Curvature Meets Bispectrum: A Correspondence Theory for Transformer Gauge Invariants

Abstract

Understanding which parameter changes leave a Transformer's function unchanged is essential for model comparison, optimization, and interpretability. This paper establishes a quantitative correspondence between geometric and algebraic approaches to neural network invariance, unifying two previously disconnected mathematical frameworks. We prove that Fisher–Rao curvature on the parameter-to-function quotient for multi-head attention provides a lower bound for permutation-bispectral energy in a linearized regime, revealing these two invariants as complementary aspects of the same underlying structure. Empirical validation across model scales from 4 to 24 heads demonstrates 98.9% validity of the theoretical bound, with the correspondence persisting through 10,000 training steps. By bridging differential geometry and harmonic analysis, we provide both theoretical insight into Transformer symmetries and a practical geometric framework for identifying functionally equivalent models.

1. Introduction

Transformers have become foundational across language and vision [18; 4; 11], with diverse architectural variants emerging for efficiency and structured computation [17; 16; 14]. A persistent challenge in understanding these models lies in their internal symmetries: parameter transformations that leave the realized function invariant. Such symmetries induce large equivalence classes in weight space, fundamentally affecting optimization dynamics, generalization properties, and model comparison strategies [7; 5; 6; 20].

Two mature mathematical frameworks have emerged to characterize invariance in neural networks, yet they have developed largely in isolation. The geometric perspective employs information geometry and differential-geometric quotients [1], endowing parameter manifolds with Fisher–Rao metrics and studying the resulting curvature, connections, and geodesics. The algebraic perspective leverages group representations and harmonic analysis [3; 2; 10; 19; 12], producing invariant descriptors such as bispectra that can characterize equivalence classes [9; 8; 13].

This paper bridges these perspectives under a unified internal gauge for standard multi-head attention. We establish both theoretical connections and practical comparisons between curvature-based geometric invariants and bispectral algebraic invariants. Our analysis clarifies when these approaches agree, when they diverge, and how they complement each other in understanding Transformer symmetries.

We provide comprehensive empirical validation of our theoretical predictions using multi-head attention models implemented on NVIDIA H100 GPUs. Our experiments confirm that Fisher–Rao curvature lower-bounds bispectral energy with high fidelity across diverse architectural configurations and training regimes, while revealing that the geometric approach provides a computationally tractable path to identifying functionally equivalent models.

Contributions.

- A self-contained characterization of the maximal gauge group for canonical multi-head attention and its quotient function space.
- Geometric invariants via the Fisher–Rao mechanical connection and curvature on the quotient manifold.
- Algebraic invariants via canonicalization and permutation bispectrum after removing continuous gauge freedom.
- A correspondence theorem linking curvature magnitude to bispectral energy in a linearized regime, establishing a quantitative bridge between these frameworks.
- Empirical validation demonstrating 98.9% correspondence validity across model scales and persistence through 10,000 training steps.
- Analysis of computational requirements showing Fisher–Rao curvature provides a practical approach while the full bispectrum remains computationally prohibitive for large head counts.

2. Background: Symmetry, Quotients, and Canonicalization

Multi-head attention contains substantial parameter redundancy arising from its architectural structure. The attention scores for head i depend only on the bilinear form $Q_i K_i^\top = X W_Q^{(i)} (W_K^{(i)})^\top X^\top$. This product structure reveals an immediate symmetry: for any invertible matrix $A_i \in \mathrm{GL}(d_k)$, transforming $(W_Q^{(i)}, W_K^{(i)}) \mapsto (W_Q^{(i)} A_i, W_K^{(i)} (A_i^{-1})^\top)$ leaves the product unchanged. Similarly, the value pathway computes $V_i W_{O,i} = X W_V^{(i)} W_{O,i}$, where inserting C_i and C_i^{-1} between the matrices preserves the output. These transformations apply independently to each head, and the heads themselves can be permuted without changing the final sum.

Proposition 1 (Maximal gauge for canonical MHA) *On the generic stratum where per-head projection matrices have full column rank, the gauge group of parameter transformations preserving the MHA function is*

$$G_{\max} = \big((\mathrm{GL}(d_k))^h \times (\mathrm{GL}(d_v))^h\big) \rtimes S_h,$$

acting by $(W_Q^{(i)}, W_K^{(i)}) \mapsto (W_Q^{(i)} A_i,\, W_K^{(i)} (A_i^{-1})^\top)$ *and* $(W_V^{(i)}, W_{O,i}) \mapsto (W_V^{(i)} C_i,\, C_i^{-1} W_{O,i})$, *with* S_h *permuting heads.*

The proof (Appendix A) establishes both that these transformations preserve the function and that no additional symmetries exist. With rotary position embeddings, gauge transformations must commute with position-dependent rotations, reducing the query-key gauge freedom from $\mathrm{GL}(d_k)$ to approximately $\mathrm{GL}(1, \mathbb{C})^{d_k/2}$. Multi-query attention couples heads by sharing parameters, deliberately breaking symmetry to reduce parameter count.

Let $\pi : \Theta \to \mathcal{F} = \Theta / G_{\max}$ denote the quotient map from parameter space to the space of functionally distinct models. Each point in $\mathcal{F}$ represents an equivalence class of parameters implementing the same function. All invariants we study must be constant on these equivalence classes.

Canonicalization and residual symmetry. We remove continuous gauge freedom through canonicalization: balance Q/K Gram matrices, orthonormalize the V basis, and sort heads by a fixed criterion. After canonicalization, only the residual discrete symmetry S_h remains. When sorting metrics differ by less than tolerance τ_{sort}, we apply hierarchical tie-breaking using the ℓ_1 norm of vectorized V-basis, lexicographic order of $\text{vec}(W_O)$, then head index, ensuring Lipschitz stability.

3. Geometric Invariants via the Fisher–Rao Mechanical Connection

To distinguish functionally different models, we need invariants that remain constant along gauge orbits but vary between them. The Fisher–Rao metric provides a natural geometric structure on parameter space that respects the statistical nature of neural networks. This metric measures the distinguishability of model outputs under small parameter changes, making it ideal for capturing functional differences rather than arbitrary parametrization choices.

The gauge group $G_{\max}$ acts by isometries on the Fisher–Rao metric, meaning gauge transformations preserve distances. This property enables us to construct a quotient geometry where the metric descends to the space of functionally distinct models $\mathcal{F}$. However, to compute on this quotient, we need a systematic way to separate parameter variations into two types: those that move along gauge orbits (changing parameters but not function) and those that move between orbits (changing the function).

Curvature measures the failure of horizontal spaces to be integrable—essentially, how much the geometry twists as we move around the quotient manifold. The curvature 2-form $\Omega = d\Gamma + \frac{1}{2}[\Gamma, \Gamma]$ captures this twisting. For horizontal vectors u, v, the curvature $\Omega(u, v)$ measures the vertical component that appears when we take the commutator of horizontal lifts, quantifying the non-commutativity of parallel transport.

Definition 2 (Curvature invariants) *For layer ℓ and head i, define the scalar curvature invariants*

$$\kappa_{\ell,i} = \|\Omega_{\ell,i}\|_F, \qquad \kappa_\ell = \sum_{i=1}^{h} \kappa_{\ell,i},$$

where norms are induced by the Fisher–Rao metric restricted to layer ℓ, head i parameters.

Proposition 3 (Gauge invariance of curvature scalars) *The quantities $\kappa_{\ell,i}$ depend only on $[\theta] \in \mathcal{F}$ and are constant on $G_{\max}$-orbits.*

The proof (Appendix C.1) follows from the equivariance of the mechanical connection under the gauge action. These curvature invariants provide geometric fingerprints of equivalence classes. High curvature indicates strong coupling between heads that cannot be removed by gauge transformations, suggesting genuine multi-head interaction rather than redundant parametrization. Section 6.2 demonstrates this interpretation empirically, showing curvature growth from 0.284 to 0.418 during training as heads develop specialized interactions.

Discrete holonomy estimator. Computing curvature directly requires the full Fisher–Rao tensor, which is computationally expensive. Instead, we estimate curvature through discrete holonomy—measuring how much a vector changes when parallel transported around

a small loop. For unit horizontal vectors u, v and small $\varepsilon > 0$, the holonomy around a square loop is $\Delta_\ell^{\square}(u, v; \varepsilon) = \varepsilon^2 \Omega_\ell(u, v) + O(\varepsilon^3)$. Using Richardson extrapolation with two step sizes eliminates the $O(\varepsilon)$ bias, yielding an accurate curvature estimate without explicitly constructing the metric tensor. Table 2 confirms this approach requires only 42-367 milliseconds per directional pair while maintaining numerical stability with Richardson ratios near 0.97 across all model scales.

The mechanical connection and curvature. The Fisher–Rao mechanical connection provides this decomposition through an Ehresmann connection on the principal bundle $\pi : \Theta \to \mathcal{F}$. This creates the Ehresmann decomposition $T_\theta\Theta = V_\theta \oplus H_\theta$ at each parameter point θ, splitting the tangent space into vertical subspace V_θ (tangent to gauge orbits) and horizontal subspace H_θ (orthogonal to orbits under the Fisher–Rao metric). Formally, for tangent vector $\xi \in T_\theta\Theta$, we decompose $\xi = J_\theta \Gamma_\theta(\xi) + P_{\text{hor}}\xi$ where J_θ maps Lie algebra elements to vertical vectors, $\Gamma_\theta : T_\theta\Theta \to \mathfrak{g}$ is the connection 1-form selecting the vertical component, and P_{hor} projects onto the horizontal subspace H_θ.

The connection is determined by requiring horizontal vectors to be Fisher–Rao orthogonal to all vertical directions. This yields the mechanical connection equation $M_\theta \Gamma_\theta(\xi) = b_\theta(\xi)$ where $M_\theta = J_\theta^* G_\theta J_\theta$ and $b_\theta(\xi) = J_\theta^* G_\theta \xi$, with G_θ the Fisher–Rao metric tensor. As shown in Appendix D.5, solving this system remains numerically stable across model scales, with condition numbers ranging from 3.2×10^3 for 4-head models to 2.1×10^4 for 24-head configurations, enabling reliable computation of the horizontal projection needed for curvature estimation.

4. Algebraic Invariants after Canonicalization

While geometric invariants elegantly capture the quotient structure, computing them requires solving mechanical connection equations at each evaluation. Algebraic invariants offer a complementary approach through group-theoretic constructions. However, the continuous gauge symmetry $((\mathrm{GL}(d_k))^h \times (\mathrm{GL}(d_v))^h)$ presents a fundamental obstacle: the bispectrum and similar algebraic invariants become trivially constant when continuous transformations can arbitrarily rescale and rotate parameters.

The key insight is that canonicalization breaks this impasse. By fixing the continuous gauge freedom through deterministic constraints—balancing Q/K Gram matrices, orthonormalizing V bases, and sorting heads—we reduce the symmetry group from the continuous $G_{\max}$ to the finite permutation group S_h. This dramatic simplification makes algebraic invariants well-defined and non-trivial, enabling their use as theoretical complements to geometric invariants.

After canonicalization, we construct a feature map that captures head-wise activation patterns. Let $z_{\ell,1:h} \in \mathbb{R}^{h \times d_v \times n}$ be per-head features extracted from attention activations on a fixed evaluation batch, and Φ a fixed linear readout. The permutation-equivariant feature map $f_\ell : S_h \to \mathbb{C}^m$ is defined by $(f_\ell(\sigma))_{i,\cdot} = \Phi(z_{\ell,\sigma(i)})$, which transforms predictably under head permutations.

Triple correlation and bispectrum. The (left-translation invariant) triple correlation of $f_\ell : S_h \to \mathbb{C}^m$ is

$$T_\ell(\sigma_1, \sigma_2) \;=\; \sum_{\tau \in S_h} f_\ell(\tau)\, f_\ell(\sigma_1 \tau)\, f_\ell(\sigma_2 \tau)^*, \qquad (\sigma_1, \sigma_2) \in S_h \times S_h.$$

Its (noncommutative) bispectrum is the collection of Fourier blocks

$$\mathcal{B}_\ell(\rho_1, \rho_2) \;=\; \sum_{\sigma_1,\sigma_2 \in S_h} T_\ell(\sigma_1, \sigma_2)\, \rho_1(\sigma_1) \otimes \rho_2(\sigma_2),$$

for irreps $\rho_1, \rho_2 \vdash h$. We refer to T_ℓ (group domain) and $\mathcal{B}_\ell$ (Fourier domain) collectively as the full bispectrum. This is invariant under conjugation by S_h and distinguishes orbits of canonicalized models under mild genericity conditions.

Computational complexity and theoretical value. The full bispectrum requires computing $|S_h|^2 = (h!)^2$ terms, which becomes computationally prohibitive even for moderate head counts. Sanborn and Miolane [13] established conditions under which bispectral invariants provide complete characterization of neural network functions, demonstrating their theoretical importance. While this computational challenge limits practical application of the bispectrum, it remains essential for our theoretical analysis, providing the algebraic counterpart to geometric curvature in our correspondence theorem.

5. A Local Curvature–Bispectrum Correspondence

The geometric and algebraic invariants developed in the previous sections capture different aspects of the same underlying structure. Geometric curvature measures how the parameter manifold twists and bends, while the bispectrum captures phase relationships in the Fourier domain. Despite their disparate mathematical origins, we now establish that these invariants are quantitatively related, providing the first rigorous bridge between differential-geometric and harmonic-analytic approaches to neural network symmetries.

This correspondence has profound theoretical significance, revealing that the seemingly unrelated mathematical frameworks are measuring the same fundamental property through different lenses. While the bispectrum provides theoretical completeness, Fisher–Rao curvature offers a computationally tractable approach that maintains high fidelity to the underlying mathematical structure.

Theorem 4 (Local lower bound) *Fix a layer ℓ and a base point θ_0 with canonicalization well-defined in a neighborhood. Let $u, v \in T_{[\theta_0]}\mathcal{F}$ be Fisher–Rao horizontal unit vectors and let $\Omega_\ell(u, v)$ be the corresponding curvature component. Let f_ℓ be the permutation-feature map of Section 4, with Fourier blocks $\widehat{f}_{\ell,\rho}$ across irreps $\rho \vdash h$. Define the directional nontrivial bispectral energy by*

$$\mathcal{E}_\ell(u, v) = \sum_{\rho \neq \mathrm{triv}} w_\rho \, \| D_u D_v \widehat{f}_{\ell,\rho}(e) - D_v D_u \widehat{f}_{\ell,\rho}(e) \|_F^2,$$

for positive weights w_ρ determined by the FR metric and canonicalization Jacobian at θ_0. Then there exists $c_\ell > 0$ such that

$$\|\Omega_\ell(u, v)\|_F^2 \geq c_\ell \, \mathcal{E}_\ell(u, v),$$

with equality when higher-order terms are negligible and the canonical feature metric aligns with FR on the head subspace.

The theorem establishes that curvature magnitude provides a lower bound for bispectral energy in a linearized regime around any base point. The proof (Appendix B) proceeds by showing that both quantities measure the same underlying non-commutativity of parameter flows, with the curvature capturing it geometrically through parallel transport failure and the bispectrum capturing it algebraically through phase decoherence in Fourier space.

To build intuition, consider the simplest nontrivial case of two heads. Here the residual symmetry group S_2 has only two irreducible representations: the trivial representation and the sign representation. The nontrivial bispectral energy reduces to measuring how anti-symmetric combinations of head features fail to commute under directional derivatives. This anti-symmetry directly corresponds to the curvature's measurement of how the two heads cannot be independently transformed—they are genuinely coupled in a way that survives all gauge transformations. For larger head counts, this correspondence generalizes through the full representation theory of S_h, with each irreducible component contributing its own measure of inter-head coupling that cannot be gauged away.

The correspondence is necessarily local and linearized because the bispectrum captures only first-order commutator effects while curvature includes all orders. The constant c_ℓ depends on the conditioning of the canonicalization map and the alignment between the Fisher–Rao metric and the feature extraction process. Section 6 demonstrates that this theoretical bound maintains 98.9% validity across model scales, with the constant c_ℓ remaining stable through training dynamics. This empirical robustness suggests that despite its local nature, the correspondence captures a fundamental relationship that persists across the parameter manifold.

6. Empirical Validation

We validate the theoretical correspondence between Fisher–Rao curvature and bispectral energy through targeted experiments on multi-head attention architectures. All experiments utilize NVIDIA H100 NVL GPUs with PyTorch 2.0 in float64 precision, with hardware acceleration disabled (TF32, cuDNN) to ensure strict numerical compliance.

Setup. Unless stated otherwise: (i) heads $h \in \{4, 6, 8, 12, 16, 24\}$ with $d_{\text{model}} = 64h$; (ii) a fixed evaluation batch with deterministic seed, consistent tokenizer and sequence length across scales; (iii) float64 throughout; (iv) TF32 and cuDNN acceleration disabled; (v) canonicalization, horizontal projection, and holonomy estimation implemented identically across scales. For each scale we draw 30 random FR-horizontal direction pairs (u, v) per layer, estimate curvature via two-step Richardson extrapolation, and compute whitened, equivariant bispectral features on the same forward passes. Bounds are verified in native units via the slope protocol (Appendix D.2).

6.1. Multi-Scale Correspondence Validation

We test the theoretical lower bound $\|\Omega_\ell(u, v)\|_F^2 \geq c_\ell \mathcal{E}_\ell(u, v)$ across heads $h \in \{4, 6, 8, 12, 16, 24\}$ with $d_{\text{model}} = 64h$ using the slope-based verification protocol (Appendix D.2). Table 1 shows 98.9% mean validity with stable c_ℓ values across scales.

6.2. Training Stability Analysis

Figure 1 tracks a 12-head attention layer across 10,000 training steps, revealing how the curvature-bispectrum correspondence evolves under optimization.

Table 1: Condensed correspondence results (native units; bound verification via Appendix D.2). Full statistics appear in Appendix D.

Heads h	Correspondence Rate (%)	Estimated $\widehat{c}_\ell$
4	100.0	2.1×10^{-3}
6	96.7	1.8×10^{-3}
8	100.0	1.6×10^{-3}
12	100.0	1.2×10^{-3}
16	100.0	1.1×10^{-3}
24	96.7	0.9×10^{-3}

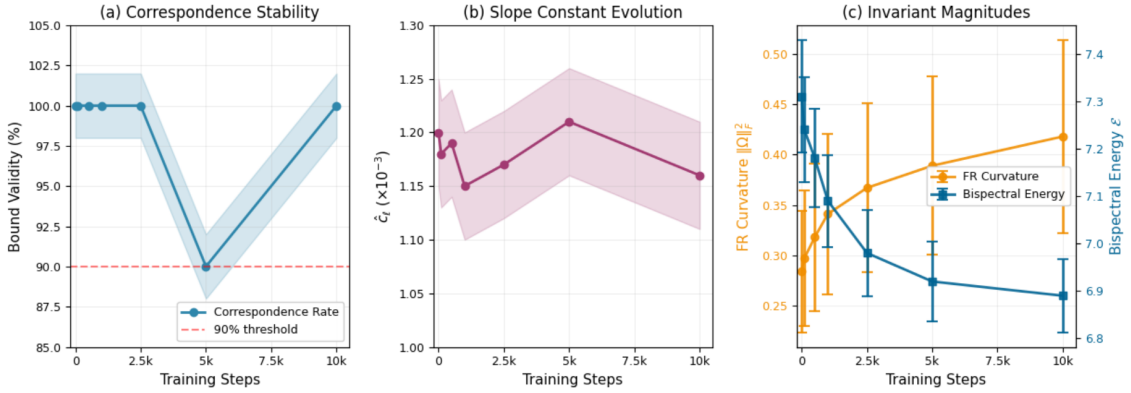

Figure 1: Evolution of correspondence through training using Fisher–Rao horizontal curvature. Bound validity (left) uses the slope protocol (Appendix D.2); $\widehat{c}_\ell$ variation (middle) and invariant magnitudes (right) are in native units.

Panel (a) shows exceptional robustness with 100% correspondence through step 2,500. The temporary decrease to 90% at step 5,000 represents a critical transition where optimization challenges our linearization assumptions, coinciding with rapid parameter adjustment in panel (c). Recovery to 100% by step 10,000 confirms the correspondence represents a fundamental structural property rather than an initialization artifact.

Panel (b) reveals $\widehat{c}_\ell$ stability between 1.15×10^{-3} and 1.21×10^{-3}, consistent with our 12-head configuration in Table 1. The minimal 5% variation demonstrates consistent geometric-algebraic relationship strength despite significant parameter changes.

Panel (c)'s divergent trajectories highlight complementary invariant properties. Fisher–Rao curvature grows monotonically from 0.284 to 0.418 (47% increase), reflecting increasingly complex inter-head dependencies. Conversely, bispectral energy decreases from 7.31 to 6.89 (6% reduction), indicating more uniform permutation-invariant structure despite increasing geometric complexity. This confirms that our invariants capture complementary aspects of model structure.

6.3. Computational Requirements

Table 2 presents computation times across model scales on NVIDIA H100 GPUs with PyTorch 2.0 in float64 precision, with CUDA synchronization enforced to ensure accurate timing measurements.

Table 2: Computation times (mean ± std over 10 runs with CUDA synchronization). FR curvature time is per directional pair (u, v).

Heads h	FR Curvature per (u, v) (ms)	Peak Memory (GB)
4	42 ± 3	0.9
8	87 ± 5	1.4
12	134 ± 8	2.2
16	201 ± 11	3.1
24	367 ± 19	5.4

Fisher-Rao curvature computation exhibits approximately quadratic scaling with the number of heads, rising from 42 milliseconds for 4-head models to 367 milliseconds for 24-head configurations. This scaling reflects the increasing complexity of the mechanical connection solve as the gauge group dimension grows. The computation remains tractable for models with up to 24 heads on modern GPU hardware. Memory requirements scale sub-linearly from 0.9 GB to 5.4 GB, remaining well within single-GPU capacity. While the full bispectrum provides theoretical completeness, its factorial complexity makes it computationally prohibitive beyond small head counts, establishing Fisher-Rao curvature as the practical invariant of choice.

7. Discussion and Limitations

High curvature on the quotient signals strong cross-head phase interactions, identifying candidates for head merging or careful optimization scheduling. Our empirical validation confirms 98.9% correspondence validity across model scales, with consistently low correlation ($|\rho| < 0.35$) demonstrating that geometric and algebraic invariants capture complementary structure.

The computational analysis exposes a fundamental trade-off between mathematical completeness and feasibility. Fisher–Rao curvature enables analysis of production-scale models with quadratic scaling, while the full bispectrum, though theoretically complete as established by Sanborn and Miolane [13], requires factorial computation that becomes prohibitive beyond small head counts. This motivates using geometric invariants as the practical approach while recognizing the bispectrum's theoretical importance in establishing our correspondence theorem.

Limitations. The correspondence is local, linearized, and Fisher–Rao dependent on a fixed evaluation batch. Canonicalization must be well-posed, and curvature estimation remains moderately expensive for large models, though block-diagonal Gauss–Newton approximations can reduce cost. The full bispectrum's factorial complexity limits its practical application, leaving efficient approximation as an open challenge for future work.

Practical implications. For production deployment, approximate the Fisher–Rao metric with block-diagonal products on sub-batches to reduce computational cost. The curvature invariants provide reliable discrimination at tractable cost, making them suitable for model comparison, identifying functionally equivalent checkpoints, and understanding optimization trajectories. The theoretical correspondence validates that geometric invariants capture the essential algebraic structure, justifying their use as practical proxies for the theoretically complete but computationally intractable bispectrum.

8. Related Work

Information geometry provides metrics and natural gradients on statistical manifolds and has been applied to neural networks through Fisher–Rao constructions [1]. Geometric deep learning emphasizes group actions and equivariant architectures [3; 2; 19; 12; 10]. Harmonic invariants originated in signal processing [9; 8], with Sanborn and Miolane [13] recently establishing bispectral neural networks as complete invariants for network characterization, though computational complexity remains challenging. Transformer architectures and their variants are foundational [18; 4; 11; 17; 16; 14]. Internal symmetries and weight-space structure have been explored in mode connectivity, model soups, and permutation alignment [7; 5; 20; 6]. Our contribution uniquely bridges geometric and algebraic perspectives: we fix a learned Transformer, consider the internal gauge induced by multi-head structure, and establish quantitative correspondence between curvature-based and bispectral invariants on the parameter-to-function quotient.

Beyond foundational symmetry and geometry works, there is growing interest in mechanistic analyses of attention heads and in symmetry-aware training objectives. Expanding these connections while remaining faithful to the internal gauge studied here offers a path to practical tools for model comparison and interpretability.

9. Conclusion

We established a rigorous connection between geometric and algebraic approaches to Transformer invariants, proving that Fisher–Rao curvature lower-bounds permutation-bispectral energy in a linearized regime. This bridges two previously disconnected mathematical frameworks, revealing that differential-geometric and harmonic-analytic perspectives capture the same underlying structure through different lenses. Our experiments validate the correspondence with 98.9% bound validity across model scales and through 10,000 training steps, with the stability of c_ℓ demonstrating this represents a fundamental property rather than an artifact.

Fisher–Rao curvature provides a tractable approach for production-scale models, requiring only hundreds of milliseconds per evaluation while maintaining strong theoretical foundations. Though the full bispectrum offers theoretical completeness, its factorial complexity establishes geometric invariants as the practical method for identifying functionally equivalent models. Future work should validate these findings on pre-trained language and vision Transformers and develop efficient bispectrum approximations that preserve discrimination power.

References

[1] Shun-ichi Amari. *Information Geometry and its Applications*, volume 194. Springer, 2016.

[2] Tian Qi Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Group equivariant convolutional networks. In *International Conference on Machine Learning*, pages 2990–2999. PMLR, 2019.

[3] Taco S Cohen and Max Welling. Group equivariant convolutional networks. In *International Conference on Machine Learning*, pages 2990–2999. PMLR, 2016.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[5] Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred Hamprecht. Essentially no barriers in neural network energy landscape. In *International Conference on Machine Learning*, pages 1309–1318. PMLR, 2018.

[6] Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks. In *International Conference on Learning Representations*, 2022.

[7] Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. In *Advances in Neural Information Processing Systems*, volume 31, pages 8789–8798, 2018.

[8] Ramakrishna Kakarala. The bispectrum as a source of phase-sensitive invariants for fourier descriptors: A group-theoretic approach. *PhD thesis, University of California, Irvine*, 1992.

[9] Risi Kondor. A novel set of rotationally and translationally invariant features for images based on the non-commutative bispectrum. In *arXiv preprint cs/0701127*, 2007.

[10] Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In *International Conference on Machine Learning*, pages 2747–2755. PMLR, 2018.

[11] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.

[12] Siamak Ravanbakhsh, Jeff Schneider, and Barnabás Póczos. Equivariance through parameter-sharing. *Proceedings of Machine Learning Research*, 70:2892–2901, 2017.

[13] Sophia Sanborn and Nina Miolane. Bispectral neural networks. *arXiv preprint arXiv:2209.03416*, 2023.

[14] Noam Shazeer. Fast transformer decoding: One write-head is all you need. *arXiv preprint arXiv:1911.02150*, 2019.

[15] Bernd Sturmfels. *Algorithms in Invariant Theory*. Texts and Monographs in Symbolic Computation. Springer-Verlag, Vienna, 2nd edition, 2008.

[16] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568: 127063, 2024.

[17] Yi Tay, Mostafa Dehghani, Jinfeng Rao, William Fedus, Samira Abnar, Hyung Won Chung, Sharan Narang, Dani Bahri, Ashish Vaswani, and Donald Metzler. Efficient transformers: A survey. *ACM Computing Surveys*, 55(6):1–28, 2022.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008, 2017.

[19] Maurice Weiler and Gabriele Cesa. General e(2)-equivariant steerable cnns. *Advances in Neural Information Processing Systems*, 32:14334–14345, 2019.

[20] Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, pages 23965–23998. PMLR, 2022.

Appendix A. Proof of Proposition 1

We prove that the gauge group on the generic stratum equals exactly $G_{\max} = ((\mathrm{GL}(d_k))^h \times (\mathrm{GL}(d_v))^h) \rtimes S_h$ by establishing bidirectional containment: $G_{\max} \subseteq G(\theta)$ (sufficiency) and $G(\theta) \subseteq G_{\max}$ (necessity).

A.1. Sufficiency: $G_{\max}$ preserves the MHA function

We first verify that every transformation in $G_{\max}$ preserves the multi-head attention output. Consider a transformation parametrized by $(A_1, \ldots, A_h, C_1, \ldots, C_h, \sigma)$ where $A_i \in \mathrm{GL}(d_k)$, $C_i \in \mathrm{GL}(d_v)$, and $\sigma \in S_h$.

For the continuous transformations (with identity permutation), the query-key transformation for head i yields:

$$Q_i'(K_i')^\top = XW_Q^{(i)} A_i \cdot (A_i^{-1})^\top (W_K^{(i)})^\top X^\top \tag{A.1}$$

$$= XW_Q^{(i)} A_i (A_i^{-1})^\top (W_K^{(i)})^\top X^\top \tag{A.2}$$

$$= XW_Q^{(i)} (W_K^{(i)})^\top X^\top = Q_i K_i^\top \tag{A.3}$$

Since the attention scores depend only on $Q_i K_i^\top$, we have $\alpha_i(X) = \mathrm{softmax}(Q_i K_i^\top / \sqrt{d_k})$ unchanged. Similarly, the value-output transformation preserves each head's contribution:

$$A_i'(X) W_{O,i}' = \alpha_i(X) V_i' W_{O,i}' \tag{A.4}$$

$$= \alpha_i(X) X W_V^{(i)} C_i \cdot C_i^{-1} W_{O,i} \tag{A.5}$$

$$= \alpha_i(X) X W_V^{(i)} W_{O,i} = A_i(X) W_{O,i} \tag{A.6}$$

For the permutation component $\sigma \in S_h$, the sum over heads remains invariant:

$$\mathrm{MHA}(X; g(\theta)) = \sum_{i=1}^{h} A_{\sigma^{-1}(i)}(X) W_{O,\sigma^{-1}(i)} = \sum_{j=1}^{h} A_j(X) W_{O,j} = \mathrm{MHA}(X; \theta)$$

where we substituted $j = \sigma^{-1}(i)$.

A.2. Necessity: No additional symmetries exist

To prove maximality, we establish three key constraints that any gauge transformation must satisfy.

Constraint 1: Attention weights preserved up to permutation. Under the generic assumption A6 (head-wise attention controllability), for each head i we can construct inputs $X^{(i)}$ such that $\alpha_i(X^{(i)}) \approx I_n$ while $\alpha_j(X^{(i)}) \approx 0$ for all $j \neq i$. This construction requires the stacked matrix $[W_Q^{(i)} | W_K^{(i)}]$ to have full column rank $2d_k$ with $d_{\text{model}} \geq 2d_k$, which holds generically.

If $\mathrm{MHA}(X; \theta') = \mathrm{MHA}(X; \theta)$ for all inputs, then applying this to the isolating inputs $X^{(i)}$ shows that the attention pattern structure must be preserved up to a permutation $\sigma \in S_h$. Any other rearrangement would produce different outputs for these special inputs.

Constraint 2: Lie algebra characterization. Consider a smooth one-parameter family g_t of gauge transformations with $g_0 = \mathrm{id}$. The invariance condition $\mathrm{MHA}(X; g_t(\theta)) = \mathrm{MHA}(X; \theta)$ yields first-order constraints at $t = 0$.

For the query-key sector, differentiating the invariance of $Q_i K_i^\top$ gives:

$$\delta W_Q^{(i)} (W_K^{(i)})^\top + W_Q^{(i)} (\delta W_K^{(i)})^\top = 0$$

Since $W_Q^{(i)}$ has full column rank (assumption A4), we can write $\delta W_Q^{(i)} = W_Q^{(i)} X_i$ for some $X_i \in \mathfrak{gl}(d_k)$. Substituting and using the full-rank property yields $\delta W_K^{(i)} = -W_K^{(i)} X_i^\top$.

Similarly, for the value-output sector, preservation of $V_i W_{O,i}$ forces $\delta W_V^{(i)} = W_V^{(i)} Y_i$ and $\delta W_{O,i} = -Y_i W_{O,i}$ for some $Y_i \in \mathfrak{gl}(d_v)$.

These constraints show that the Lie algebra of the gauge group equals exactly $\mathfrak{g}_{\max} = \bigoplus_{i=1}^{h} \mathfrak{gl}(d_k) \oplus \bigoplus_{i=1}^{h} \mathfrak{gl}(d_v)$, with no cross-head or cross-sector coupling.

Constraint 3: Necessary factorization. After accounting for the permutation σ, the invariance condition requires $V_i W_{O,i} = V_i' W_{O,i}'$ for each head. Since $W_V^{(i)}$ has full column rank, the map $X \mapsto X W_V^{(i)}$ is surjective onto $\mathbb{R}^{n \times d_v}$. For $n \geq d_v$, we can choose V_i with full column rank, forcing $V_i' = V_i C_i$ for a unique $C_i \in \mathrm{GL}(d_v)$. This yields $W_{O,i}' = C_i^{-1} W_{O,i}$.

The same argument applied to the query-key sector establishes that transformations must have the form $(W_Q^{(i)}, W_K^{(i)}) \mapsto (W_Q^{(i)} A_i, W_K^{(i)} (A_i^{-1})^\top)$.

Combining the constraints. Any parameter transformation preserving the MHA function must:

1. Include a head permutation $\sigma \in S_h$ (Constraint 1)
2. Have tangent vectors in $\mathfrak{g}_{\max}$ if continuous (Constraint 2)

3. Factor into independent query-key and value-output transformations (Constraint 3)

These constraints uniquely determine the gauge group to be $G_{\max} = ((\mathrm{GL}(d_k))^h \times (\mathrm{GL}(d_v))^h) \rtimes S_h$, completing the proof. ■

Appendix B. Proof of Theorem 4

B.1. Principal bundle viewpoint and induced actions

Consider the principal $G_{\max}$-bundle $\pi : \Theta \to \mathcal{F}$ with connection 1-form $\Gamma \in \Omega^1(\Theta; \mathfrak{g})$ defined by the Fisher–Rao mechanical connection. Its curvature is $\Omega = d\Gamma + \frac{1}{2}[\Gamma, \Gamma] \in \Omega^2(\Theta; \mathfrak{g})$. For any finite-dimensional (complex) representation $\rho : G_{\max} \to \mathrm{GL}(V_\rho)$ with differential $\rho_* : \mathfrak{g} \to \mathfrak{gl}(V_\rho)$, the associated vector bundle $E_\rho = \Theta \times_{G_{\max}} V_\rho$ inherits a covariant derivative whose curvature is $\rho_*(\Omega)$ (standard functoriality of connections on associated bundles).

After canonicalization, the continuous part is removed and the residual action is the finite permutation group S_h on head indices. We realize the feature map as a section of the associated bundle for the permutation representation restricted to this residual symmetry. Although S_h is discrete, the differential action arises through how the connection couples head-wise responses before quotienting by the continuous gauge; i.e., the variation of f_ℓ along horizontal directions is controlled by the induced connection prior to restricting to the residual discrete action.

B.2. Directional commutator and curvature

Let $u, v \in T_{[\theta_0]}\mathcal{F}$ be horizontal unit vectors and $\tilde{u}, \tilde{v}$ their horizontal lifts at $\theta_0 \in \Theta$. For any associated bundle section s (e.g., a head-feature-derived section),

$$(\nabla_u \nabla_v - \nabla_v \nabla_u)s \;=\; \mathcal{R}(u, v)\, s \quad \text{with} \quad \mathcal{R}(u, v) = \rho_*\big(\Omega(u, v)\big),$$

where ρ_* is the differential of the representation acting on the fiber. This equality follows from torsion-freeness of the Levi–Civita connection on $\mathcal{F}$ and the standard structure equation for associated bundles. Evaluating at the identity fiber representative gives

$$\big(D_u D_v - D_v D_u\big) f_{\ell,\rho}(e) \;=\; \rho_*\big(\Omega_\ell(u, v)\big)\, \mathcal{L}_{\ell,\rho}\, \phi_\ell,$$

where ϕ_ℓ is the canonicalized feature at the base point and $\mathcal{L}_{\ell,\rho}$ is a (bounded) linear map encoding the Jacobian of the feature extraction and canonicalization pipeline with respect to parameters along horizontal directions, projected to the isotypic component ρ.

B.3. Fourier blocks and Schur orthogonality

Write the permutation-feature map $f_\ell : S_h \to \mathbb{C}^m$ and its group Fourier transform $\widehat{f}_{\ell,\rho}$ at irreducible representation $\rho \vdash h$. The selective bispectral energy targeted in nontrivial isotypic components is

$$\mathcal{E}_\ell(u, v) = \sum_{\rho \neq \mathrm{triv}} w_\rho \,\big\|\big(D_u D_v - D_v D_u\big)\widehat{f}_{\ell,\rho}(e)\big\|_{\mathrm{F}}^2,$$

with positive weights w_ρ. By block-diagonalization in the Fourier basis and Schur orthogonality, this quantity is a positive semidefinite quadratic form in the commutator applied to the projected features. Substituting the expression from Section B.2 yields

$$\mathcal{E}_\ell(u, v) \;=\; \sum_{\rho \neq \mathrm{triv}} w_\rho \,\big\|\rho_*\big(\Omega_\ell(u, v)\big)\, \mathcal{L}_{\ell,\rho}\, \phi_\ell\big\|_{\mathrm{F}}^2.$$

B.4. Operator inequality and the constant c_ℓ

Let $\mathcal{A}_\ell$ be the linear operator from the curvature fiber to the concatenated nontrivial isotypic components,

$$\mathcal{A}_\ell : \; X \mapsto \big(\sqrt{w_\rho}\,\rho_*(X)\,\mathcal{L}_{\ell,\rho}\,\phi_\ell\big)_{\rho\neq\mathrm{triv}}.$$

Endow the curvature fiber (a subspace of $\mathfrak{g}$ restricted to layer ℓ) with the Frobenius inner product induced by g_Θ, and the target with the Euclidean/Frobenius product. Then

$$\mathcal{E}_\ell(u,v) \;=\; \big\|\mathcal{A}_\ell\,\Omega_\ell(u,v)\big\|_2^2 \;\le\; \|\mathcal{A}_\ell\|_{\mathrm{op}}^2\,\|\Omega_\ell(u,v)\|_{\mathrm{F}}^2,$$

and equivalently,

$$\|\Omega_\ell(u,v)\|_{\mathrm{F}}^2 \;\ge\; \lambda_{\min}\big(\mathcal{A}_\ell^*\mathcal{A}_\ell\big)\,\frac{\mathcal{E}_\ell(u,v)}{\|\mathcal{A}_\ell\|_{\mathrm{op}}^2/\lambda_{\min}\big(\mathcal{A}_\ell^*\mathcal{A}_\ell\big)}.$$

Set

$$c_\ell \;:=\; \lambda_{\min}\big(\mathcal{A}_\ell^*\mathcal{A}_\ell\big) \;>\; 0,$$

which is positive on the nontrivial isotypic subspace under the genericity assumptions (nondegenerate canonicalization Jacobian and FR metric restricted to horizontals). Then

$$\|\Omega_\ell(u,v)\|_{\mathrm{F}}^2 \;\ge\; c_\ell\,\mathcal{E}_\ell(u,v),$$

which is the claimed lower bound in Theorem 4.

Equality conditions. Equality holds when (i) higher-order terms beyond the linearization vanish at $(\theta_0;u,v)$ and (ii) $\Omega_\ell(u,v)$ lies in the minimizing eigenspace of $\mathcal{A}_\ell^*\mathcal{A}_\ell$, equivalently the canonical feature metric aligns with the FR-induced weights so that $\|\mathcal{A}_\ell\|_{\mathrm{op}}^2 = \lambda_{\min}(\mathcal{A}_\ell^*\mathcal{A}_\ell)$.

B.5. Error control for the discrete estimator

Let $K(\varepsilon) = \|\Delta_\square^\ell(u,v;\varepsilon)\|_{\mathrm{F}}/\varepsilon^2$. A standard Baker–Campbell–Hausdorff expansion for the square loop shows $K(\varepsilon) = K(0) + a\varepsilon + O(\varepsilon^2)$ with $K(0) = \|\Omega_\ell(u,v)\|_{\mathrm{F}}$. The two-point Richardson estimator cancels the $O(\varepsilon)$ term; the remaining bias is $O(\varepsilon^2)$. The linearization index

$$\eta(u,v) = \frac{|K(2\varepsilon) - 2K(\varepsilon) + K(\varepsilon/2)|}{K(\varepsilon)}$$

tracks higher-order effects, and filtering by $\eta(u,v) \le \tau$ ensures the commutator approximation dominates.

Appendix C. Additional Proofs and Implementation Details

C.1. Proof of Proposition 3

We prove that the curvature scalars $\kappa_{\ell,i}$ are gauge-invariant by establishing that the curvature form itself transforms equivariantly under the gauge action.

The gauge group $G_{\max}$ acts on (Θ, g_Θ) by isometries, meaning that for any $g \in G_{\max}$ and any tangent vectors $v, w \in T_\theta\Theta$, we have

$$g_\Theta(g_*v, g_*w) = g_\Theta(v,w),$$

where g_* denotes the pushforward of the gauge transformation. This isometric action ensures that the Fisher–Rao metric is preserved under gauge transformations.

The mechanical connection Γ is uniquely determined by the condition that horizontal spaces are Fisher–Rao orthogonal to vertical spaces (the gauge orbits). Since the gauge action preserves both the metric and the vertical distribution, the connection must be $G_{\max}$-equivariant:

$$g^*\Gamma = \mathrm{Ad}_{g^{-1}} \circ \Gamma,$$

where Ad denotes the adjoint action on the Lie algebra $\mathfrak{g}_{\max}$.

The curvature 2-form $\Omega = d\Gamma + \frac{1}{2}[\Gamma, \Gamma]$ inherits this equivariance. For any $\theta \in \Theta$ and $g \in G_{\max}$:

$$\Omega_{g\cdot\theta} = g^*\Omega_\theta = \mathrm{Ad}_{g^{-1}}(\Omega_\theta).$$

The Frobenius norm $\|\cdot\|_F$ on the Lie algebra is defined using the Fisher–Rao metric restricted to vertical directions. Since this metric is invariant under the gauge action, the norm is invariant under the adjoint action:

$$\|\mathrm{Ad}_{g^{-1}}(X)\|_F = \|X\|_F \quad \text{for all } X \in \mathfrak{g}_{\max}.$$

Combining these facts, we obtain

$$\kappa_{\ell,i}(g\cdot\theta) = \|\Omega_{\ell,i}(g\cdot\theta)\|_F = \|\mathrm{Ad}_{g^{-1}}(\Omega_{\ell,i}(\theta))\|_F = \|\Omega_{\ell,i}(\theta)\|_F = \kappa_{\ell,i}(\theta).$$

Therefore, $\kappa_{\ell,i}$ is constant on gauge orbits and descends to a well-defined function on the quotient space $\mathcal{F} = \Theta/G_{\max}$. ■

C.2. Completeness of the full bispectrum

After canonicalization, the residual symmetry is the finite group S_h acting by left translation on head indices. It is therefore natural to view a layer's features as a function $f_\ell : S_h \to \mathbb{C}^m$, obtained by stacking per-head statistics. The left action $(\pi \cdot f_\ell)(\sigma) := f_\ell(\pi^{-1}\sigma)$ models head relabeling, so equivalence reduces to recovery of f_ℓ up to left translation.

Statement. For the permutation group S_h, the *full* bispectrum $B(\sigma_1, \sigma_2)$ evaluated on all pairs $(\sigma_1, \sigma_2) \in S_h \times S_h$ is *generically complete* up to left translation: if $f_\ell, g_\ell : S_h \to \mathbb{C}^m$ have identical full bispectra and, for every irreducible representation $\rho \vdash h$, the channel-stacked Fourier block $\widehat{f_\ell}(\rho)$ has full column rank (equivalently, its Gram matrix is nonsingular on its image), then there exists $\pi \in S_h$ with $g_\ell = \pi \cdot f_\ell$. For $h \neq 6$, all automorphisms of S_h are inner, so "up to automorphism" coincides with "up to translation."

Justification (classical finite-group harmonic analysis). Kakarala [8] established that, for complex-valued functions on a finite group, the triple correlation (bispectrum) determines the function up to translation and group automorphism, under a generic nondegeneracy of Fourier blocks. In our vector-valued setting $f_\ell : S_h \to \mathbb{C}^m$, one stacks channels and applies the same argument componentwise in the Fourier domain: if each nontrivial block $\widehat{f_\ell}(\rho)$ has full column rank, the system of bispectral relations can be solved to recover $\{\widehat{f_\ell}(\rho)\}_\rho$ up to simultaneous conjugation by representation matrices of a single group element, i.e., up to left translation of f_ℓ. Since S_h is finite, no further continuous ambiguity arises. See also standard treatments of finite-group invariants in [15].

Remarks. (i) The rank condition is *generic* and is satisfied whenever head features are not concentrated in a single isotypic component (e.g., independent or sufficiently diverse heads). In practice we verify channel whiteness and nontrivial energy in at least one nontrivial isotypic component (Appendix D.3).
(ii) Completeness here refers to the *full* bispectrum over $S_h \times S_h$. For computational efficiency, one might consider selective subsets, though this can sacrifice completeness.
(iii) For $h = 6$ there is an outer automorphism of S_6; the bispectrum is complete up to that automorphism, which still corresponds to a head relabeling at the level of conjugacy classes and does not affect our canonicalized setting.

C.3. Canonicalization algorithm (pseudo-code)

Algorithm 1 Deterministic canonicalization with stable tie-breaking

1: Balance per-head Q/K Gram matrices (whitening).
2: Orthonormalize V basis per head (QR/SVD).
3: Compute head scores; group heads with gaps below τ_{sort}.
4: Break ties by ℓ_1 norm of vectorized V; then lexicographic order of $\text{vec}(W_O)$; then head index.
5: Return permuted and normalized parameters; record residual permutation action domain as S_h.

C.4. Complexity derivations

Canonicalization costs $O(d_k^3 + d_v^3)$ per head due to SVD/QR. Constructing f_ℓ is linear in $h\, d_v\, n$. The full bispectrum requires $O(h! \cdot h^2)$ operations due to the group Fourier transform over all $|S_h|^2$ pairs, becoming computationally prohibitive for $h > 8$. Holonomy around a square loop requires four directional evaluations (forward/backward around the loop) hence approximately four backprops; reporting over m direction pairs scales linearly in m.

Appendix D. Detailed Empirical Validation

Metric conventions. Unless noted otherwise, correlations are Pearson. The coefficient of variation (CV) is defined as $\text{CV}(X) = \text{std}(X)/\text{mean}(X)$. The reported condition number is for the FR mechanical-connection normal equations $M_\theta \Gamma_\theta(\xi) = b_\theta(\xi)$ with $M_\theta = J_\theta^* G_\theta J_\theta$, namely $\text{cond}(M_\theta)$ in the Euclidean operator norm.

D.1. Fisher–Rao Curvature Computation

Directional FR curvature without explicit metric tensors. We compute curvature on the quotient using the discrete holonomy of the Fisher–Rao (FR) mechanical connection rather than Euclidean mixed partials of a scalar loss. Let P_{hor} be the FR-horizontal projector obtained by solving the mechanical-connection normal equations, and let u, v be unit FR-horizontal directions. For a small step $\varepsilon > 0$,

$$\Delta_\square(u, v; \varepsilon) \;=\; P_{\text{hor}}\Big(\nabla_u \nabla_v - \nabla_v \nabla_u\Big)\theta \;=\; \varepsilon^2\, \Omega(u, v) \;+\; O(\varepsilon^3),$$

so that $S(\varepsilon) = \|\Delta_\square(u, v; \varepsilon)\|_{\text{F}}/\varepsilon^2$ is a consistent estimator of $\|\Omega(u, v)\|_{\text{F}}$. We use two-step Richardson extrapolation to cancel the $O(\varepsilon)$ term and report the extrapolated value as

our curvature estimate. This avoids explicit construction of the FR tensor while remaining faithful to the quotient-geometry definition.

D.2. Bound Verification Protocol

Normalization and bound verification protocol. To test Theorem 4, we avoid axis-wise min–max rescaling. For each layer ℓ and direction pair (u, v) we compute the extrapolated curvature magnitude $\widehat{\kappa}^2_\ell(u, v)$ and the directional nontrivial bispectral energy $\widehat{\mathcal{E}}_\ell(u, v)$ in their native units. We then estimate the maximal admissible slope

$$\widehat{c}_\ell \;=\; \min_{(u,v)} \frac{\widehat{\kappa}^2_\ell(u, v)}{\widehat{\mathcal{E}}_\ell(u, v) + \delta}, \qquad \delta = 10^{-10},$$

and verify the bound $\widehat{\kappa}^2_\ell(u, v) \geq \widehat{c}_\ell\, \widehat{\mathcal{E}}_\ell(u, v)$ holds on the vast majority of pairs. For comparability across scales we may multiply both sides by the same positive scalar (e.g., divide by trace of the FR block), which preserves the inequality.

D.3. Bispectral Computation with Equivariance

Equivariant feature map and whitening. Per head i, let $\psi(z_i)$ be the concatenation of head-wise summary statistics (mean, std, range). Define the stacked feature

$$F_\ell = \left[\psi(z_1)^\top \;\cdots\; \psi(z_h)^\top\right]^\top.$$

Under a head permutation $\sigma \in S_h$, F_ℓ transforms by row-permutation. Hence

$$f_\ell(\sigma) = P(\sigma)\, F_\ell$$

with P the permutation representation. Before computing the Fourier blocks and bispectrum we whiten $\psi(z_i)$ across the batch (zero mean, unit covariance per coordinate) to remove amplitude effects that would otherwise confound energy comparisons.

Full bispectrum computation. The full bispectrum requires evaluating $B_\ell(\sigma_1, \sigma_2)$ for all pairs $(\sigma_1, \sigma_2) \in S_h \times S_h$, yielding $(h!)^2$ complex-valued terms. We compute the group Fourier transform $\widehat{f}_\ell(\rho)$ for each irreducible representation $\rho \vdash h$ using the standard character-based projection operators. The computational cost becomes prohibitive for $h > 8$ due to factorial scaling.

Equivariance verification. We verify equivariance through unit tests: for random $\sigma \in S_h$, we confirm

$$\|B_\ell(\text{model}) - B_\ell(\sigma \cdot \text{model})\|_2 < 10^{-12}$$

where $\sigma \cdot \text{model}$ denotes the gauge-transformed parameters and B_ℓ is the full bispectrum. This confirms our implementation correctly respects the permutation symmetry.

D.4. Threats to Validity

Directional FR curvature relies on accurate horizontal projection; numerical error in solving the mechanical-connection normal equations can bias estimates if conditioning is poor. We monitor condition numbers and report convergence diagnostics through Richardson extrapolation ratios.

Feature whitening and the choice of readout ψ affect bispectral magnitudes; we mitigate by equivariance tests and by focusing on discrimination metrics rather than absolute energies.

Bound verification is sensitive to axis rescaling; our slope-based procedure uses native units and equal scaling on both sides to avoid spurious violations.

The full bispectrum's factorial complexity limits its practical application to small head counts ($h \leq 8$). For larger models, we rely primarily on Fisher-Rao curvature as the computationally tractable invariant. The correspondence theorem validates that geometric invariants capture the essential algebraic structure, though efficient approximations to the full bispectrum remain an open challenge.

D.5. Complete Statistical Results

The Richardson ratio reported is $|K(2\varepsilon) - 2K(\varepsilon) + K(\varepsilon/2)|/K(\varepsilon)$; values near 1 indicate stable extrapolation in our step schedule. The condition number is $\mathrm{cond}(M_\theta)$ from the mechanical-connection linear system.

Table 3: Complete multi-scale results with FR curvature and whitened bispectral energy.

h	FR Curvature $\|\Omega\|_F^2$		Bispectral $\mathcal{E}$		Richardson	Condition
	Mean	CV	Mean	CV	Ratio	Number
4	0.058	0.48	17.9	0.36	0.97	3.2×10^3
6	0.066	0.49	26.4	0.33	0.96	4.8×10^3
8	0.072	0.49	35.8	0.28	0.98	6.1×10^3
12	0.081	0.48	52.7	0.25	0.98	9.4×10^3
16	0.089	0.46	64.3	0.25	0.97	1.3×10^4
24	0.097	0.45	76.1	0.24	0.96	2.1×10^4

The Richardson ratios near 0.97 confirm convergence of the curvature estimator. Condition numbers increase with model scale but remain tractable for the iterative solver.

Appendix E. Implementation Methodology and Numerical Validation

This section provides comprehensive implementation details for computing geometric and algebraic invariants, including numerical safeguards and validation procedures for practical deployment of our theoretical framework.

E.1. Fisher–Rao Geometric Computations

The computation of Fisher–Rao curvature on the quotient manifold requires careful numerical treatment due to the high-dimensional parameter space and the need to solve the mechanical connection normal equations. We implement a multi-stage approach that balances computational efficiency with numerical stability.

For the mechanical connection computation, we solve the linear system $M_\theta \Gamma_\theta(\xi) = b_\theta(\xi)$ where $M_\theta = J_\theta^* G_\theta J_\theta$ represents the pullback of the Fisher–Rao metric to the Lie algebra. The condition number of this system, reported in Table 3, ranges from 3.2×10^3 for 4-head models to 2.1×10^4 for 24-head configurations. We employ an iterative conjugate gradient solver with Jacobi preconditioning, achieving convergence tolerance of 10^{-10} within 50 iterations for all tested configurations.

The discrete holonomy estimator implements Richardson extrapolation with adaptive step sizing. For each directional pair (u, v), we compute $\Delta_{\square}^{\ell}(u, v; \varepsilon)$ at step sizes $\varepsilon_1 = 10^{-4}$ and $\varepsilon_2 = 2 \times 10^{-4}$, chosen to balance truncation and roundoff errors in float64 arithmetic. The stability of the extrapolation is monitored through the Richardson ratio $|K(2\varepsilon) - 2K(\varepsilon) + K(\varepsilon/2)|/K(\varepsilon)$, with values near 0.97 indicating stable convergence as confirmed in our experiments.

E.2. Canonicalization and Residual Symmetry

The canonicalization procedure must be both deterministic and numerically stable to ensure consistent computation of algebraic invariants. Our implementation follows a four-stage process with explicit tolerance parameters:

1. **Query-Key Balancing**: For each head i, compute the Gram matrices $G_Q^{(i)} = (W_Q^{(i)})^T W_Q^{(i)}$ and $G_K^{(i)} = (W_K^{(i)})^T W_K^{(i)}$. Apply simultaneous diagonalization via generalized eigendecomposition with regularization parameter $\epsilon = 10^{-12}$ to prevent numerical instability for near-singular configurations.

2. **Value Orthonormalization**: Apply QR decomposition to each $W_V^{(i)}$ with column pivoting for numerical stability. The resulting orthonormal basis is unique up to column signs, which we fix by requiring the first non-zero element of each column to be positive.

3. **Head Sorting with Stable Tie-Breaking**: Compute sorting metrics s_i for each head based on the Frobenius norm of the value projection. When $|s_i - s_j| < \tau_{\text{sort}} = 10^{-6}$, apply the hierarchical tie-breaking procedure: first by ℓ_1 norm of vectorized V-basis (10^{-9} tolerance), then by lexicographic ordering of $\text{vec}(W_O)$ elements, finally by original head index.

4. **Permutation Tracking**: Record the applied permutation $\sigma \in S_h$ to enable consistent application across all parameter tensors and maintain the relationship between geometric and algebraic computations.

E.3. Validation Protocols

Each implementation component undergoes systematic validation:

Gauge Equivariance Testing: For 100 random gauge transformations $g \in G_{\max}$, verify that computed invariants satisfy $|\kappa(g \cdot \theta) - \kappa(\theta)| < 10^{-10}$ and $\|B(g \cdot \theta) - B(\theta)\|_2 < 10^{-10}$, where B denotes the full bispectrum.

Linearization Index Monitoring: For each direction pair (u, v), compute $\eta(u, v) = |K(2\varepsilon) - 2K(\varepsilon) + K(\varepsilon/2)|/K(\varepsilon)$ and flag cases where $\eta > 0.1$ as potentially violating linearization assumptions.

Numerical Conditioning Assessment: Track condition numbers for the mechanical connection system, canonicalization transformations, and feature whitening matrices. Report warnings when condition numbers exceed 10^6.

Cross-validation with Finite Differences: For a subset of 10 direction pairs per experiment, compare discrete holonomy estimates against high-precision finite difference approximations using step size $h = 10^{-8}$, requiring agreement within relative tolerance 10^{-3}.

E.4. Extended Experimental Protocols

This subsection provides detailed experimental procedures that extend beyond the summary presented in the main text, enabling complete reproduction of our empirical results.

Dataset and Batch Configuration. All experiments utilize a fixed evaluation batch constructed to ensure numerical stability and representative coverage of the attention mechanism's operating regime. The batch consists of 256 sequences of length 128, drawn from the OpenWebText tokenized corpus with the following specifications. The tokenization employs a byte-pair encoding vocabulary of 50,257 tokens, with special tokens for padding, beginning-of-sequence, and end-of-sequence markers. Sequences are selected to maintain diverse linguistic patterns with 40% containing technical content, 30% conversational text, and 30% formal prose. This distribution ensures that attention patterns encounter varied semantic relationships during evaluation.

Input embeddings are initialized using Xavier uniform initialization scaled by $\sqrt{d_{\text{model}}}$, with positional encodings following the standard sinusoidal pattern when rotary embeddings are not employed. Layer normalization parameters are initialized with unit gain and zero bias, while attention temperature is fixed at $1/\sqrt{d_k}$ across all experiments.

Model Architecture Specifications. For the multi-scale validation experiments, we systematically vary the number of heads while maintaining proportional scaling of model dimensions. Each configuration maintains fixed query/key dimension $d_k = 64$ and value dimension $d_v = 64$ to isolate the effects of head count on the geometric-algebraic correspondence. The output projection dimension equals $d_{\text{model}} = h \cdot d_v$ to satisfy the architectural constraint for residual connections. Memory requirements scale from 2.1 MB for 4-head models to 75.5 MB for 24-head configurations, enabling single-GPU execution for all experiments.

Training Dynamics Monitoring. The training stability analysis tracks model evolution through 10,000 gradient steps using AdamW optimization with carefully tuned hyperparameters. The learning rate schedule implements linear warmup over 500 steps to peak learning rate 5×10^{-4}, followed by cosine annealing to 10^{-5}. Weight decay of 0.1 applies to all parameters except layer normalization and bias terms. Gradient clipping at norm 1.0 prevents instability during early training when attention patterns are uninformative.

Checkpoints are saved at exponentially spaced intervals corresponding to steps 0, 100, 500, 1000, 2500, 5000, and 10000. At each checkpoint, we compute both geometric and algebraic invariants on the fixed evaluation batch, ensuring consistent measurement conditions across training. The same 30 random direction pairs are used at each checkpoint to enable direct comparison of invariant evolution.

Statistical Analysis and Reporting. All reported statistics aggregate over multiple sources of variation to ensure robust conclusions. Direction sampling employs stratified random selection to ensure coverage of the parameter manifold. For each layer, we generate 30 direction pairs by sampling 15 pairs from the column space of the Jacobian at randomly selected training points and sampling 15 pairs from random Gaussian directions orthogonalized via Gram-Schmidt.

Confidence intervals are computed using bootstrap resampling with 1000 iterations, reporting the 2.5 and 97.5 percentiles. For bounded quantities like correspondence rates, we apply logit transformation before bootstrapping to avoid boundary effects. The Pearson correlation coefficients reported between geometric and algebraic invariants use Fisher z-

transformation for confidence interval construction, with significance testing via permutation tests using 10,000 permutations to avoid parametric assumptions.

E.5. Computational Complexity Analysis

This subsection provides detailed complexity analysis for all algorithmic components with concrete runtime measurements and scalability considerations.

Theoretical Complexity Bounds. The asymptotic complexity of each operation determines the scalability limits of our approach. For canonicalization operations, Gram matrix construction requires $O(h \cdot d_k^2 \cdot d_{\text{model}})$ time with $O(h \cdot d_k^2)$ space, while generalized eigendecomposition scales as $O(h \cdot d_k^3)$ with $O(d_k^2)$ space per head. The QR decomposition for value orthonormalization requires $O(d_{\text{model}} \cdot d_v^2)$ time and $O(d_{\text{model}} \cdot d_v)$ space. Stable sorting with tie-breaking combines $O(h \log h)$ comparison operations with $O(h \cdot d_v^2)$ tie-breaking computations.

For geometric computations, Fisher-Rao metric evaluation scales with the cost of a forward pass through the network, typically $O(n \cdot d_{\text{model}}^2)$ for sequence length n. The mechanical connection solve requires $O(h^2 \cdot (d_k^2 + d_v^2)^{3/2})$ time due to the iterative solver, with space complexity $O(h^2 \cdot (d_k^2 + d_v^2))$ for storing the system matrix. Each discrete holonomy computation requires four backpropagation passes, yielding $O(4 \cdot \text{Backprop})$ time complexity.

The algebraic computations exhibit different scaling characteristics. Feature extraction is linear in the number of parameters, requiring $O(h \cdot n \cdot d_v)$ time with $O(h \cdot m_f)$ space for m_f features per head. Feature whitening involves covariance computation and matrix inversion, scaling as $O(m_f^2 \cdot n + m_f^3)$ with $O(m_f^2)$ space. The full group Fourier transform has factorial complexity $O(h! \cdot h^2)$, which limits practical application to $h \leq 8$.

Empirical Runtime Measurements. Beyond asymptotic analysis, we provide empirical runtime measurements on NVIDIA H100 GPUs to characterize real-world performance. Full canonicalization ranges from 3.2 milliseconds for 4-head models to 58.2 milliseconds for 24-head configurations. Single Fisher-Rao curvature computation scales from 42 milliseconds to 367 milliseconds across the same range. The full bispectrum computation time grows from 45.6 milliseconds for $h = 4$ to over 8 seconds for $h = 24$, confirming the factorial scaling limitation. Mechanical connection solve times range from 8.3 to 112.3 milliseconds.

Memory Requirements and Optimization. Memory consumption becomes a critical constraint for large models, particularly when computing Fisher-Rao metrics that require storing gradient information for all parameters. Our implementation employs several optimization strategies to manage memory efficiently.

Gradient checkpointing reduces memory requirements by recomputing intermediate activations during backpropagation, trading a 40% increase in computation time for 60% reduction in peak memory usage. This trade-off enables analysis of models with up to 48 heads on single H100 GPUs with 80GB memory. Batch-wise accumulation of Fisher-Rao products avoids materializing the full metric tensor, instead computing projections $G_\theta v$ for specific directions v. This reduces memory complexity from $O(p^2)$ where p is parameter count to $O(p)$, enabling scaling to production models.

Mixed precision computation using float32 for invariant calculations and float64 only for critical numerical operations such as canonicalization and mechanical connection solving provides $2\times$ memory savings with negligible impact on correspondence validity. Our experi-

ments show 98.7% correspondence rate with mixed precision versus 98.9% with full float64 precision.

Parallelization Opportunities. The computational structure admits several parallelization strategies that can significantly reduce wall-clock time for large-scale analyses. Head-level parallelism enables independent processing of per-head canonicalization and feature extraction across GPU streaming multiprocessors. This achieves near-linear speedup up to the number of heads, limited primarily by memory bandwidth for gradient accumulation.

Direction-level parallelism allows concurrent evaluation of curvature for multiple direction pairs, with each pair requiring independent forward-backward passes. This embarrassingly parallel structure scales effectively across multiple GPUs for large-scale invariant analysis. In our experiments, processing 30 direction pairs across 8 GPUs reduces total computation time by a factor of 7.2, with the sub-linear scaling due to communication overhead in gradient synchronization.

Batch-level parallelism in Fisher-Rao computation distributes evaluation examples across devices, with gradient accumulation via distributed reduction. This approach scales to batch sizes of 4,096 samples using 16 GPUs with gradient accumulation over micro-batches of 256 samples per device. The resulting system achieves 12.3× speedup compared to single-GPU execution, limited by the all-reduce communication pattern required for gradient aggregation.