Gating Enables Curvature: A Geometric Expressivity Gap in Attention

Anonymous Authors[1]

Abstract

Multiplicative gating is widely used in neural architectures, but its use in attention is recent and its geometric role remains unclear. We model attention outputs as Gaussian means and study their Fisher Rao geometry. At the operator level, ungated attention induces flat manifolds through affine value mixing. Gating enables curved geometries, including positive curvature. This reveals a geometric expressivity gap. Furthermore, we identify a structured regime where curvature accumulates under composition, leading to a systematic amplification effect with depth. Empirically, gated models show higher curvature and perform better on nonlinear tasks, with no consistent gains on linear ones.

1. Introduction

Attention mechanisms are a core part of transformer models used for sequence modeling and large language models (Vaswani et al., 2017). Prior work studies the expressivity of attention mechanisms, including universality and depth separation (Yun et al., 2020; Levine et al., 2020; Pérez et al., 2019; Wang & E, 2024). These works study what input-output mappings attention models can represent, independent of training.

Yet, the intrinsic geometry of attention representations remains largely unexplored. It offers a complementary notion of expressivity, focused on internal organization rather than computable functions.

We study the geometry induced by attention layers by modeling their outputs as parameters of a statistical manifold equipped with the Fisher–Rao metric, whose curvature captures intrinsic properties of the representation space (Amari, 1985; Amari & Nagaoka, 2000; Amari, 2016; Liang et al., 2019; Amari et al., 2019; Kim et al., 2022). For Gaussian decoders, the Fisher geometry associated with the mean parameter is Euclidean, so any nonzero curvature arises solely from the structure of the attention mapping itself. This choice isolates the geometric structure induced by the architecture, rather than by the statistical model (Rao, 1945; Amari, 1985).

Recent architectural developments have introduced attention mechanisms augmented with multiplicative gating, reporting improvements in training stability, scaling behavior, and long-context modeling (Qiu et al., 2025; Zhang et al., 2024; Cho et al., 2014; Danihelka et al., 2016; Wu et al., 2016; Dauphin et al., 2017; Krishnamurthy et al., 2022). While gating shows practical benefits, its geometric role remains unclear. Ungated attention is affine and induces flat statistical manifolds. Multiplicative gating removes this constraint and enables non-flat geometries.

Figure 1 illustrates this distinction. We study gated and ungated attention geometrically, with the following contributions:

- Using Fisher–Rao geometry, we show operator-level flatness of ungated attention due to affine value mixing.
- We prove multiplicative gating enables non-flat geometries unattainable in the ungated setting.
- We show this separation persists in standard content-aware attention.
- We establish robustness and structured residual amplification of gated curvature.
- We provide empirical evidence that gating increases curvature and improves performance on nonlinear tasks.

Representation-level implication. Ungated attention yields flat representations, limiting non-affine structure within a single block. Multiplicative gating removes this constraint, enabling intrinsically curved and richer geometries, especially for nonlinear data. These results characterize single-block geometry, independent of depth, and parallel combinatorial expressivity results (Montúfar et al., 2014), where depth increases complexity, while gating provides a geometric mechanism for inducing and amplifying curvature.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

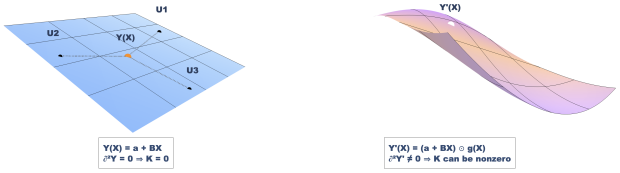

Figure 1. **Geometric intuition for curvature generation in attention.** Left: Ungated attention produces affine combinations of value vectors, so outputs lie in the affine hull $\mathrm{aff}\{U_1, U_2, U_3\}$, yielding a flat representation manifold with zero curvature. Right: Gating introduces element-wise modulation $Y'(X) = Y(X) \odot g(X)$, breaking affine structure and enabling nonzero curvature in the induced representation manifold.

2. Preliminaries

2.1. Notations

Linear algebra. All vectors are real-valued. We write $\mathbb{R}^D$ for the ambient representation space and use lowercase letters for vectors and uppercase letters for matrices. Inner products and norms are taken with respect to the standard Euclidean structure on $\mathbb{R}^D$. For vectors $a, b \in \mathbb{R}^D$, $a \odot b$ denotes their componentwise (Hadamard) product.

Probability simplex. The probability simplex in $\mathbb{R}^n$ is denoted by

$$\Delta^{n-1} := \left\{ \alpha \in \mathbb{R}^n \mid \alpha_i \geq 0,\ \sum_{i=1}^{n} \alpha_i = 1 \right\}.$$

Gaussian statistical manifolds. We consider Gaussian distributions of the form

$$p(y \mid \mu) = \mathcal{N}(y; \mu, I_D),$$

where the mean parameter $\mu \in \mathbb{R}^D$ varies and the covariance is fixed. This defines a statistical manifold parameterized by μ.

Fisher–Rao geometry and fixed covariance. The Fisher–Rao metric provides a canonical, parameterization-invariant notion of geometry for statistical models. For a Gaussian location family with fixed covariance $\Sigma \succ 0$,

$$p(y \mid \mu) = \mathcal{N}(y; \mu, \Sigma),$$

the Fisher information is

$$I(\mu) = \Sigma^{-1}.$$

For any parameterization $\mu = \mu(\phi)$, the induced Riemannian metric is

$$g_{ij}(\phi) = \partial_i \mu(\phi)^\top \Sigma^{-1} \partial_j \mu(\phi),$$

corresponding to a Mahalanobis inner product. Defining the whitened mean $\tilde{\mu}(\phi) = \Sigma^{-1/2}\mu(\phi)$, this reduces to

$$g_{ij}(\phi) = \langle \partial_i \tilde{\mu}(\phi), \partial_j \tilde{\mu}(\phi) \rangle,$$

showing that any fixed non-isotropic covariance is equivalent to the isotropic case under a linear change of coordinates.

As a result, the intrinsic geometry of the statistical manifold is determined entirely by the embedding $\phi \mapsto \mu(\phi) \subset \mathbb{R}^D$. Since the metric is constant, any nonzero curvature arises from the embedding $\mu(\phi)$ rather than from the statistical model. Allowing the covariance to vary would introduce additional curvature through the metric itself. Fixing covariance therefore isolates geometric effects attributable to the attention mapping.

The Gaussian construction should be viewed as a coordinate-invariant realization of a constant Euclidean geometry on the representation space, rather than a modeling assumption on the data distribution.

Curvature. We use standard notions from Riemannian geometry. The Riemann curvature tensor is denoted by $\mathcal{R}$ and sectional curvature by K. A manifold is flat if $\mathcal{R} \equiv 0$.

Attention outputs. Throughout the paper, attention outputs $Y(X) \in \mathbb{R}^D$ are interpreted as mean parameters of Gaussian distributions. Precise definitions of ungated and gated attention are given in Section 2.2.

2.2. Problem Setup and Formulation

We formalize attention architectures and the geometric objects they induce. Our goal is to compare the intrinsic geometry of representations produced by ungated and gated attention mechanisms, independent of training dynamics and downstream tasks.

Attention architecture. We consider a single-head attention layer of the form introduced by Vaswani et al. (2017). Let X denote an input sequence, and let

$$U_1, \ldots, U_n \in \mathbb{R}^D$$

be value vectors obtained after value and output projections. The ungated attention output is

$$Y(X) = \sum_{i=1}^{n} \alpha_i(X) U_i, \quad \alpha(X) \in \Delta^{n-1},$$

where the attention weights $\alpha_i(X)$ are produced by a softmax over query–key interactions. For fixed value vectors, $Y(X)$ lies in $\mathrm{aff}\{U_1, \ldots, U_n\}$. The softmax only parameterizes how this affine hull is traversed.

Gated attention. We consider attention augmented with a multiplicative gate. Let $X_g(X) \in \mathbb{R}^m$ denote a gating input, $W_\theta \in \mathbb{R}^{m\times D}$ a trainable matrix, and $\sigma : \mathbb{R} \to (0, 1)$ a smooth elementwise nonlinearity. The gated attention output is

$$Y'(X) = Y(X) \odot \sigma(X_g(X)W_\theta).$$

When the gate is constant, this reduces to the ungated case up to a fixed rescaling. Allowing the gate to vary introduces nonlinear modulation of the attention output.

Representation geometry. To study the geometry of attention representations, we associate each output with a Gaussian distribution whose mean is given by the attention output,

$$p(y \mid \mu) = \mathcal{N}(y; \mu, I_D), \quad \mu(X) = Y(X).$$

As discussed above, the Fisher–Rao metric is Euclidean on the mean parameter, so the induced geometry is determined entirely by the embedding $\mu(X) \subset \mathbb{R}^D$. In particular, curvature reflects nonlinear structure in the mapping $X \mapsto Y(X)$.

This provides a principled way to study the intrinsic geometry of representations produced by attention layers, independent of parameterization.

Immersions. A smooth map $F : U \subset \mathbb{R}^d \to \mathbb{R}^D$ is called an immersion if its Jacobian $DF(\phi)$ has full rank at every point. This ensures that the induced metric is nondegenerate and that curvature is well-defined on the image manifold.

Problem formulation. We compare the statistical manifolds induced by ungated and gated attention,

$$\mathcal{M}_{\text{ung}} := \{\mathcal{N}(Y(X), I_D) \mid X \in \mathcal{X}\},$$
$$\mathcal{M}_{\text{gat}} := \{\mathcal{N}(Y'(X), I_D) \mid X \in \mathcal{X}\}.$$

Our goal is to characterize the geometric constraints imposed by the affine structure of ungated attention, and to determine whether multiplicative gating enlarges the class of realizable geometries. In particular, we ask:

- What geometric structures are imposed by the affine form of ungated attention?
- Can multiplicative gating enable representations with nonzero intrinsic curvature?
- Does gating strictly enlarge the range of geometries realizable by attention?

3. Main Results

This section develops the main theoretical results describing the intrinsic geometry of statistical manifolds induced by ungated and gated attention operators. Proofs of all results in this section are deferred to the appendix.

Roadmap. We first show that ungated attention induces flat manifolds (Section 3.1). We then show that gating breaks this affine constraint and enables non-flat geometries, creating an expressivity gap (Section 3.2). Next, we show that curvature can accumulate with depth (Section 3.3). Finally, we show that this curvature is robust and generic (Section 3.4).

3.1. Flatness of Ungated Attention

Theorem 3.1 (Flatness of ungated attention manifolds)**.** *An ungated attention operator whose outputs are affine combinations of value vectors induces an intrinsically flat statistical manifold.*

Let $U_1, \dots, U_n \in \mathbb{R}^D$ be fixed value vectors and define the convex hull

$$\mathcal{U} := \operatorname{conv}(U_1, \dots, U_n) = \left\{ \sum_{i=1}^{n} \alpha_i U_i \;\middle|\; \alpha \in \Delta_{n-1} \right\} \subset \mathbb{R}^D.$$

Let $k := \dim \operatorname{aff}\{U_1, \dots, U_n\} = \dim \operatorname{span}\{U_1 - U_n, \dots, U_{n-1} - U_n\}$. On the relative interior $\operatorname{relint}(\mathcal{U})$, choose affine coordinates

$$\varphi : \Omega \subset \mathbb{R}^k \to \operatorname{relint}(\mathcal{U}),$$
$$\mu_{\text{ung}}(\phi) := \varphi(\phi) = a + B\phi, \quad \text{for } \phi \in \Omega.$$

where $B \in \mathbb{R}^{D\times k}$ has full column rank. Consider the statistical manifold

$$\mathcal{M}_{\text{ung}} := \{\mathcal{N}(\mu_{\text{ung}}(\phi), I_D) \mid \phi \in \Omega\},$$

equipped with the Fisher–Rao metric. Then the induced Riemannian metric on Ω is constant:

$$g^{\text{ung}}_{ij}(\phi) = \langle \partial_i \mu_{\text{ung}}(\phi), \partial_j \mu_{\text{ung}}(\phi) \rangle = C_{ij},$$

where $C = (C_{ij})$ is a fixed positive-definite matrix. Consequently, the Levi–Civita connection has vanishing Christoffel symbols, and the Riemann curvature tensor and sectional curvatures satisfy

$$R^{\text{ung}} \equiv 0, \qquad K_{\text{ung}} = 0 \quad \text{on } \Omega.$$

Thus, $\mathcal{M}_{\text{ung}}$ is isometric to an open subset of the Euclidean space $(\mathbb{R}^k, C)$ and is intrinsically flat.

This flatness is invariant under reparameterization. For any diffeomorphism $\psi : \hat{\Omega} \to \Omega$, the pullback metric $\tilde{g} = \psi^* g^{\text{ung}}$ satisfies

$$\mathcal{R}(\tilde{g}) = \psi^* \mathcal{R}(g^{\text{ung}}) \equiv 0.$$

Thus no smooth reparameterization can induce nonzero curvature.

The vanishing curvature therefore shows a fundamental geometric limitation of ungated attention, arising from the affine form of its output together with the Euclidean Fisher geometry of the Gaussian mean family.

Theorem 3.1 shows that ungated attention produces intrinsically flat representations. As a result, any target geometry with nonzero curvature lies outside this class. We now formalize this as a quantitative separation.

Theorem 3.2 (Curvature approximation lower bound)**.** *Let $F^\star : U \to \mathbb{R}^D$ be a smooth immersion defined on an open set $U \subset \mathbb{R}^m$, and let $K_{F^\star}(u, \Pi)$ denote the sectional curvature at u along a 2-plane $\Pi \subset T_u U$. Assume*

$$K_{F^\star}(u, \Pi) \geq \kappa_0 > 0 \quad \forall u \in U, \forall \Pi \subset T_u U, \dim(\Pi) = 2.$$

Let F_{ung} be any representation realized by a single ungated attention block. Then

$$\sup_{u \in U} \sup_{\substack{\Pi \subset T_u U \\ \dim(\Pi)=2}} \left| K_{F_{\text{ung}}}(u, \Pi) - K_{F^\star}(u, \Pi) \right| \geq \kappa_0.$$

Corollary 3.3 (One-block efficiency gap)**.** *Let $F^\star : U \to \mathbb{R}^D$ be a smooth immersion whose induced representation manifold satisfies*

$$K_{F^\star}(u, \Pi) \geq \kappa_0 > 0 \quad \forall u \in U, \forall \Pi \subset T_u U, \dim(\Pi) = 2.$$

Then for any $\varepsilon < \kappa_0$, no single ungated attention operator block can approximate $F^\star$ within curvature error ε.

Remark 3.4 (Connection to low-rank structure)**.** The affine form of ungated attention implies that its outputs lie in a low-dimensional affine subspace determined by the value vectors. This recovers the well-known low-rank bottleneck in attention from a geometric perspective. In our formulation, this affine constraint induces a constant metric and hence intrinsic flatness of the representation manifold. [See Appendix A.6 for detailed discussion.]

Ungated attention is flat due to its affine structure. We ask whether multiplicative gating preserves this constraint. The next lemma shows that it generically breaks affinity, enabling nontrivial geometry.

3.2. Gating Enables Curvature

Lemma 3.5 (Multiplicative gating generically destroys affine structure)**.** *Let $U \subset \mathbb{R}^d$ be a connected open set with $d \geq 1$. Let*

$$Y(\phi) = a + B\phi, \qquad B \in \mathbb{R}^{D \times d}, \ \operatorname{rank}(B) = d,$$

and denote by $B_i \in \mathbb{R}^D$ the i-th column of B. For $g \in C^2(U, \mathbb{R}^D)$ define

$$\mu(\phi) = Y(\phi) \odot g(\phi),$$

where $\odot$ denotes the componentwise (Hadamard) product. Then for all i, j,

$$\partial_{ij}\mu = B_i \odot \partial_j g + B_j \odot \partial_i g + Y \odot \partial_{ij} g.$$

Moreover, the set

$$\mathcal{A} = \{g \in C^2(U, \mathbb{R}^D) : \mu \text{ is affine on } U\}$$

is a proper closed linear subspace of $C^2(U, \mathbb{R}^D)$ (with its Fréchet topology). Hence $\mathcal{A}$ is nowhere dense and meagre, and for residual $g \in C^2(U, \mathbb{R}^D)$, the map μ is non-affine. Equivalently, non-affinity holds generically in the Baire category sense.

This shows that multiplicative gating breaks the affine structure of ungated attention. Since flatness arises from affinity, removing this constraint allows the induced metric to vary, enabling nontrivial curvature. We now construct explicit examples and show that such curvature arises within standard content-aware attention.

Theorem 3.6 (Concrete witness of non-flat geometry under multiplicative gating in content-aware attention)**.** *Consider a single-head attention layer with standard projections*

$$\begin{aligned} Q(X) &= XW_Q, \\ K(X) &= XW_K, \\ V(X) &= XW_V, \\ U(X) &= V(X)W_O. \end{aligned}$$

Let the output at position j be

$$Y_j(X) = \sum_{i=1}^{n} \alpha_{ji}(X) U_i(X),$$

where the weights $\alpha_{ji}(X)$ are obtained by applying the softmax to $Q(X)K(X)^\top$. and define the gated output

$$Y'_j(X) = Y_j(X) \odot \sigma(X_g(X)W_\theta).$$

Assume a Gaussian decoder $p(y \mid \mu) = \mathcal{N}(y; \mu, I_3)$.

Then there exist an open set $U \subset \mathbb{R}^2$, a smooth map $\phi \mapsto X(\phi)$, projection matrices, and $W_\theta \in \mathbb{R}^{m \times 3}$ with $\operatorname{rank}(W_\theta) = 3$ such that:

(i) The ungated map is affine,

$$Y_j(X(\phi)) = a + B\phi,$$

and induces a flat manifold with

$$K_{\text{ung}}(\phi) = 0.$$

(ii) The gated map parametrizes a spherical patch,

$$Y'_j(X(\phi)) = s(\phi), \quad s(\phi) = \begin{pmatrix} \cos\phi_1 \cos\phi_2 \\ \cos\phi_1 \sin\phi_2 \\ \sin\phi_1 \end{pmatrix},$$

and satisfies

$$K_{\mathrm{gat}}(\phi) = 1.$$

Corollary 3.7 (Geometric expressivity of multiplicative gating)**.** *Fix W_Q, W_K, W_V, W_O and the input map $\phi \mapsto X(\phi)$ from Theorem 3.6, and let W_θ be the only trainable parameter in*

$$Y_j'(X) = Y_j(X) \odot \sigma(X_g(X)W_\theta).$$

Then the family of manifolds obtained by varying W_θ strictly contains the ungated family (recovered by a constant gate). In particular, there exist W_θ such that

$$K_{\mathrm{ung}} = 0, \qquad K_{\mathrm{gat}} \equiv 1 > 0.$$

Thus, multiplicative gating strictly enlarges the class of realizable manifolds.

Corollary 3.8 (Extension to higher-dimensional decoders)**.** *The construction of Theorem 3.6 extends to Gaussian decoders in dimension $D \geq 3$.*

Specifically, embedding the spherical patch $s(\phi)$ in the first three coordinates of $\mathbb{R}^D$ and keeping the remaining coordinates constant yields a gated statistical manifold whose sectional curvatures in planes tangent to the spherical directions are strictly positive, while the ungated manifold remains flat.

Thus, the curvature gap persists for all $D \geq 3$.

Theorem 3.9 (Strict geometric separation)**.** *Let $\mathcal{G}_{\mathrm{ung}}$ and $\mathcal{G}_{\mathrm{gat}}$ denote the classes of local representation geometries induced (under the Gaussian embedding) by single ungated and gated attention blocks, respectively. Then*

$$\exists\, \mathcal{M} \in \mathcal{G}_{\mathrm{gat}} \quad \text{such that} \quad \mathcal{M} \not\cong_{\mathrm{loc}} \mathcal{N} \quad \forall\, \mathcal{N} \in \mathcal{G}_{\mathrm{ung}},$$

where $\cong_{\mathrm{loc}}$ denotes local isometry.

3.3. Depth Amplifies Curvature

Depth amplification under composition. Multiplicative gating induces curvature in a single layer, but its behavior under composition is unclear. We analyze a regime with coherent multiplicative structure across layers, yielding a local normal form in which curvature can be tracked and shown to amplify with depth.

Lemma 3.10 (Local gated-attention normal form)**.** *Let $F_L : \mathcal{X} \to \mathbb{R}^D$ be the representation map of an L-layer gated attention stack. Suppose there exist $x_0 \in \mathcal{X}$, an open set $x_0 \in U$, and local charts*

$$\chi_{\mathrm{in}} : U \to V \subset \mathbb{R}^2, \quad \chi_{\mathrm{out}} : F_L(U) \to W \subset \mathbb{R}^D,$$

with $\chi_{\mathrm{in}}(x_0) = 0$, such that

$$\chi_{\mathrm{out}} \circ F_L \circ \chi_{\mathrm{in}}^{-1}(u, v) = \left(u, v, \sum_{\ell=1}^{L} a_\ell \psi(u, v), 0, \ldots, 0\right),$$

for some $a_\ell > 0$ and $\psi \in C^2(V)$.

Let $A_L := \sum_{\ell=1}^{L} a_\ell$. Then

$$\chi_{\mathrm{out}} \circ F_L \circ \chi_{\mathrm{in}}^{-1}(u, v) = (u, v, A_L \psi(u, v), 0, \ldots, 0)\,.$$

The representation in Lemma 3.10 is a local structural reduction, not a universal characterization. It isolates a regime where multiplicative gating yields coherent accumulation across layers. The results characterize this mechanism. Such conditional analyses are standard for isolating expressivity.

The normal form is realizable by gated attention stacks in which residual connections add aligned gated contributions across layers. Thus, the amplification result should not be read as a consequence of gating or residual connections alone.

Lemma 3.11 (Realizability of the local normal form)**.** *Fix $L \geq 1$ and $a_1, \ldots, a_L > 0$. Let $U \subset \mathbb{R}^2$ be open and $\psi \in C^2(U)$ satisfy*

$$0 < a_\ell \psi(u, v) < 1 \quad \forall (u, v) \in U,\ \ell = 1, \ldots, L.$$

Then there exists an L-layer gated attention stack with

$$Y' = Y \odot \sigma(X_g W_\theta),$$

residual connections, and a local parametrization $(u, v) \mapsto X(u, v)$ such that

$$(u, v) \mapsto \left(u, v, \sum_{\ell=1}^{L} a_\ell \psi(u, v), 0, \ldots, 0\right).$$

Lemma 3.12 (Gaussian curvature of the graph normal form)**.** *Let $U \subset \mathbb{R}^2$ be open with $0 \in U$, and let $\psi \in C^2(U)$ satisfy $\nabla\psi(0) = 0$. For $A_L > 0$, define*

$$F_L(u, v) = (u, v, A_L \psi(u, v), 0, \ldots, 0) \in \mathbb{R}^D,$$
$$M_L := F_L(U).$$

Then the Gaussian curvature of M_L at $F_L(0)$ is

$$K_{M_L}(F_L(0)) = A_L^2 \det D^2\psi(0).$$

Theorem 3.13 (Depth-amplified intrinsic curvature under a gated-attention normal form)**.** *Under the assumptions of the preceding two lemmas,*

$$K_{M_L}\big(F_L(0)\big) = A_L^2 \det D^2\psi(0).$$

In particular, if there exists $a_0 > 0$ such that $a_\ell \geq a_0$ for all ℓ, then

$$\big|K_{M_L}\big(F_L(0)\big)\big| \geq a_0^2 \big|\det D^2\psi(0)\big|\, L^2.$$

If in addition $\det D^2\psi(0) > 0$, then

$$K_{M_L}\big(F_L(0)\big) \geq a_0^2 \det D^2\psi(0)\, L^2.$$

Curvature scales with the cumulative amplitude $A_L = \sum_{\ell=1}^{L} a_\ell$, not depth alone. A single layer with sufficiently large amplitude can achieve the same curvature. Thus, depth provides a systematic mechanism for accumulating curvature, rather than being strictly necessary.

Corollary 3.14 (Higher-dimensional embedding)**.** *Let $D \geq 3$ and define*

$$\tilde{F}_L(u,v) = (u, v, A_L\psi(u,v), 0, \ldots, 0) \in \mathbb{R}^D.$$

Then

$$K_{\tilde{M}_L}(\tilde{F}_L(0)) = A_L^2 \det D^2\psi(0).$$

Thus, embedding in higher dimensions does not affect the curvature scaling.

Interpretation. Additional coordinates that are constant do not contribute to the induced metric.

From scalar to vector-valued representations. The previous result considers aligned scalar contributions. We extend this to vector-valued graphs without alignment across layers.

Theorem 3.15 (Curvature of vector-valued graph representations)**.** *Let $U \subset \mathbb{R}^2$ be open with $0 \in U$, and let $\Phi_1, \ldots, \Phi_L : U \to \mathbb{R}^{D-2}$ be C^2. Define*

$$\Phi_L^{\mathrm{tot}}(u,v) := \sum_{\ell=1}^{L} \Phi_\ell(u,v),$$

$$F_L(u,v) := (u, v, \Phi_L^{\mathrm{tot}}(u,v)) \in \mathbb{R}^D.$$

and $M_L := F_L(U)$. If $D\Phi_L^{\mathrm{tot}}(0) = 0$, then

$$K_{M_L}(F_L(0)) = \sum_{\alpha=1}^{D-2} \det\big(D^2(\Phi_L^{\mathrm{tot}})^\alpha(0)\big) = \sum_{\alpha=1}^{D-2} \det\Big(\sum_{\ell=1}^{L} D^2\Phi_\ell^\alpha(0)\Big).$$

Remark 3.16. Although $D^2\Phi_L^{\mathrm{tot}} = \sum_\ell D^2\Phi_\ell$ is additive, curvature depends on determinants and is therefore not additive. Instead, cross-layer interactions appear, reflecting the combined second-order structure.

Corollary 3.17 (Aligned scalar regime)**.** *Under the assumptions of Theorem 3.15, suppose $\Phi_\ell(u,v) = a_\ell\psi(u,v)c$ for some $\psi \in C^2(U)$, scalars $a_\ell \geq a_0 > 0$, and fixed $c \in \mathbb{R}^{D-2}$. Assume $\nabla\psi(0) = 0$ and let $A_L = \sum_{\ell=1}^{L} a_\ell$. Then*

$$K_{M_L}(F_L(0)) = A_L^2\|c\|^2 \det D^2\psi(0),$$

and

$$|K_{M_L}(F_L(0))| \geq a_0^2 L^2 \|c\|^2 |\det D^2\psi(0)|.$$

In this aligned regime, second-order contributions share a common direction, eliminating cross-layer interactions and yielding coherent quadratic scaling. [Proof. See Appendix A.1.11]

From existence to robustness. The preceding results establish positively curved realizations and their amplification under composition. We further show that such curvature is stable under perturbations, forming a nonempty open set. Thus, the curvature gap is robust and locally generic.

3.4. Robustness and Genericity

Theorem 3.18 (Curvature gap and local robustness under multiplicative gating)**.** *Let $D \geq 3$ and $m \geq D$. There exist an open set $U \subset \mathbb{R}^2$, an affine map*

$$Y(\phi) = a + B\phi, \quad \mathrm{rank}(B) = 2,$$

and a smooth map $X^* : U \to \mathbb{R}^m$ such that for $W_\theta \in \mathbb{R}^{m\times D}$,

$$\mu_{W_\theta}(\phi) := Y(\phi) \odot \sigma(X^*(\phi)W_\theta)$$

defines a Gaussian family $\mathcal{F}(W_\theta) = \{\mathcal{N}(\mu_{W_\theta}(\phi), I_D)\}$ with induced metric

$$g_{ij}(\phi; W_\theta) = \langle \partial_i\mu_{W_\theta}(\phi), \partial_j\mu_{W_\theta}(\phi)\rangle.$$

(i) ***Flatness.*** *If the gate is indentically 1, then $\mu_{W_\theta}(\phi) = Y(\phi)$ and $K(\phi) \equiv 0$.*

(ii) ***Positive curvature.*** *There exists W_θ^* such that*

$$K(\phi; W_\theta^*) \equiv 1 \quad \forall\phi \in U.$$

(iii) ***Local robustness.*** *There exist a compact $K \subset U$, $c > 0$, and a neighborhood $\mathcal{O}$ of W_θ^* such that*

$$(\phi, W_\theta) \in R \quad \textit{(where } R \textit{ is defined in Lemma 3.19.)},$$
$$K(\phi; W_\theta) \geq \frac{c}{2} > 0 \quad \forall\phi \in K,\ W_\theta \in \mathcal{O}.$$

Lemma 3.19 (Regular locus and continuity of curvature)**.** *Let $R := \{(\phi, W_\theta) : \det g(\phi; W_\theta) > 0\}$. Then R is open, curvature is well-defined on R, and $(\phi, W_\theta) \mapsto K(\phi; W_\theta)$ is continuous on R.*

Lemma 3.20 (Uniform robustness on compact sets)**.** *Let $K \subset U$ be compact. Suppose*

$$(\phi, W_\theta^*) \in R \quad \textit{and} \quad K(\phi; W_\theta^*) \geq c > 0 \quad \forall\phi \in K.$$

Then there exists an open neighborhood $\mathcal{O}$ of W_θ^* such that

$$(\phi, W_\theta) \in R,$$
$$K(\phi; W_\theta) \geq \frac{c}{2} > 0 \quad \forall\phi \in K,\ W_\theta \in \mathcal{O}.$$

From robustness to genericity. The previous result shows that strictly positive curvature persists under parameter perturbations. We now show that non-flatness is not only stable but typical. At the level of local curvature, non-flat representations arise generically once affine structure is removed.

Theorem 3.21 (Pointwise generic non-flatness under multiplicative gating)**.** *Let $U \subset \mathbb{R}^d$ be open with $d \geq 2$. Let*

$$\mu_g(\phi) = Y(\phi) \odot g(\phi), \quad g \in C^2(U, \mathbb{R}^D),$$

where

$$Y(\phi) = a + B\phi, \quad a \in \mathbb{R}^D,\ B \in \mathbb{R}^{D\times d},\ \mathrm{rank}(B) = d.$$

Fix $\phi_0 \in U$, and assume that at least $d+1$ coordinates of $Y(\phi_0)$ are nonzero. Define

$$\mathcal{I}_{\phi_0} := \{g \in C^2(U, \mathbb{R}^D) : \mathrm{rank}\, D\mu_g(\phi_0) = d\},$$

and

$$\mathcal{N}_{\phi_0} := \{g \in \mathcal{I}_{\phi_0} : R_g(\phi_0) \neq 0\}.$$

Then $\mathcal{N}_{\phi_0}$ is open and dense in $\mathcal{I}_{\phi_0}$ in the C^2 topology.

Remark on assumptions. The immersion condition ensures a nondegenerate metric at ϕ_0, while the nonzero coordinate condition guarantees an accessible normal direction under multiplicative perturbations.

Robustness and genericity. Positive curvature induced by multiplicative gating is not a fragile artifact of specific constructions. Theorem 3.18 establishes that strictly positive curvature persists under parameter perturbations, forming an open set in parameter space. Complementarily, Theorem 3.21 shows that non-flatness is generic in the function-space model $\mu_g(\phi) = Y(\phi) \odot g(\phi)$, occurring on an open dense subset of C^2.

Together, these results imply that once the affine constraint of ungated attention is removed, flatness becomes non-generic, and curved geometries arise robustly under small perturbations.

4. Experiments

Our theory predicts that ungated attention yields low-curvature representations, while gating enables higher curvature. Since ungated models may exhibit small nonzero curvature in practice, we interpret curvature comparatively. We test whether this expressivity gap appears in learned models on a controlled task.

Setup. We consider a sequence classification task with a curved decision boundary. Each sample is generated by drawing a latent center $c \in [-2, 2]^2$ and forming a sequence of length 8 as noisy observations $x_i = c + \epsilon_i$, $\epsilon_i \sim \mathcal{N}(0, \sigma^2 I)$. Labels depend only on c: letting $r = \|c\|$ and $\theta = \mathrm{atan2}(c_2, c_1)$,

$$s(c) = \sin(2.5\theta) + 0.6(r - 1.2), \quad y = \mathbf{1}[s(c) > 0].$$

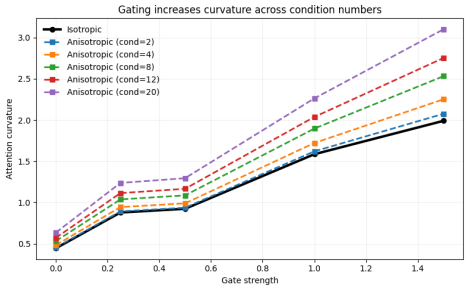

Figure 2. **Gating increases representation curvature.** Isotropic curvature is invariant across condition numbers, while anisotropic curvature varies with conditioning. Higher gate strength consistently increases curvature.

We use a single-block attention model with a classifier head. Inputs in $\mathbb{R}^2$ are projected to 64 dimensions, processed by scaled dot-product attention, mean-pooled, and passed to a two-layer MLP. We compare ungated attention, a pointwise SiLU nonlinearity, and multiplicative gating. Gated models use

$$Y' = Y \odot (1 + \alpha(\sigma(WY) - 1)),$$

where α controls gate strength ($\alpha = 1$ recovers the theoretical form). A residual connection and layer normalization follow the attention block.

Models are trained with AdamW over multiple seeds. Curvature is measured using a finite-difference proxy

$$\kappa(x) = \left\| \frac{f(x + \epsilon v) - 2f(x) + f(x - \epsilon v)}{\epsilon^2} \right\|,$$

averaged over random directions v. This measures second-order variation and serves as a proxy for deviation from affine behavior. We report curvature of the attention output mean and also evaluate robustness under diagonal precision matrices with condition numbers $2, 4, 8, 12, 20$.

4.1. Results

This section shows how gating affects curvature and performance.

Figure 2 shows curvature versus gate strength. Increasing gate strength consistently raises curvature. Isotropic curves coincide across condition numbers, indicating intrinsic geometry, while anisotropic curvature scales with conditioning but preserves the same trend. Thus, gating controls curvature, while the metric primarily affects its scale.

Figure 3 shows the learned decision boundaries. The ground-truth boundary exhibits a coupled radial–angular nonlinear structure. The ungated model produces approximately piecewise linear regions that deviate from this geometry, while the gated model more closely aligns with the underlying nonlinear structure.

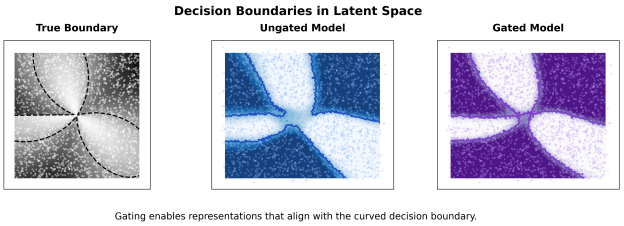

Figure 3. Decision boundaries in latent space on the synthetic curved classification task. The ungated model fails to capture the nonlinear boundary, while the gated model better aligns with the curved structure.

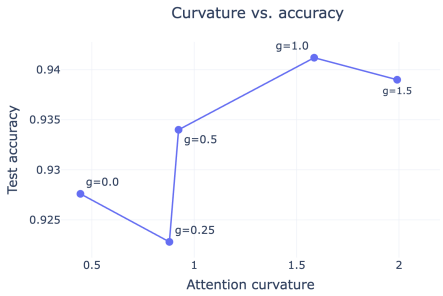

Figure 4. **Curvature correlates with performance.** Test accuracy versus isotropic attention curvature across gate strengths. Accuracy increases with curvature, with mild saturation.

Figure 4 shows a strong positive correlation between curvature and test accuracy ($r \approx 0.79$). Accuracy increases with curvature and saturates at higher values, indicating that greater geometric expressivity improves performance. Results under anisotropic metrics (Appendix) show consistent trends.

Linear control task. To test whether gains arise from generic nonlinearity, we evaluate a control task with a linear decision boundary. Gating provides no consistent advantage (Appendix A.4), indicating its benefits are specific to tasks requiring nonlinear structure.

4.2. Ablation study

Effect of multiplicative gating. We compare attention variants to isolate the role of multiplicative gating. Let Y denote the attention output. The ungated variant uses Y directly. The SiLU variant applies the pointwise nonlinearity

$$Y' = \mathrm{SiLU}(Y), \qquad \mathrm{SiLU}(x) = x \cdot \sigma(x),$$

where σ denotes the sigmoid function. The gated-sigmoid variant applies multiplicative modulation of the form

$$Y' = Y \odot \sigma(WY),$$

and the gated-nonsparse variant uses the same multiplicative structure with a rescaled gate

$$Y' = Y \odot (0.5 + 0.5\,\sigma(WY))\,.$$

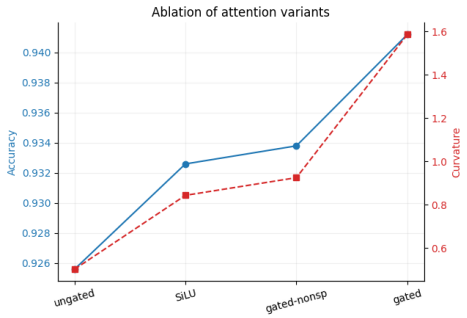

Figure 5. **Ablation of attention variants.** Test accuracy (left axis) and isotropic curvature (right axis) for different attention variants. Ungated attention yields the lowest curvature and accuracy. Adding a pointwise SiLU nonlinearity increases both modestly, while multiplicative gating produces substantially higher curvature and improved accuracy. This shows that gains in geometric expressivity arise specifically from multiplicative gating rather than generic nonlinear transformations.

In the ablation study, all gated variants use gate strength 1 for consistency. Figure 5 shows test accuracy and isotropic curvature across variants. Ungated attention yields the lowest curvature and accuracy, while a SiLU nonlinearity provides only modest improvements. In contrast, multiplicative gating produces substantially higher curvature and improved accuracy, indicating that geometric expressivity arises specifically from gating rather than generic nonlinear transformations. We report isotropic curvature to isolate representation effects, and include anisotropic results in the appendix, which show consistent trends.

These observations are consistent with the geometric interpretation, where gating better captures nonlinear structure.

5. Conclusion

We study attention geometry via curvature and show that multiplicative gating strictly enlarges the class of realizable geometries, enabling nonzero intrinsic curvature unattainable in the ungated setting. We also identify a regime where curvature accumulates under composition, yielding depth amplification. Empirically, gated models exhibit higher curvature than ungated and pointwise nonlinear variants, and isotropic curvature correlates with performance with mild saturation.

Impact Statement

This paper presents theoretical results on the geometric properties of attention mechanisms, with the goal of advancing fundamental understanding in machine learning. The work is purely mathematical and does not involve data collection, deployment, or direct interaction with users.

As such, we do not anticipate immediate societal or ethical risks arising directly from this work. However, like many advances in machine learning theory, improved understanding of model architectures may contribute indirectly to the development of more capable systems, which could have both positive and negative downstream impacts depending on their application.

We encourage future work to consider the broader implications of such advances in applied settings.

References

Amari, S.-i. *Differential-Geometrical Methods in Statistics*. Springer, 1985.

Amari, S.-i. *Information Geometry and Its Applications*. Springer, 2016.

Amari, S.-i. and Nagaoka, H. *Methods of Information Geometry*. American Mathematical Society, 2000.

Amari, S.-i., Karakida, R., and Oizumi, M. Fisher information and natural gradient learning in random deep networks. In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 694–702. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/amari19a.html.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. Learning phrase representations using rnn encoder–decoder for statistical machine translation. *Proceedings of EMNLP*, 2014.

Danihelka, I., Wayne, G., Uria, B., Kalchbrenner, N., and Graves, A. Associative long short-term memory. *Proceedings of ICML*, 2016.

Dauphin, Y. N., Fan, A., Auli, M., and Grangier, D. Language modeling with gated convolutional networks. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 933–941. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/dauphin17a.html.

Kim, M., Li, D., Hu, S. X., and Hospedales, T. Fisher SAM: Information geometry and sharpness aware minimisation. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 11148–11161. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/kim22f.html.

Krishnamurthy, K., Can, T., and Schwab, D. J. Theory of gating in recurrent neural networks. *Phys. Rev. X*, 12:011011, Jan 2022. doi: 10.1103/PhysRevX.12.011011. URL https://link.aps.org/doi/10.1103/PhysRevX.12.011011.

Levine, Y., Wies, N., Sharir, O., Bata, H., and Shashua, A. Limits to depth efficiencies of self-attention. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 22640–22651. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/ff4dfdf5904e920ce52b48c1cef97829-Paper.pdf.

Liang, T., Poggio, T., Rakhlin, A., and Stokes, J. Fisher-rao metric, geometry, and complexity of neural networks. In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 888–896. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/liang19a.html.

Montúfar, G., Pascanu, R., Cho, K., and Bengio, Y. On the number of linear regions of deep neural networks. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/fa6f2a469cc4d61a92d96e74617c3d2a-Paper.pdf.

Pérez, J., Marinković, J., and Barceló, P. On the turing completeness of modern neural network architectures. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HyGBdo0qFm.

Qiu, Z., Wang, Z., Zheng, B., Huang, Z., Wen, K., Yang, S., Men, R., Yu, L., Huang, F., Huang, S., Liu, D.,

Zhou, J., and Lin, J. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=1b7whO4SfY.

Rao, C. R. Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37:81–91, 1945.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Wang, M. and E, W. Understanding the expressive power and mechanisms of transformer for sequence modeling. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=0o7Rd5jngV.

Wu, Y., Zhang, S., Zhang, Y., Bengio, Y., and Salakhutdinov, R. R. On multiplicative integration with recurrent neural networks. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/f69e505b08403ad2298b9f262659929a-Paper.pdf.

Yun, C., Bhojanapalli, S., Rawat, A. S., Reddi, S., and Kumar, S. Are transformers universal approximators of sequence-to-sequence functions? In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ByxRM0Ntvr.

Zhang, Y., Yang, S., Zhu, R., Zhang, Y., Cui, L., Wang, Y., Wang, B., Shi, F., Wang, B., Bi, W., Zhou, P., and Fu, G. Gated slot attention for efficient linear-time sequence modeling. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 116870–116898. Curran Associates, Inc., 2024. doi: 10.52202/079017-3710. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/d3f39e51f5f634fb16cc3e658f8512b9-Paper-Conference.pdf.

A. Appendix

A.1. Proofs of Main Results

A.1.1 Proof of Theorem 3.1

Proof. The attention output takes the form

$$\mu = \sum_{i=1}^{n} \alpha_i U_i, \quad \alpha \in \Delta_{n-1},$$

so the set of outputs lies in the convex hull of $\{U_1, \dots, U_n\}$.

On the relative interior, we can eliminate one coordinate using the constraint $\sum_i \alpha_i = 1$. Writing

$$\alpha_n = 1 - \sum_{i=1}^{n-1} \alpha_i,$$

we obtain

$$\mu = \sum_{i=1}^{n-1} \alpha_i U_i + \left(1 - \sum_{i=1}^{n-1} \alpha_i\right) U_n = U_n + \sum_{i=1}^{n-1} \alpha_i (U_i - U_n).$$

Let

$$k = \dim \operatorname{aff}\{U_1, \dots, U_n\}.$$

On the relative interior of the convex hull, choose local affine coordinates

$$\phi \in \Omega \subset \mathbb{R}^k,$$

under which the map admits an affine parameterization

$$\mu_{\text{ung}}(\phi) = a + B\phi,$$

where $a \in \mathbb{R}^D$ and $B \in \mathbb{R}^{D \times k}$ has full column rank k. Its partial derivatives are constant:

$$\partial_i \mu_{\text{ung}}(\phi) = B_i \in \mathbb{R}^D,$$

where B_i denotes the i-th column of B.

Equipping the Gaussian family $\mathcal{N}(\mu, I_D)$ with the Fisher–Rao metric, the induced Riemannian metric on the parameter domain Ω is

$$g_{ij}^{\text{ung}}(\phi) = \langle \partial_i \mu_{\text{ung}}(\phi), \partial_j \mu_{\text{ung}}(\phi) \rangle_{\mathbb{R}^D} = \langle B_i, B_j \rangle_{\mathbb{R}^D} =: C_{ij}.$$

Since $B \in \mathbb{R}^{D \times k}$ has full column rank, the matrix

$$C = B^\top B$$

is symmetric positive definite. Hence, the metric tensor is independent of ϕ.

For a constant metric tensor in global coordinates, the Christoffel symbols of the Levi–Civita connection vanish identically:

$$\Gamma_{ij}^k = \tfrac{1}{2} g^{k\ell} \big(\partial_i g_{j\ell} + \partial_j g_{i\ell} - \partial_\ell g_{ij} \big) = 0.$$

Consequently, the Riemann curvature tensor,

$$R^k_{\ell ij} = \partial_i \Gamma^k_{\ell j} - \partial_j \Gamma^k_{\ell i} + \Gamma^k_{im} \Gamma^m_{\ell j} - \Gamma^k_{jm} \Gamma^m_{\ell i},$$

vanishes identically. Therefore, all sectional curvatures are zero, and the statistical manifold $\mathcal{M}_{\text{ung}}$ is intrinsically flat. □

A.1.2 Proof of Theorem 3.2

Proof. By Theorem 3.1, the manifold induced by any ungated attention representation is intrinsically flat. Hence

$$K_{F_{\text{ung}}}(u, \Pi) = 0 \quad \forall u \in U,\ \forall \Pi \subset T_u U,\ \dim(\Pi) = 2.$$

By assumption,

$$K_{F^\star}(u, \Pi) \geq \kappa_0 > 0 \quad \forall u \in U,\ \forall \Pi \subset T_u U,\ \dim(\Pi) = 2.$$

Therefore,

$$\left|K_{F_{\text{ung}}}(u, \Pi) - K_{F^\star}(u, \Pi)\right| = |0 - K_{F^\star}(u, \Pi)| = K_{F^\star}(u, \Pi) \geq \kappa_0$$

for all $u \in U$ and all 2-planes $\Pi \subset T_u U$.

Taking nested suprema over u and Π yields the result. □

A.1.3 Proof of Lemma 3.5

Proof. Differentiation gives the stated formula. Define

$$T : C^2(U, \mathbb{R}^D) \to C^0(U, \mathbb{R}^{D \times d \times d}), \qquad T(g) = D^2(Y \odot g).$$

Then T is continuous and linear, and $\mathcal{A} = \ker T$ since μ is affine if and only if $D^2\mu \equiv 0$ on connected U. Thus $\mathcal{A}$ is a closed linear subspace.

To see $\mathcal{A} \neq C^2(U, \mathbb{R}^D)$, note that since $\operatorname{rank}(B) = d \geq 1$, we have $B \neq 0$, hence $Y(\phi) = a + B\phi$ is not identically zero. Therefore there exist $m \in \{1, \ldots, D\}$ and $\phi_0 \in U$ such that $Y_m(\phi_0) \neq 0$. Take $g = e_m \psi$ with $\psi \in C_c^\infty(U)$ satisfying

$$\psi(\phi_0) = 0, \qquad \partial_i \psi(\phi_0) = 0, \qquad \partial_{11} \psi(\phi_0) = 1.$$

Then

$$(Tg)_{11,m}(\phi_0) = Y_m(\phi_0)\, \partial_{11}\psi(\phi_0) = Y_m(\phi_0) \neq 0,$$

so $g \notin \ker T$. Hence $\mathcal{A}$ is a proper closed linear subspace, and therefore nowhere dense and meagre. □

A.1.4 Proof of Theorem 3.6

Proof. The proof gives an explicit witness inside a standard content-aware transformer block.

Let

$$U := \left(0, \frac{\pi}{4}\right) \times \left(0, \frac{\pi}{4}\right) \subset \mathbb{R}^2, \qquad \phi = (\phi_1, \phi_2).$$

Define

$$s(\phi_1, \phi_2) = \begin{pmatrix} \cos\phi_1 \cos\phi_2 \\ \cos\phi_1 \sin\phi_2 \\ \sin\phi_1 \end{pmatrix}.$$

Since $\phi_1, \phi_2 \in (0, \pi/4)$, all three coordinates of $s(\phi)$ are strictly positive, so

$$s(\phi) \in (0, 1)^3 \qquad \text{for all } \phi \in U.$$

This is the standard local parametrization of a patch of the unit sphere $S^2 \subset \mathbb{R}^3$, and its induced Gaussian curvature is identically 1.

Step 1: Enforce affine structure in the ungated output. Set

$$W_Q = 0, \qquad W_K = 0.$$

Note that content-dependent attention mechanisms strictly contain this uniform-weight case as a special instance. Therefore, any geometric expressivity achievable under this construction is realizable within the full content-aware attention class.

Then $Q(X) = 0$ and $K(X) = 0$ for every input X, so all attention scores are equal and the softmax produces uniform weights

$$\alpha_{ji}(X) = \frac{1}{n} \qquad \text{for all } X,\ i = 1, \dots, n.$$

Choose W_V and W_O so that the composed linear map

$$L := W_V W_O$$

has rank 3. Fix a vector $a \in \mathbb{R}^3$ and a matrix $B \in \mathbb{R}^{3\times 2}$ of rank 2, chosen so that

$$0 < \frac{s_k(\phi)}{(a + B\phi)_k} < 1 \qquad \text{for all } \phi \in U,\ k = 1, 2, 3.$$

Such a choice is possible because $s(\phi)$ is bounded on the relatively compact set $\overline{U}$, so choosing a sufficiently large componentwise ensures

$$(a + B\phi)_k > s_k(\phi) > 0 \qquad \text{for all } \phi \in U,\ k = 1, 2, 3.$$

Now define

$$U_1(X(\phi)) := n(a + B\phi), \qquad U_i(X(\phi)) := 0 \quad \text{for } i = 2, \dots, n.$$

Because L has rank 3, the input token embeddings $X(\phi)$ can be chosen smoothly so that these value vectors are realized exactly. With these choices,

$$Y_j(X(\phi)) = \sum_{i=1}^{n} \alpha_{ji}(X(\phi)) U_i(X(\phi)) = \frac{1}{n} U_1(X(\phi)) = a + B\phi.$$

Hence the ungated map $\phi \mapsto Y_j(X(\phi))$ is affine. Since the decoder is the Gaussian location family with covariance I_3, the Fisher–Rao metric on mean space is Euclidean, and therefore the induced manifold

$$\mathcal{M}_{\text{ung}} = \{\mathcal{N}(Y_j(X(\phi)), I_3) : \phi \in U\}$$

is isometric to an open subset of Euclidean space. Thus

$$K_{\text{ung}}(\phi) = 0 \qquad \text{for all } \phi \in U.$$

Step 2: Construct an exact multiplicative gate realizing a sphere. Define the componentwise ratio

$$\tilde{z}(\phi) := \frac{s(\phi)}{a + B\phi},$$

where the division is taken coordinatewise. By the choice of a and B,

$$\tilde{z}(\phi) \in (0, 1)^3 \qquad \text{for all } \phi \in U.$$

Let $\sigma^{-1} : (0, 1) \to \mathbb{R}$ denote the inverse of the scalar nonlinearity, applied coordinatewise, and define

$$t(\phi) := \sigma^{-1}(\tilde{z}(\phi)) \in \mathbb{R}^3.$$

Then t is smooth and satisfies

$$\sigma(t(\phi)) = \tilde{z}(\phi) \qquad \text{for all } \phi \in U.$$

Choose any matrix

$$W_\theta \in \mathbb{R}^{m\times 3} \qquad \text{with } \operatorname{rank}(W_\theta) = 3.$$

Since $W_\theta \in \mathbb{R}^{m\times 3}$ has full column rank, it admits a left inverse $W_\theta^\dagger \in \mathbb{R}^{3\times m}$ such that

$$W_\theta^\dagger W_\theta = I_3.$$

Let $t(\phi) \in \mathbb{R}^3$ and view $t(\phi)^\top \in \mathbb{R}^{1\times 3}$. Define

$$X_g(\phi) := t(\phi)^\top W_\theta^\dagger \in \mathbb{R}^{1\times m}.$$

Then

$$X_g(\phi)W_\theta = t(\phi)^\top W_\theta^\dagger W_\theta = t(\phi)^\top,$$

and therefore

$$\sigma(X_g(\phi)W_\theta) = \sigma(t(\phi)) = \tilde{z}(\phi)$$

holds exactly. Consequently,

$$Y_j'(X(\phi)) = Y_j(X(\phi)) \odot \sigma(X_g(\phi)W_\theta) = (a + B\phi) \odot \tilde{z}(\phi) = s(\phi).$$

Thus the gated output coincides exactly with the spherical parametrization $s(\phi)$.

Step 3: Deduce curvature. Since $s(\phi)$ parametrizes a smooth patch of the unit sphere S^2, the induced Euclidean metric on its image is the spherical metric, whose Gaussian curvature is identically 1. Because the Fisher–Rao metric for the Gaussian location family coincides with the Euclidean pullback metric on the mean manifold, the induced gated statistical manifold

$$\mathcal{M}_{\text{gat}} = \{\mathcal{N}(Y_j'(X(\phi)), I_3) : \phi \in U\}$$

is isometric to a spherical patch and satisfies

$$K_{\text{gat}}(\phi) = 1 \qquad \text{for all } \phi \in U.$$

Combining the flatness of $\mathcal{M}_{\text{ung}}$ with the positive curvature of $\mathcal{M}_{\text{gat}}$ yields the claimed curvature gap inside standard content-aware attention. The construction above is an explicit witness that multiplicative gating strictly enlarges the class of realizable representation geometries in this architectural family. □

A.1.5 Proof of Theorem 3.9

Proof. By Theorem 3.1, all sectional curvatures vanish identically on any manifold induced by a single ungated attention block.

By Theorem 3.6, there exists $\mathcal{M} \in \mathcal{G}_{\text{gat}}$ whose sectional curvature is strictly positive on an open set.

Since sectional curvature is invariant under local isometry, $\mathcal{M}$ cannot be locally isometric to any $\mathcal{N} \in \mathcal{G}_{\text{ung}}$. The result follows. □

A.1.6 Proof of Lemma 3.10

Proof. By linearity of summation,

$$\sum_{\ell=1}^{L} a_\ell \psi(u,v) = \Big(\sum_{\ell=1}^{L} a_\ell\Big)\psi(u,v) = A_L\psi(u,v).$$

Substituting this identity into the chart expression yields

$$\chi_{\text{out}} \circ F_L \circ \chi_{\text{in}}^{-1}(u,v) = \big(u,\ v,\ A_L\psi(u,v),\ 0,\dots,0\big).$$

This proves the claim. □

A.1.7 Proof of Lemma 3.11

Proof. We construct the model explicitly.

Fix an ambient dimension $D \geq 4$. Define the initial state

$$h^{(0)}(u, v) = (u, v, 0, 1, 0, \ldots, 0) \in \mathbb{R}^D,$$

and define the input parametrization by

$$X(u, v) = h^{(0)}(u, v).$$

The fourth coordinate is a constant auxiliary feature.

Attention mechanism. We instantiate each attention block on a single-token input, so the softmax weight is identically equal to 1. This is a valid special case of the standard attention mechanism.

Choose value and output projections so that the ungated attention output is exactly

$$Y(u, v) = e_3 = (0, 0, 1, 0, \ldots, 0).$$

This is achieved by letting the value/output map read only the constant fourth coordinate and send it to the third coordinate.

Gate construction. Define

$$X_g^{(\ell)}(u, v) = \mathrm{logit}(a_\ell \psi(u, v))\, e_3, \qquad W_\theta^{(\ell)} = I_D.$$

Then

$$X_g^{(\ell)}(u, v) W_\theta^{(\ell)} = \mathrm{logit}(a_\ell \psi(u, v))\, e_3.$$

Applying σ componentwise yields

$$\sigma\big(X_g^{(\ell)} W_\theta^{(\ell)}\big) = \big(\tfrac{1}{2}, \tfrac{1}{2}, a_\ell \psi(u, v), \tfrac{1}{2}, \ldots, \tfrac{1}{2}\big).$$

Gated output. Since $Y(u, v) = e_3$, we obtain

$$Y'(u, v) = e_3 \odot \sigma(\cdot) = a_\ell \psi(u, v)\, e_3.$$

Residual accumulation. Define

$$h^{(\ell)} = h^{(\ell-1)} + a_\ell \psi(u, v)\, e_3.$$

By induction,

$$h^{(L)}(u, v) = \Big(u,\ v,\ \sum_{\ell=1}^{L} a_\ell \psi(u, v),\ 1,\ 0, \ldots, 0\Big).$$

Final output map. Define a linear projection $P : \mathbb{R}^D \to \mathbb{R}^D$ by

$$P(x_1, x_2, x_3, x_4, x_5, \ldots, x_D) = (x_1, x_2, x_3, 0, \ldots, 0).$$

Then

$$P\big(h^{(L)}(u, v)\big) = \Big(u,\ v,\ \sum_{\ell=1}^{L} a_\ell \psi(u, v),\ 0, \ldots, 0\Big),$$

which is the claimed normal form. □

A.1.8 Proof of Lemma 3.12

Proof. Because the last $D-3$ coordinates are constant, the intrinsic geometry of M_L is the same as that of the graph surface

$$X(u,v) = (u, v, f_L(u,v)) \subset \mathbb{R}^3, \qquad f_L(u,v) := A_L \psi(u,v).$$

Hence it suffices to compute the Gaussian curvature of the graph surface X.

The tangent vectors are

$$X_u = (1, 0, f_u), \qquad X_v = (0, 1, f_v).$$

Thus the coefficients of the first fundamental form are

$$E = \langle X_u, X_u \rangle = 1 + f_u^2, \qquad F = \langle X_u, X_v \rangle = f_u f_v, \qquad G = \langle X_v, X_v \rangle = 1 + f_v^2.$$

A unit normal vector is

$$N = \frac{(-f_u, -f_v, 1)}{\sqrt{1 + f_u^2 + f_v^2}}.$$

The second derivatives are

$$X_{uu} = (0, 0, f_{uu}), \qquad X_{uv} = (0, 0, f_{uv}), \qquad X_{vv} = (0, 0, f_{vv}),$$

so the coefficients of the second fundamental form are

$$e = \langle X_{uu}, N \rangle = \frac{f_{uu}}{\sqrt{1 + f_u^2 + f_v^2}},$$

$$f = \langle X_{uv}, N \rangle = \frac{f_{uv}}{\sqrt{1 + f_u^2 + f_v^2}},$$

$$g = \langle X_{vv}, N \rangle = \frac{f_{vv}}{\sqrt{1 + f_u^2 + f_v^2}}.$$

Therefore,

$$K = \frac{eg - f^2}{EG - F^2} = \frac{f_{uu} f_{vv} - f_{uv}^2}{\big(1 + f_u^2 + f_v^2\big)^2}.$$

Now evaluate at the origin. Since $\nabla \psi(0) = 0$,

$$\nabla f_L(0) = A_L \nabla \psi(0) = 0, \qquad D^2 f_L(0) = A_L D^2 \psi(0).$$

Hence

$$K_{M_L}\big(F_L(0)\big) = \det D^2 f_L(0).$$

Using the scaling of the determinant for a 2×2 matrix,

$$\det D^2 f_L(0) = \det\big(A_L D^2 \psi(0)\big) = A_L^2 \det D^2 \psi(0).$$

Thus

$$K_{M_L}\big(F_L(0)\big) = A_L^2 \ \det D^2 \psi(0).$$

□

A.1.9 Proof of Theorem 3.13

Proof. By the local gated-attention normal form lemma,

$$\chi_{\text{out}} \circ F_L \circ \chi_{\text{in}}^{-1}(u, v) = \big(u,\ v,\ A_L\psi(u, v),\ 0, \dots, 0\big).$$

Therefore the graph-curvature lemma applies and gives

$$K_{M_L}\big(F_L(0)\big) = A_L^2\ \det D^2\psi(0).$$

If $a_\ell \geq a_0 > 0$ for all ℓ, then

$$A_L = \sum_{\ell=1}^{L} a_\ell \geq a_0 L.$$

Taking absolute values yields

$$\big|K_{M_L}\big(F_L(0)\big)\big| = A_L^2\ \big|\det D^2\psi(0)\big| \geq a_0^2\ \big|\det D^2\psi(0)\big|\ L^2.$$

If furthermore $\det D^2\psi(0) > 0$, then the absolute values may be dropped and we obtain

$$K_{M_L}\big(F_L(0)\big) \geq a_0^2\ \det D^2\psi(0)\, L^2.$$

□

A.1.10 Proof of Theorem 3.15

Proof. Write

$$\Phi_L^{\text{tot}} = (f^1, \dots, f^{D-2}),$$

so that

$$F_L(u, v) = (u, v, f^1(u, v), \dots, f^{D-2}(u, v)).$$

First derivatives are

$$F_u = (1, 0, f_u^1, \dots, f_u^{D-2}), \quad F_v = (0, 1, f_v^1, \dots, f_v^{D-2}).$$

Since $D\Phi_L^{\text{tot}}(0) = 0$, we have $f_u^\alpha(0) = f_v^\alpha(0) = 0$, so

$$F_u(0) = (1, 0, 0, \dots, 0), \quad F_v(0) = (0, 1, 0, \dots, 0).$$

Thus the induced metric satisfies $g_{ij}(0) = \delta_{ij}$.

The normal space at $F_L(0)$ is spanned by

$$n_\alpha = e_{2+\alpha}, \quad \alpha = 1, \dots, D-2.$$

Second derivatives are

$$F_{uu} = (0, 0, f_{uu}^1, \dots, f_{uu}^{D-2}), \quad F_{uv} = (0, 0, f_{uv}^1, \dots, f_{uv}^{D-2}), \quad F_{vv} = (0, 0, f_{vv}^1, \dots, f_{vv}^{D-2}).$$

Thus the second fundamental form in direction n_α is

$$II^{(\alpha)} = \begin{pmatrix} f_{uu}^\alpha(0) & f_{uv}^\alpha(0) \\ f_{uv}^\alpha(0) & f_{vv}^\alpha(0) \end{pmatrix} = D^2 f^\alpha(0).$$

At the base point, the Gaussian curvature is given by the Gauss equation

$$K = \sum_{\alpha=1}^{D-2} \big(h_{11}^\alpha h_{22}^\alpha - (h_{12}^\alpha)^2\big) = \sum_{\alpha=1}^{D-2} \det\big(II^{(\alpha)}\big),$$

where $h_{ij}^{\alpha} = \langle F_{ij}, n_\alpha \rangle$.

Substituting $II^{(\alpha)} = D^2 f^\alpha(0)$ gives

$$K_{M_L}\big(F_L(0)\big) = \sum_{\alpha=1}^{D-2} \det\Big(D^2(\Phi_L^{\mathrm{tot}})^\alpha(0)\Big).$$

Linearity of the Hessian yields

$$D^2(\Phi_L^{\mathrm{tot}})^\alpha(0) = \sum_{\ell=1}^{L} D^2\Phi_\ell^\alpha(0),$$

which gives the stated formula. □

A.1.11 Proof of corollary 3.17

Proof. We have

$$\Phi_L^{\mathrm{tot}}(u,v) = \sum_{\ell=1}^{L} a_\ell \psi(u,v)c = A_L\psi(u,v)c.$$

Thus

$$D^2(\Phi_L^{\mathrm{tot}})^\alpha(0) = A_L c_\alpha D^2\psi(0).$$

Using $\det(\lambda H) = \lambda^2 \det(H)$ for 2×2 matrices,

$$\det\big(D^2(\Phi_L^{\mathrm{tot}})^\alpha(0)\big) = A_L^2 c_\alpha^2 \det D^2\psi(0).$$

Summing over α gives

$$K_{M_L}\big(F_L(0)\big) = A_L^2\|c\|^2 \det D^2\psi(0).$$

Since $a_\ell \geq a_0 > 0$, we have $A_L \geq a_0 L$, hence

$$A_L^2 \geq a_0^2 L^2,$$

which gives the bound. □

A.1.12 Proof of Lemma 3.19

Proof. The map $(\phi, W_\theta) \mapsto \mu_{W_\theta}(\phi)$ is smooth, hence $g_{ij}(\phi; W_\theta)$ depends continuously on (ϕ, W_θ). Therefore $\det g$ is continuous and R is open.

Since $g = (D\mu_{W_\theta})^\top D\mu_{W_\theta}$, positivity implies that $D\mu_{W_\theta}$ has rank 2, so μ_{W_θ} is an immersion and curvature is well-defined.

Gaussian curvature depends smoothly on the metric and its derivatives, hence is continuous on R. □

A.1.13 Proof of Lemma 3.20

Proof. Since $(\phi, W_\theta) \mapsto K(\phi; W_\theta)$ is continuous on R (Lemma 3.19), and

$$K(\phi; W_\theta^*) \geq c > 0 \quad \forall \phi \in K,$$

it follows that

$$\inf_{\phi\in K} K(\phi; W_\theta^*) \geq c.$$

By continuity in both variables, $K(\phi; W_\theta)$ is uniformly continuous on a neighborhood of the compact set $K \times \{W_\theta^*\}$. Hence there exists an open neighborhood $\mathcal{O}$ of W_θ^* such that

$$|K(\phi; W_\theta) - K(\phi; W_\theta^*)| < \frac{c}{2} \quad \forall \phi \in K,\ W_\theta \in \mathcal{O}.$$

Therefore

$$K(\phi; W_\theta) \geq \frac{c}{2} > 0.$$

Since R is open and contains (ϕ, W_θ^*), possibly shrinking $\mathcal{O}$ ensures $(\phi, W_\theta) \in R$ for all $\phi \in K$, completing the proof. □

A.1.14 Proof of Theorem 3.18

Proof. For each W_θ, the map $\mu_{W_\theta} : U \to \mathbb{R}^D$ is smooth since Y, X^*, and σ are smooth. Hence

$$\phi \mapsto \mathcal{N}(\mu_{W_\theta}(\phi), I_D)$$

defines a smooth Gaussian location family. On any open set where $D\mu_{W_\theta}$ has rank 2, the map is an immersion. The Fisher–Rao metric reduces to the Euclidean pullback,

$$g_{ij}(\phi; W_\theta) = \langle \partial_i \mu_{W_\theta}(\phi), \partial_j \mu_{W_\theta}(\phi) \rangle.$$

(i) Flatness. If the gate is identically 1, then

$$\mu_{W_\theta}(\phi) = Y(\phi) = a + B\phi.$$

Thus $\partial_i \mu_{W_\theta} = Be_i$ is constant, so the metric is constant. Hence all Christoffel symbols vanish and the Gaussian curvature is identically zero.

(ii) Positive curvature realization. Let $s : U \to S^2 \subset \mathbb{R}^3$ be a smooth spherical patch. Define

$$\tilde{s}(\phi) = (s_1(\phi), s_2(\phi), s_3(\phi), c_4, \ldots, c_D),$$

with constants $c_4, \ldots, c_D > 0$. Since the added coordinates are constant, the first fundamental form is unchanged, so $\tilde{s}$ has curvature identically 1.

After possibly shrinking U, we may assume $\tilde{s}$ is bounded. Choose the affine map $Y(\phi) = a + B\phi$ so that

$$0 < \tilde{s}_j(\phi) < Y_j(\phi) \qquad \forall \phi \in U,\ j = 1, \ldots, D.$$

Define

$$z(\phi) = \tilde{s}(\phi) \oslash Y(\phi) \in (0,1)^D, \qquad \ell(\phi) = \sigma^{-1}(z(\phi)).$$

Set

$$X^*(\phi) = (\ell(\phi), 0, \ldots, 0), \qquad W_\theta^* = \begin{bmatrix} I_D \\ 0 \end{bmatrix}.$$

Then $X^*(\phi) W_\theta^* = \ell(\phi)$, hence

$$\mu_{W_\theta^*}(\phi) = Y(\phi) \odot \sigma(\ell(\phi)) = \tilde{s}(\phi).$$

Thus $K(\phi; W_\theta^*) \equiv 1$.

(iii) Local robustness. Fix a nonempty compact set $K \subset U$. Since $K(\phi; W_\theta^*) = 1$ for all $\phi \in K$, the result follows from Lemma 3.20. □

A.1.15 Proof of theorem 3.21

Proof. **Openness.** Let $g \in \mathcal{N}_{\phi_0}$. Since $g \in \mathcal{I}_{\phi_0}$, the induced metric is nondegenerate at ϕ_0. On $\mathcal{I}_{\phi_0}$, the curvature tensor at ϕ_0 depends smoothly on the 2-jet of μ_g, hence continuously on g in the C^2 topology. Therefore the condition $R_g(\phi_0) \neq 0$ is stable under sufficiently small perturbations, so $\mathcal{N}_{\phi_0}$ is open.

Density. Fix $g \in \mathcal{I}_{\phi_0}$ and set $F := \mu_g$. Let

$$T := \operatorname{Im} DF(\phi_0).$$

Let

$$S := \{j : Y_j(\phi_0) \neq 0\}, \quad E := \operatorname{span}\{e_j : j \in S\}.$$

Since $\dim E \geq d+1$ and $\dim T = d$, we have

$$\dim(E \cap T^\perp) \geq \dim E + \dim T^\perp - D \geq 1.$$

Choose a unit vector $n \in E \cap T^\perp$.

Since $Y_j(\phi_0) \neq 0$ for $j \in S$, by continuity there exists a neighborhood $V \ni \phi_0$ such that $Y_j(\phi) \neq 0$ on V for all $j \in S$. Choose $\chi \in C_c^\infty(V)$ with $\chi \equiv 1$ near ϕ_0.

Define

$$q_\varepsilon(\phi) := \frac{\varepsilon}{2}\|\phi - \phi_0\|^2.$$

Define

$$(h_\varepsilon)_j(\phi) = \begin{cases} \chi(\phi)\, q_\varepsilon(\phi)\, \dfrac{n_j}{Y_j(\phi)}, & j \in S, \\ 0, & j \notin S. \end{cases}$$

Set $\tilde{g}_\varepsilon = g + h_\varepsilon$, $\tilde{F}_\varepsilon = \mu_{\tilde{g}_\varepsilon}$.

Then

$$\tilde{F}_\varepsilon = F + \chi q_\varepsilon n.$$

Since q_ε vanishes to first order at ϕ_0, we have

$$D(\chi q_\varepsilon n)(\phi_0) = 0,$$

so

$$D\tilde{F}_\varepsilon(\phi_0) = DF(\phi_0).$$

Thus $\tilde{g}_\varepsilon \in \mathcal{I}_{\phi_0}$.

Second fundamental form. At ϕ_0, choose normal coordinates on the parameter domain and orthonormal bases of T and $T^\perp$, with $n = n^1$. In these coordinates, the Christoffel symbols vanish at ϕ_0, so

$$B_{ij}^\alpha = \langle \partial_{ij} F(\phi_0), n^\alpha \rangle.$$

Since $\chi \equiv 1$ near ϕ_0,

$$D^2\tilde{F}_\varepsilon(\phi_0) = D^2 F(\phi_0) + \varepsilon I_d \otimes n.$$

Thus

$$\tilde{B}_{ij}^1 = B_{ij}^1 + \varepsilon\delta_{ij}, \quad \tilde{B}_{ij}^\alpha = B_{ij}^\alpha \ (\alpha \geq 2).$$

Curvature computation. By the Gauss equation,

$$R_{ijkl} = \sum_\alpha (B_{ik}^\alpha B_{jl}^\alpha - B_{il}^\alpha B_{jk}^\alpha).$$

Fix any $i \neq j$ (which exists since $d \geq 2$). Then

$$\tilde{R}_{ijij}(\varepsilon) = (B_{ii}^1 + \varepsilon)(B_{jj}^1 + \varepsilon) - (B_{ij}^1)^2 + \sum_{\alpha \geq 2} (B_{ii}^\alpha B_{jj}^\alpha - (B_{ij}^\alpha)^2).$$

Hence

$$\tilde{R}_{ijij}(\varepsilon) = R_{ijij}(0) + \varepsilon(B_{ii}^1 + B_{jj}^1) + \varepsilon^2.$$

Since n is a unit vector, the coefficient of the ε^2 term is exactly 1. Thus this is a nonzero polynomial in ε, so it has only finitely many roots.

Therefore, for all sufficiently small ε outside this finite set,

$$\tilde{R}_{ijij}(\varepsilon) \neq 0.$$

Since a single nonzero curvature component implies $R_{\tilde{g}_\varepsilon}(\phi_0) \neq 0$, we obtain $\tilde{g}_\varepsilon \in \mathcal{N}_{\phi_0}$.

Since $h_\varepsilon \to 0$ in C^2, every neighborhood of g contains such perturbations. Thus $\mathcal{N}_{\phi_0}$ is dense in $\mathcal{I}_{\phi_0}$. □

A.2. Additional analysis of anisotropy under varying conditioning.

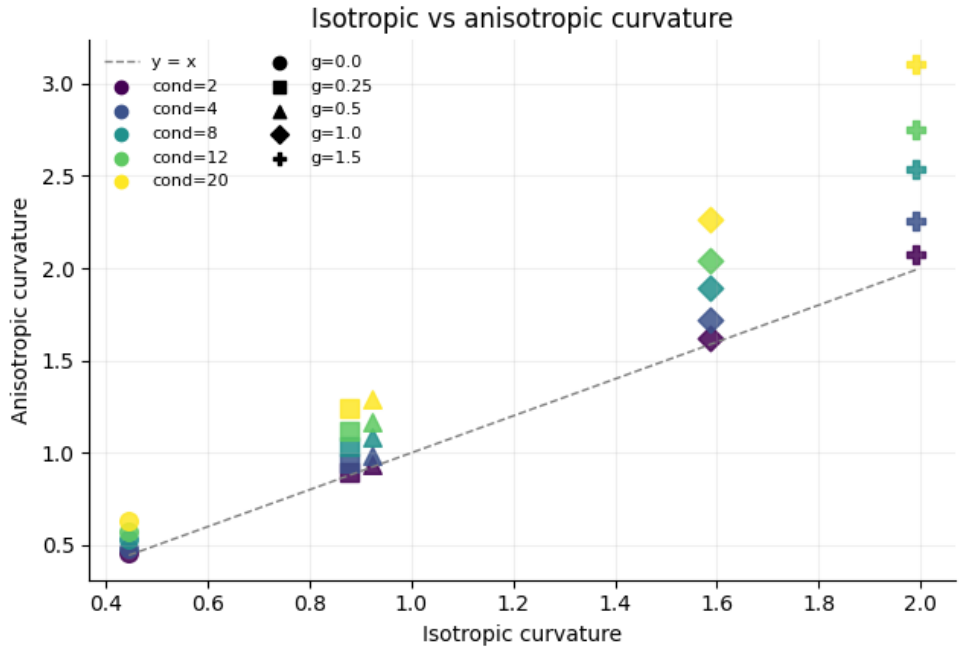

Figure 6. **Isotropic vs anisotropic curvature.** Points correspond to different gate strengths and condition numbers, with marker shape indicating gate strength and color indicating condition number. The two curvature measures are nearly perfectly correlated, while anisotropic curvature differs in scale due to metric effects.

Curvature is intrinsic to the representation mapping. Figure 6 compares curvature measured under isotropic and anisotropic metrics across gate strengths and condition numbers. Each point corresponds to a specific combination of gate strength and condition number. We observe that isotropic and anisotropic curvature are nearly perfectly correlated across all settings. While anisotropic curvature increases with the condition number due to metric scaling, the relative ordering induced by gate strength is preserved. This shows that curvature is primarily determined by the representation mapping, while the choice of metric mainly affects its scale, consistent with our theoretical analysis.

A.3. Ablation under anisotropic metrics

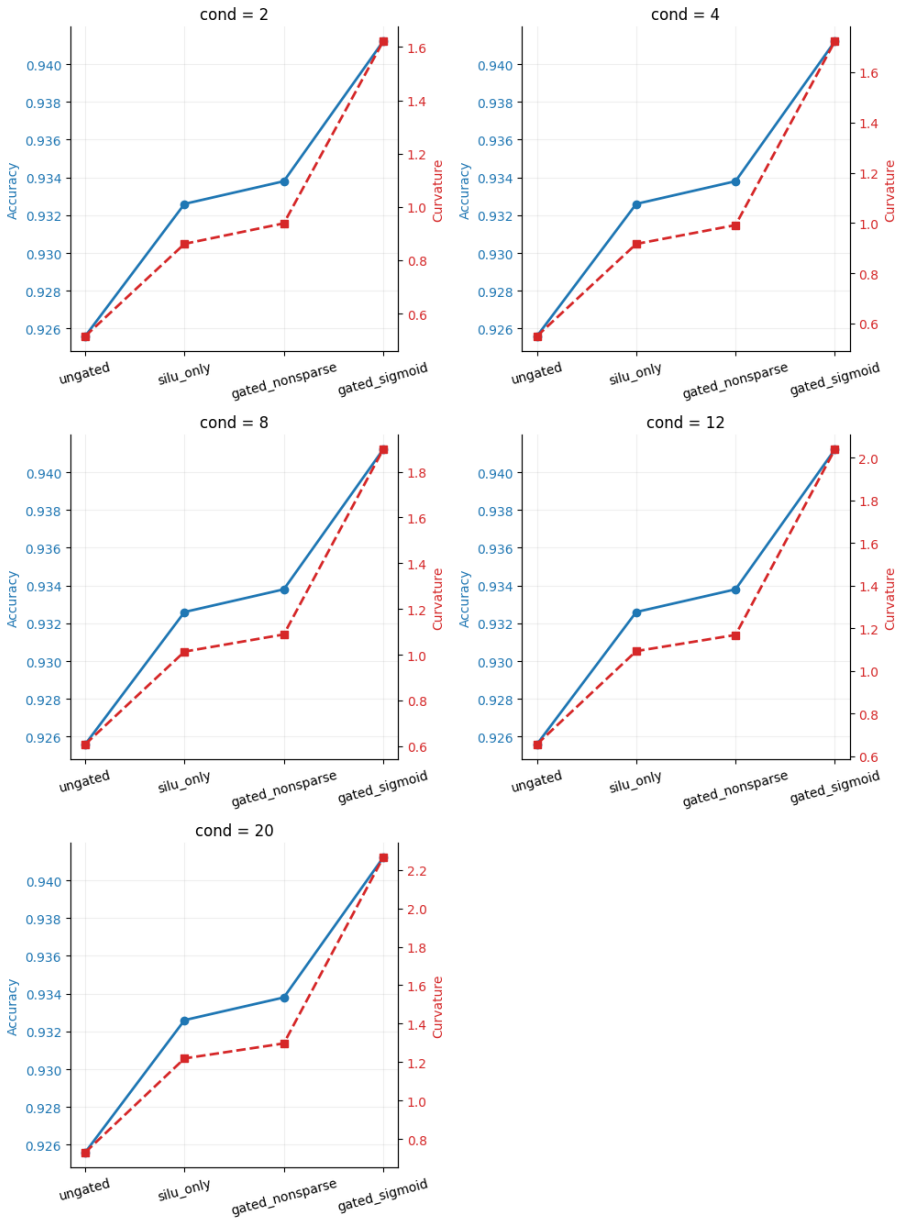

Figure 7. **Ablation under anisotropic metrics.** Each subplot corresponds to a different condition number. While curvature increases with the condition number, the relative ordering of variants and their accuracy remain unchanged.

We repeat the ablation analysis under anisotropic metrics with varying condition numbers. Figure 7 shows test accuracy and anisotropic curvature for different attention variants across condition numbers.

We observe that the qualitative trends remain unchanged. Multiplicative gating consistently yields higher curvature and improved accuracy, while ungated attention and pointwise nonlinearities exhibit lower curvature and weaker performance. Increasing the condition number rescales the magnitude of curvature but does not affect accuracy or the relative ordering of variants. These results further support that geometric expressivity is driven by the representation mapping rather than the choice of metric.

A.4. Linear Control Task

To assess whether the observed gains arise from generic additional nonlinearity, we consider a control task in which the decision boundary is linear and does not require curved representations. As in the main experiment, each example is generated by sampling a latent center $c \in [-2, 2]^2$ and forming a sequence of noisy observations. The label is determined by a linear function of the latent center,

$$y = \mathbf{1}(w^\top c > 0),$$

for a fixed vector $w \in \mathbb{R}^2$.

Table 1 reports results across different gating strengths. We observe that all variants achieve similar accuracy, with differences well within the standard deviation across seeds. In particular, accuracy remains nearly constant despite substantial variation in representation curvature.

Figure 8 shows that increasing gate strength leads to a clear increase in attention curvature under both isotropic and anisotropic metrics (with condition number 12). However, as shown in Figure 9, these increases in curvature do not translate into improved performance. The relationship between curvature and accuracy is not consistently positive and is weakly negative in this setting.

Taken together, these results indicate that increased geometric expressivity does not provide a performance benefit when the task does not require nonlinear structure. This supports the interpretation that the gains observed in the main experiment arise specifically from the ability of gated attention to realize curved representations, rather than from generic additional nonlinearity.

Table 1. Linear control task. Mean ± std over seeds.

Gate strength	Accuracy	Attn curvature (iso)	Attn curvature (aniso cond = 12)
0.00	0.9656 ± 0.0081	0.9883	1.2919
0.25	0.9664 ± 0.0079	0.9624	1.2695
0.50	0.9654 ± 0.0081	0.9434	1.2598
1.00	0.9662 ± 0.0082	1.0920	1.4584
1.50	0.9636 ± 0.0088	1.3995	1.8438

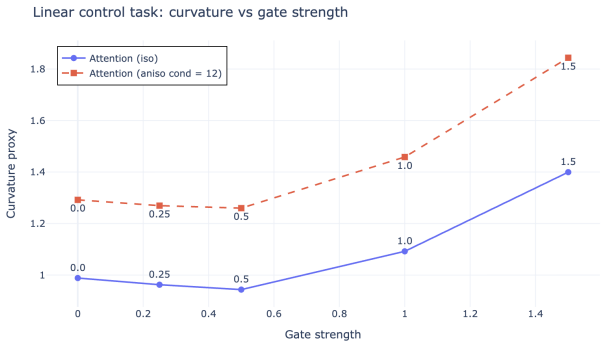

Figure 8. Linear control task. Attention curvature as a function of gate strength under isotropic and anisotropic metrics (condition number 12). While curvature increases overall with gate strength, the relationship is not strictly monotonic, reflecting the absence of task pressure for nonlinear structure.

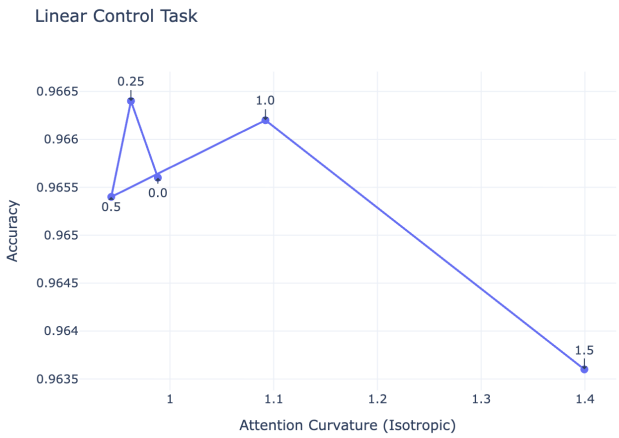

Figure 9. Linear control task. Accuracy as a function of isotropic attention curvature across gate strengths. No consistent positive relationship is observed, indicating that increased curvature does not improve performance when the task does not require nonlinear structure.

We note that, unlike in the curved task, curvature does not increase monotonically with gate strength in the linear control setting. This suggests that, in the absence of task pressure for nonlinear structure, curvature is not systematically reinforced during training.

A.5. Experimental Details

A.5.1. Model Architecture

We use a minimal attention-based model consisting of a single attention block followed by a classifier head. Inputs in $\mathbb{R}^2$ are projected to a hidden dimension of 64 and arranged into sequences of length 8. The projected inputs are processed by scaled dot-product attention, followed by mean pooling and a two-layer MLP classifier.

We consider multiple variants of the attention output: (i) ungated attention, (ii) a pointwise SiLU nonlinearity applied to the attention output, and (iii) multiplicative gating with different parameterizations.

For gated models, we vary a gate strength parameter $\alpha \in \{0, 0.25, 0.5, 1.0, 1.5\}$, which controls the degree of multiplicative modulation.

A.5.2. Training Setup

All models are trained using AdamW with learning rate 2×10^{-3}, weight decay 10^{-4}, and batch size 128 for 20 epochs. Results are averaged over 5 random seeds $\{0, 1, 2, 3, 4\}$. All model variants are trained under identical settings to ensure fair comparison.

A.5.3. Data Generation

We evaluate models on a synthetic sequence classification task designed to require nonlinear structure. Each dataset consists of 4000 training samples and 1000 test samples. Inputs are generated in $\mathbb{R}^2$ with latent centers sampled uniformly from $[-2, 2]^2$, and observation noise with standard deviation 0.20 is added to each sample.

A.5.4. Curvature Estimation

We estimate representation curvature using a finite-difference proxy:

$$\kappa(x) = \left\| \frac{f(x + \varepsilon v) - 2f(x) + f(x - \varepsilon v)}{\varepsilon^2} \right\|,$$

where v is a random unit direction and $\varepsilon = 10^{-2}$. The estimate is averaged over 64 random directions.

This quantity measures the magnitude of second-order variation of the representation map f and serves as a proxy for local

nonlinearity. It is not an invariant notion of Riemannian curvature and does not directly estimate sectional curvature. Rather, it is used for comparative analysis across model variants.

We report curvature of the attention output mean, and additionally compute curvature under a Fisher–Rao square-root embedding.

A.5.5. Robustness Under Anisotropic Metrics

To assess robustness under different metric scalings, we evaluate curvature under diagonal precision matrices with condition numbers $\{2, 4, 8, 12, 20\}$. This tests whether relative curvature comparisons are stable under anisotropic rescaling of the representation space.

A.6. Low-Rank Bottleneck and Geometric Interpretation

Recent empirical work (Qiu et al., 2025) identifies a limitation of standard softmax attention: the composition of value and output projections induces a low-rank linear bottleneck. In particular, the attention output can be written as $o_i^k = \sum_j S_{ij}^k X_j (W_V^k W_O^k)$, where the effective transformation is governed by $W_V^k W_O^k$, whose rank is at most $d_k \ll d_{\text{model}}$. Thus, attention operates through a low-rank linear map. This admits a geometric interpretation: since attention outputs are convex combinations of value vectors ($\alpha(X) \in \Delta^{n-1}$), they lie in the affine hull of these vectors and admit an affine parameterization $\mu_{\text{ung}}(\phi) = a + B\phi$. Consequently, the representation is confined to a k-dimensional affine subspace $a + \text{span}(B) \subset \mathbb{R}^D$, where $k = \text{rank}(B) \ll D$ in practice. As we show, such affine embeddings induce constant metrics and intrinsically flat statistical manifolds. In this sense, the low-rank bottleneck is not merely algebraic but geometric, corresponding to a restriction to flat representation geometry. Multiplicative gating removes this constraint via input-dependent nonlinear modulation, enabling non-flat (curved) geometries.