phases. Concretely, let $R = \mathrm{diag}(e^{i\phi_1}, \ldots, e^{i\phi_N})$ act on edge objects by

$$U_{ij} \longmapsto U'_{ij} = R_i \, U_{ij} \, R_j^{-1} = e^{\,i(\phi_i - \phi_j)} \, U_{ij}. \tag{19}$$

More generally, an edge matrix E_{ij} of *weight* $\alpha \in \mathbb{R}$ transforms as

$$E_{ij} \longmapsto E'_{ij} = R_i^{\alpha} \, E_{ij} \, R_j^{-\alpha} = e^{\,i\alpha(\phi_i - \phi_j)} \, E_{ij}. \tag{20}$$

The map $X \mapsto RXR^{-1}$ is a similarity, hence it changes eigenvectors but preserves eigenvalues; in particular, the spectrum is gauge-invariant under equation 19[4]. A key convenience of the Abelian case is that entrywise products respect the weight. Writing $(B \odot C)_{ij} := B_{ij} \, C_{ij}$, if B has weight α and C has weight β, then $B \odot C$ has weight $\alpha + \beta$:

$$(B \odot C)'_{ij} = (R_i^{\alpha} B_{ij} R_j^{-\alpha}) \, (R_i^{\beta} C_{ij} R_j^{-\beta}) = R_i^{\alpha+\beta} \, (B \odot C)_{ij} \, R_j^{-(\alpha+\beta)}. \tag{21}$$

In particular,

$$E \text{ and } C \text{ gauge–equivariant} \quad \Longrightarrow \quad E \odot C \text{ gauge–equivariant (weights add).} \tag{22}$$

Two immediate uses are: (i) the adjacency Ω_{ij} is invariant (weight 0), so $\Omega \odot E$ inherits the weight of E; and (ii) the masked edge operator

$$\tilde{\Omega}_{ij} := \Omega_{ij} \odot U_{ij}$$

is gauge-covariant of weight 1 by equation 19–equation 21. Using equation 20 and equation 22, all Abelian constructions in this section follow by inspection.

Now that our Abelian bridge is complete, we consider the feasibility of a non-Abelian bridge, motivated by equivariant work in Attention, Fuchs et al. (2020); Liao & Smidt (2023); Xu et al. (2023), and Diffusion maps, Singer & Wu (2012). Unfortunately, the twisted-Hadamard-covariance equation 22 for non-Abelian groups do not hold. For non-Abelian groups (e.g. $\mathrm{SU}(2)$) acting by $E_{ij} \mapsto R_i E_{ij} R_j^{-1}$ with noncommuting R_i, the Hadamard map is not an equivariant bilinear map in general (componentwise, duplicated free indices and noncommuting middle factors obstruct rewriting the transform as $R_i(\cdot)R_j^{-1}$). This is why Abelian "twisted-Hadamard" constructions cleanly preserve equivariance, while their non-Abelian analogues require either promotion to tensor-squared connections (on $V \otimes V$) or projection to scalar invariants; see Appendix A for a concise proof.

6 Conclusion

We established a *Diffusion–Attention Connection*: under mild and explicit conditions, diffusion maps arise as the *adjoint square* of self-attention built from the same bilinear dissimilarity. Concretely, let the (real) logits be $z_{ij} = -\beta \, d_{ij}^{+}$ induced by a Hermitian geometry operator M, and define first–order attention by a row/column softmax along a fixed axis. Then the symmetric second–order operator

$$P \propto \mathcal{A}^{(+)} \odot (\mathcal{A}^{(+)})^\dagger \tag{23}$$

is (up to a standard similarity) the diffusion-map operator constructed from the same kernel, hence it shares spectra and embeddings. This clarifies *when* attention behaves as diffusion and, equally, *when it does not* (e.g., mixed axes, missing normalization, or complex phases inside the softmax). Beyond this identity, we separated geometry from directionality via the split $\mathcal{W} = M + A$ (Hermitian M, anti-Hermitian A). The anti-Hermitian part induces a $U(1)$ edge phase

$$U_{ij} = \exp\big(iq \arg(R_i A R_j^\dagger)\big),$$

which we attach *outside* the softmax and re-normalize, yielding a gauge-equivariant, row-stochastic "magnetic" attention. This places magnetic/connection Laplacians on the same footing as attention, and explains design choices that preserve Hermiticity and Markovianity. The connection suggests

[4] If $\Omega v = \lambda v$ and $\Omega' := R\Omega R^{-1}$, then $\Omega'(Rv) = R(\Omega v) = \lambda(Rv)$. We use "gauge covariance" for the transformation law equation 20; spectral invariance is a consequence, not the definition.

simple attention-style predictors for ordered data (in the spirit of Nonlinear Laplacian Spectrum Analysis (NLSA), Takens (1980); Giannakis & Majda (2012; 2013).): given a context index $c, c' \in\in \{1, 2, \cdots, L\}$ and values $W^{(V)}$, first-order attention yields

$$R'_{cX} \;=\; \mathcal{A}^{(+)}_{cc'}\, R_{c'Y}\, W^{(V)}_{YX}, \tag{24}$$

while the diffusion counterpart uses the second–order operator

$$R'_{cX} \;=\; P_{cc'}\, R_{c'Y}\, W^{(V)}_{YX} \;=\; \Delta_{cc'}\, R_{c'Y}\, W^{(V)}_{YX} \quad (+\text{ skip }). \tag{25}$$

Here P is row-stochastic (random-walk), and $\Delta = I - P$ provides controllable smoothing; the adjoint-square view clarifies the relation among these updates.

The framework extends naturally to multi-head settings by combining headwise diffusions (e.g., mixtures $\sum_h w_h P^{(h)}$ or PoE kernels) and to sample-dependent anisotropy via $\mathcal{W}_{XY} \to \mathcal{W}_{ijXY}$, aligning with variable-bandwidth and anisotropic kernels Berry & Harlim (2016); Kushnir et al. (2012); related density-adapted practices in t-SNE Van der Maaten & Hinton (2008) fit the same lens. Attention provides a first–order, normalized view; diffusion is its symmetric second–order square. The split into M (geometry) and A (connection) yields principled, gauge-equivariant phased variants and suggests simple multi-head diffusion maps.

REFERENCES

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *ICLR*, 2014. arXiv:1409.0473.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003. doi: 10.1162/089976603321780317.

Tyrus Berry and John Harlim. Variable bandwidth diffusion kernels. *ACHA*, 40(1):68–96, 2016. doi: 10.1016/j.acha.2015.01.001.

Steven L. Brunton and J. Nathan Kutz. *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge University Press, 2nd edition, 2022.

Julio Candanedo. Linearized diffusion map. *arXiv:2507.14257*, 2025. doi: 10.48550/arXiv.2507.14257.

Ronald R Coifman and Stéphane Lafon. Diffusion maps. *ACHA*, 21(1):5–30, 2006.

Fabian B. Fuchs, Daniel E. Worrall, Volker Fischer, and Max Welling. Se(3)-transformers: 3d roto-translation equivariant attention networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. URL https://arxiv.org/abs/2006.10503.

Dimitrios Giannakis and Andrew J Majda. Nonlinear laplacian spectral analysis for time series with intermittency and low-frequency variability. *PNAS*, 109(7):2222–2227, 2012.

Dimitrios Giannakis and Andrew J Majda. Nonlinear laplacian spectral analysis: Capturing intermittent and low-frequency spatiotemporal patterns in high-dimensional data. *Stat. Anal. Data Min.*, 6(3):180–194, 2013.

Bernard O. Koopman. Hamiltonian systems and transformation in hilbert space. *PNAS*, 17(5): 315–318, 1931. doi: 10.1073/pnas.17.5.315.

Dan Kushnir, Ali Haddad, and Ronald R. Coifman. Anisotropic diffusion on sub-manifolds with application to earth structure classification. *ACHA*, 32(2):280–294, 2012. doi: 10.1016/j.acha.2011.06.002.

Yi-Lun Liao and Tess Smidt. Equiformer: Equivariant graph attention transformer for 3d atomistic graphs. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2206.11990.

Tobias Christian Nauen, Sebastián Palacio, and Andreas Dengel. Taylorshift: Shifting the complexity of self-attention from squared to linear (and back) using taylor-softmax. 15306:1–16, 2025. doi: 10.1007/978-3-031-78172-8_1. arXiv:2403.02920.

Peter J. Schmid. Dynamic mode decomposition of numerical and experimental data. *Journal of Fluid Mechanics*, 656:5–28, 2010. doi: 10.1017/S0022112010001217.

Amit Singer and Hau-Tieng Wu. Vector diffusion maps and the connection laplacian. *Communications on Pure and Applied Mathematics*, 65(8):1067–1144, 2012. doi: 10.1002/cpa.21395. arXiv:1102.0075v1.

Floris Takens. Detecting strange attractors in turbulence. *Dynamical Systems and Turbulence, Warwick 1980*, pp. 366–381, 1980.

Yao-Hung Hubert Tsai, Shaojie Bai, Makoto Yamada, Louis-Philippe Morency, and Ruslan Salakhutdinov. Transformer dissection: An unified understanding for transformer's attention via the lens of kernel. pp. 4344–4353, 2019. doi: 10.18653/v1/D19-1443. URL https://aclanthology.org/D19-1443/.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(Nov):2579–2605, 2008.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS 2017*, 30:5998–6008, 2017. arXiv:1706.03762.

Mingyu Xu, Ziyi Wu, Jiahui Hou, and Kui Jia. Geometric equivariant vision transformers. In *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 216 of *Proceedings of Machine Learning Research*, pp. 2349–2360, 2023. URL https://proceedings.mlr.press/v216/xu23b/xu23b.pdf.

A HADAMARD PRODUCT IS NOT EQUIVARIANT FOR NON-ABELIAN ACTIONS

Let G be a non-Abelian matrix group[5] with node-wise action on edge tensors

$$B_{ij,\alpha\beta} \longmapsto B'_{ij,\alpha\beta} = (R_i)_{\alpha\gamma}\, B_{ij,\gamma\delta}\, (R_j^{-1})_{\delta\beta}, \qquad R_i \in G,$$

and define the Hadamard (entrywise) product $(B \odot C)_{ij,\alpha\beta} := B_{ij,\alpha\beta} C_{ij,\alpha\beta}$. Then, in general,

$$(B \odot C)' \;\neq\; R_i\, (B \odot C)\, R_j^{-1},$$

i.e. the Hadamard map is not G–equivariant under left–right conjugation, unless the action reduces to one-dimensional characters (the Abelian case).

With C transforming analogously to B, a direct calculation gives

$$(B \odot C)'_{ij,\alpha\beta} = \big(R_i\big)_{\alpha\gamma}\big(R_i\big)_{\alpha\mu}\, B_{ij,\gamma\delta}\, C_{ij,\mu\nu}\, \big(R_j^{-1}\big)_{\delta\beta}\big(R_j^{-1}\big)_{\nu\beta}. \tag{26}$$

Equivariance would require

$$(B \odot C)'_{ij,\alpha\beta} \stackrel{?}{=} (R_i)_{\alpha\rho}\, (B \odot C)_{ij,\rho\sigma}\, (R_j^{-1})_{\sigma\beta} = (R_i)_{\alpha\rho}\, B_{ij,\rho\sigma} C_{ij,\rho\sigma}\, (R_j^{-1})_{\sigma\beta}.$$

Comparing with equation 26 reveals two structural obstructions:

(i) Duplicated free indices. The factors $(R_i)_{\alpha\gamma}(R_i)_{\alpha\mu}$ (and similarly on the right) entail two independent copies of the representation at the same free index α. To match the target form, one would need a G–equivariant bilinear map $V \otimes V \to V$ that canonically collapses γ, μ into a single index ρ (and similarly δ, ν into σ).

(ii) Noncommuting middle factors. Even abstractly attempting to factor equation 26 as $R_i(\cdot)R_j^{-1}$ fails because the two R_i's (and two R_j^{-1}'s) act on different tensor legs and does not commute through B and C in a way that produces a single conjugation of $B \odot C$.

If $G = U(1)$ acts by characters $R_i = e^{i\phi_i}$, then entries pick up commuting phases and

$$(B \odot C)'_{ij} = R_i^{\alpha+\beta}\, (B \odot C)_{ij}\, R_j^{-(\alpha+\beta)},$$

so weights add and equivariance holds.

[5] With representation matrix $(R_i)_{\alpha\beta} = \exp\big(i\, \theta_a^i\, T^a_{\alpha\beta}\big)$ (matrix exponential), Lie algebra generators T^a, and node-dependent coefficients θ_a^i.