BEYOND ROUTING: MATRIX ZONOTOPIC ATTENTION FOR SET TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Standard attention adapts token routing to the input while sharing its value projections across contexts. We study a complementary design axis for set learning: making feature transformations explicitly context-dependent and carrying transformation-derived features across layers. We introduce Matrix Zonotopic Attention (MZAttn), which parameterizes a transformation family using a learned center matrix and generator matrices modulated by bounded, permutation-invariant context gates. Rather than using this family only to produce a single-vector update, MZAttn propagates separate center and generator feature streams and connects them to prediction through a generator-aware readout. To characterize operator variation, we introduce Transformation Degrees of Freedom (TDOF), which counts the independent matrix directions along which a specified operator varies with context. At the attention-core level, we establish exact representations for specified gated operator families and quantify approximation limits within fixed matrix-bank classes. We further show that identical signed and magnitude summaries can become distinguishable after the same rowwise normalization. For the normalized magnitude readout, we characterize local information capacity and construct a complementary two-row code supporting stable local recovery and scalar linear prediction through a finite head, with an explicit approximation bound. The resulting architecture preserves the permutation structure required for set prediction. We investigate its inductive bias through set-learning experiments. Together, these contributions frame set attention around three complementary choices: where to route information, how to transform it, and what transformation information to retain.

1 INTRODUCTION

When can a set-attention model replace feature rows by a summary without changing what later layers can compute? A prediction head may use only signed sums or sums of magnitudes, while a subsequent operation depends on each row's internal structure. A summary sufficient for a present prediction can therefore be insufficient for continued computation.

Our answer is to retain transformation-derived rows through further processing, and analyze their eventual readout separately. *Matrix Zonotopic Attention* (MZAttn) implements this choice within a Set Transformer (Lee et al., 2019). Each token carries a center and *generator rows*, separately stored feature vectors. Centers determine routing and permutation-invariant context for bounded tanh gates. A learned center matrix and gated generator matrices define the transformation family. Mixed gated actions and inherited rows pass through residual, normalization, and feed-forward processing. The head reads pooled centers and the *interval-hull half-width* (IHW), the coordinate-wise sum of absolute rows. Matrix zonotopes supply the center–generator organization (Alanwar et al., 2021), not a certification guarantee for these nonlinear features. Conditional weights and multistream states already exist (Schug et al., 2025; Zhu et al., 2025); our analysis asks what this state–readout combination preserves.

We first prove that even joint signed and magnitude summaries can fail to support continued computation. With at least two rows of width at least three, strictly positive, nonconstant rows can share both summaries and their centers but separate after the same repeated rowwise LayerNorm with unit gains and zero biases, for any finite nonnegative epsilon, the normalization stabilizer. The sum-

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

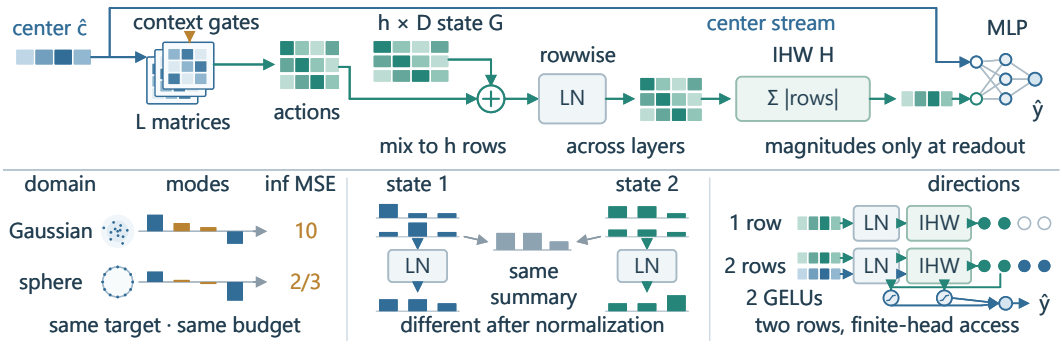

Figure 1: **Transformation, state, and readout.** Top: gated actions persist as generator rows before IHW readout. Bottom: input laws change quadratic prediction budgets (left); equal signed and magnitude summaries separate after rowwise normalization (middle); complementary rows support local recovery and finite-head prediction (right). Direction markers depict width four at zero epsilon. These are separate theoretical constructions under their stated assumptions, not empirical comparisons.

maries do not determine each row's mean and variance. No decoder of the original summaries can reproduce both continued outputs. This is a fixed-continuation obstruction, not a lower bound for an encoder redesigned from the original input (Figure 1, middle).

For repeated rowwise normalization followed by IHW, we characterize how many latent coordinates a smooth local decoder can recover. We analyze nonconstant rows with nonzero coordinates before absolute value, without informative center-side features or post-IHW normalization. At zero epsilon, one row loses two local directions; two complementary rows attain the full readout dimension for widths of at least four. Positive epsilon restores a radial direction, but at minimum single-row recovery width every smooth inverse becomes ill-conditioned as epsilon vanishes, if encoding and post-normalization derivatives stay uniformly bounded and row variance stays bounded away from zero.

A constructive result makes retained information accessible without relying on this weak direction. For widths divisible by four, we give complementary affine row encodings under unit-gain, zero-bias normalization. After an affine rescaling absorbed by the head, their magnitude readout is well-conditioned on a fixed Euclidean neighborhood whose radius does not shrink with width, uniformly as epsilon vanishes. An affine IHW head, exactly implemented by two GELU units and an output bias, approximates scalar linear latent targets. For bounded latent norms and coefficients, its leading error bound decays as the inverse square root of width.

The gate budget for scalar prediction is a separate question. For quadratic scalar targets, we derive exact mean-squared-error infima at every gate count over sums of linear features multiplied by tanh gates with affine arguments, plus a free affine term. Gaussian inputs impose separate positive and negative spectral budgets. On a uniform sphere, squared input norm is constant, changing which quadratic variations require gates. The same target can therefore need different gate budgets. Matching lower bounds and limiting constructions apply to the nonlinear class without uniform parameter-norm bounds, not only to a polynomial surrogate. Suitable gated linear unit (GLU) variants share the curves.

The results characterize specified cores and local constructions, not what training necessarily learns. They distinguish preserving a prediction, preserving continued computation, and providing stable access to retained information.

2 Related Work

Conditional transformations and structured states. CondConv mixes a fixed kernel bank using input-dependent coefficients (Yang et al., 2019); MZAttn's tanh bank is affinely reparameterizable as a sigmoid-weighted bank. Schug et al. (2025) interpret attention as configuring key–query-specific

value networks. Hyper-Connections retains interacting residual streams (Zhu et al., 2025), and Residual Matrix Transformers use keyed outer-product memories (Mak & Flanigan, 2025). These works supply conditional weights and structured storage. MZAttn combines center-controlled set routing, separate matrix-bank and stored-row budgets, and rowwise processing before IHW. Our guarantees concern these interfaces, not a comparison of the unrestricted architectures' expressivity.

Reduction and recovery. CLUE constructs linear reductions of polynomial differential equations that preserve selected observables and admit closed evolution (Ovchinnikov et al., 2021). We use the same consistency requirement, but establish its failure for joint signed/IHW summaries under a neural continuation. LayerNorm's projection and scaling geometry is established (Brody et al., 2023). Bruna et al. (2014) give stability bounds, modulo global sign, for pooled linear measurements including magnitude sums; Haider et al. (2023) reconstruct ReLU measurements locally with dual frames. Repeated rowwise normalization defines a different measurement map: fixing signs does not remove input-dependent rescaling. Complementary local directions must survive summation, and local rank alone does not provide a width-independent recovery neighborhood. Our construction controls both this geometry and the normalization remainder, while its finite-head guarantee concerns approximate scalar prediction rather than exact global inversion.

Fixed-budget gated approximation. Pearce et al. (2025) analyze a given bilinear MLP's spectrum, while Belrose & Rigg (2025) derive Gaussian least-squares polynomial approximations of given MLPs and GLUs. We instead fix a quadratic target and optimize over the entire gated-linear class at a specified gate count, including gate biases. Polynomial projection is inherited; the additional steps constrain every gate's positive and negative quadratic directions and supply matching sequences within the nonlinear class. On the sphere, equivalence up to an identity shift changes the gate budget, not merely a constant in the Gaussian estimate.

3 Matrix Zonotopic Attention

3.1 Representations and shared routing

A vector zonotope is $\{c + \sum_{g=1}^{h} \alpha_g G_g : |\alpha_g| \le 1\}$, with interval-hull half-width $\delta = \sum_g |G_g|$. A matrix zonotope is

$$\mathcal{M} = \left\{ M_c + \sum_{l=1}^{L} \xi_l M_l : |\xi_l| \le 1 \right\}. \tag{1}$$

The bank size L counts available matrix directions, while h controls the number of feature rows stored per token. Each row has width D; d is the input dimension and d_k the per-head width. Generator rows are components of a token's state, not additional set elements. For token x_i, the embedding is

$$c_i = \mathrm{LN}(\mathrm{GELU}(W_c x_i + b_c)), \qquad G_i = \frac{\mathrm{reshape}(W_g x_i + b_g)}{\sqrt{hD}} \in \mathbb{R}^{h \times D}. \tag{2}$$

Shared projections on centers give $\alpha_{ij} = \mathrm{softmax}_j(q_i^\top k_j / \sqrt{d_k})$. Both streams use these weights: $\hat{c}_i = \sum_j \alpha_{ij} v_j^c$ and $\hat{G}_i = \sum_j \alpha_{ij} v_j^G$. The routing is shared, but the center and row features are projected separately and remain distinct.

3.2 Conditional actions and persistent state

Suppressing head indices, the update is

$$o_i^c = M_c \hat{c}_i, \qquad o_i^G = \hat{G}_i M_c^\top + W_{\mathrm{mix}}\, \mathrm{stack}_{l=1}^{L}[\gamma_l(\mu) M_l \hat{c}_i], \tag{3}$$

with $W_{\mathrm{mix}} \in \mathbb{R}^{h \times L}$ and $\gamma = \tanh(W_\gamma \mu + b_\gamma)$. The key/value-center mean μ is permutation-invariant. The first generator term transforms inherited features; the second injects new gated matrix actions. Mixing maps the L actions into the same h rows, keeping storage fixed across layers while allowing L and h to vary separately. The update omits the product's cross terms: it defines learned features, not a guaranteed set enclosure (Appendix A.1).

3.3 Complete architecture and analyzed interfaces

Token embedding is followed by an ISAB encoder, PMA pooling, an SAB decoder and a prediction head, using MZ attention blocks throughout. In each block, attention outputs are added to the query states and LayerNorm is applied separately to centers and generator rows. Each stream then passes through its own two-layer GELU FFN, followed by residual addition and a second LayerNorm. The final head applies a two-layer GELU MLP to concatenated pooled centers and IHW, without post-IHW normalization in the default model. Shared tokenwise operations and invariant gates preserve set permutation structure (Proposition A.2). Centers receive no generator feedback; a center-only head cannot train generator-only parameters through the data loss (Proposition A.3). Initialization uses $M_c = I$ and small nonzero M_l.

The analysis isolates three interfaces within this computation. Sections 4.1–4.2 use a gated core with signed/affine readout to study transformations and scalar prediction. Section 4.3 fixes a continuation and compares intermediate states with equal current summaries. Section 4.4 studies repeated row-wise normalization followed by summed magnitudes. Section 5 distinguishes training the isolated function class from evaluating complete models with routing, FFNs and the actual prediction head.

4 Theoretical Analysis

We follow transformation features from their generation, through continued computation, to prediction. The comparisons below concern the stated interfaces and retain their separate assumptions.

4.1 What one gated core adds to attention

For a specified operator T, $\mathrm{TDOF}(T) = \dim \mathrm{span}\{T(c) - T(c') : c, c' \in \mathcal{C}\}$ counts its context-dependent matrix directions (Definition B.1); exact representation by a fixed bank needs at least this many directions. We analyze an isolated MZ core with uniform aggregation, zero incoming generators, one stored slot and a signed readout (Theorem B.2).

Theorem 4.1 (Separation for every nonzero matrix). For any nonzero $A \in \mathbb{R}^{d\times d}$, one such core with $L = 1$ represents $f_A(S) = [I + \tanh(\mu(S)_1)A]\mu(S)$ exactly. No attention block with affine query, key, value and output maps and softmax aggregation represents f_A on all constant-token sets: on such sets the block is affine in the token, whereas f_A is not.

Theorem 4.2 (ReLU width and depth cost). Let $Ae_1 \neq 0$, $0 < a \leq 1/2$, $u = Ae_1/\|Ae_1\|_2$ and $g(t) = u^\top f_A(\{te_1, \dots, te_1\})$. A self-attention stack on identical tokens with affine maps, residual additions and D ReLU FFNs of widths $w_1, \dots, w_D$, without LayerNorm, with mean pooling and a scalar affine output, that approximates g within ϵ uniformly on $[-a, a]$ satisfies

$$\epsilon \geq \frac{\kappa a^2}{4\prod_{t=1}^{D}(w_t+1)^2}, \qquad \kappa = 2\|Ae_1\|_2 \operatorname{sech}^2(a)(1 - a\tanh a) > 0.$$

One hidden layer therefore needs width of order $\epsilon^{-1/2}$, the rate piecewise-linear interpolation attains, whereas one gate is exact.

Theorem 4.3 (Exact depth–direction tradeoff). Let $\mu = (c, z) \sim \mathcal{N}(0, I_d)$ and $T_*(c) = I + \sum_{l=1}^{k} \tanh(c_l) e_l e_d^\top$ with $k < d$. Among residual stacks $T_D(c) = \prod_{t=1}^{D}[I + A_t(c)]$ whose layers combine q_t fixed directions from the square-zero algebra $\{ue_d^\top : u \perp e_d\}$ with square-integrable context coefficients, with no intervening maps and $Q = \sum_t q_t$,

$$\inf_{T_D} \mathbb{E}\|[T_*(c) - T_D(c)]\mu\|_2^2 = \nu\,(k - \min\{k, Q\}), \qquad \nu = \mathbb{E}_{Z\sim\mathcal{N}(0,1)} \tanh^2 Z > 0.$$

Exact representation thus needs $Q \geq k = \mathrm{TDOF}(T_*)$, and $D \geq \lceil k/q \rceil$ layers of at most q directions each, while one native core with $L = k$ is exact. All products in the algebra vanish, so a stack only adds its directions; outside the algebra this collapse fails.

Corollary 4.4 (Local approximation of any smooth operator). Let $T : \mathbb{R}^d \to \mathbb{R}^{m\times d}$ be twice continuously differentiable near the ball K of radius ρ around μ_0, with Hessian bound C_T on K and Jacobian $\partial \operatorname{vec} T(\mu_0) = \sum_{l=1}^{r} s_l u_l v_l^\top$ of rank r. For every $\tau > 0$, the native bank $T_\tau(\mu) = T(\mu_0) + \sum_{l=1}^{r} \tanh(\tau v_l^\top(\mu - \mu_0))\, \tau^{-1} s_l \operatorname{mat}(u_l)$ matches $T(\mu_0)$ and its Jacobian exactly, and on K

$$\|[T(\mu) - T_\tau(\mu)]\mu\|_2 \le \left[\tfrac{1}{2} C_T \|\mu - \mu_0\|_2^2 + \tfrac{\tau^2}{3} \sum_{l=1}^{r} s_l |v_l^\top(\mu - \mu_0)|^3\right] \|\mu\|_2.$$

Any bank with L directions has Jacobian rank at most L, so matching this one needs $L \ge r$.

On an identifiable family of banks, any parameterization whose outputs are locally Lipschitz in its P parameters needs $P \ge (L+1)md$ to represent every member exactly, which the native bank attains, and b-bit parameters need $Pb \ge (L+1)md \log_2(B/\epsilon)$ to approximate every member within $\epsilon \in (0, B)$ (Theorem B.4). These comparisons exclude normalization and, except for the ReLU result, FFNs; they are not lower bounds for general Transformers. Appendix B gives the proofs.

4.2 Scalar prediction depends on the input domain

For a square-integrable operator $T(C)$ and value probes satisfying $\mathbb{E}[VV^\top \mid C] = \Sigma \succ 0$ with fixed Σ, Theorem C.2 gives a covariance spectral-tail optimum over affine L-direction banks with unrestricted square-integrable context coefficients. This optimum is a lower bound for native affine–tanh banks; equality requires additional structure, as in Corollary C.3. Scalar prediction exposes different quantities, which we characterize below (Figure 1, left).

Let $f_Q(x) = c_* + a_*^\top x + x^\top Q x$, $Q = Q^\top$, and define

$$\mathcal{G}_L = \left\{ c + a^\top x + \sum_{l=1}^{L} (b_l^\top x) \tanh(w_l^\top x + \theta_l) \right\}.$$

A native isolated signed MZ core with an affine scalar readout realizes every member of $\mathcal{G}_L$; suitable GLU and conditional-weight models realize it too. Parameters are finite for each function, with no uniform norm budget. Write $\mathcal{E}_L^P(Q) = \inf_{g\in\mathcal{G}_L} \mathbb{E}_P[(f_Q - g)^2]$.

Theorem 4.5 (Gaussian curve). For $X \sim \mathcal{N}(0, I_d)$, let α_j be the decreasing positive eigenvalues of Q and β_j the decreasing magnitudes of its negative eigenvalues, with zero padding. Then

$$\mathcal{E}_L^{\text{Gaussian}}(Q) = 2 \sum_{j>L} (\alpha_j^2 + \beta_j^2). \tag{4}$$

The zero-infimum gate count is $\max\{n_+(Q), n_-(Q)\}$. For $\mathcal{N}(0, \Sigma)$, replace Q by $\Sigma^{1/2} Q \Sigma^{1/2}$.

Proof idea. The quadratic Gaussian projection of one gated-linear term has at most one eigenvalue of each sign, even with gate bias. Its quadratic form vanishes on the hyperplane orthogonal to the gate direction, which any two-dimensional definite subspace would intersect. With L gates, the positive and negative inertia are each at most L; orthogonal projection and spectral approximation give the two residual tails. The free affine terms fit the target's affine part. For the remaining centered quadratic component, $\mathbb{E}[(X^\top C X - \operatorname{tr} C)^2] = 2\|C\|_F^2$ for symmetric C explains the factor two. Pairing positive and negative modes and taking $\tau^{-1}\tanh(\tau a^\top x) b^\top x \to (a^\top x)(b^\top x)$ approaches the bound. Appendix C.2 supplies the complete proof and finite-τ error. This is an infimum over finite-parameter functions without a uniform norm budget, not a training convergence guarantee.

Theorem 4.6 (Spherical curve). Let X be uniform on $r\mathbb{S}^{d-1}$, $d \ge 2$, $r > 0$, and let $\lambda_1 \ge \cdots \ge \lambda_d$ be the eigenvalues of Q. If $2L < d$, let $I_L = \{L+1, \ldots, d-L\}$ and $\bar{\lambda}_L$ be their mean. Then

$$\mathcal{E}_L^{\text{sphere}}(Q) = \frac{2r^4}{d(d+2)} \sum_{i\in I_L} (\lambda_i - \bar{\lambda}_L)^2. \tag{5}$$

For $2L \ge d$ the infimum is zero. The least gate count for zero infimum is $\min_t \max\{n_+(Q - tI), n_-(Q - tI)\}$, equal to $\lfloor d/2 \rfloor$ for a simple spectrum.

Why the answer changes. On the sphere, $x^\top(tI)x = tr^2$ is constant, so quadratic information is represented modulo the identity. A gate's spherical quadratic projection has at most one direction of

each sign after an identity shift. Optimizing the shared shift lets the best L terms retain the outer L eigenvalues on each side and approximate the middle ones by their mean. Appendix C.3 gives the proof and numerical example.

Zero-epsilon LayerNorm of isotropic Gaussian features gives a uniform sphere in the centered subspace. Both curves extend to globally Lipschitz gates with a nonzero derivative at some point, including suitable GLU variants (Corollary C.8). The input law is substantive: fixed radius or rotational symmetry alone does not suffice for these curves (Appendix C.3.5).

4.3 A READABLE SNAPSHOT NEED NOT SUPPORT CONTINUED COMPUTATION

Appendix D.1 characterizes what fixed mixing preserves. Continued computation raises a further question. Let $S(G) = \sum_i G_{i,:}$ and $H(G) = \sum_i |G_{i,:}|$. A reduction R can replace a state under a fixed continuation F exactly when equal reduced states remain equal after continuing: $R(s) = R(t) \Rightarrow R(F(s)) = R(F(t))$. This standard exact-reduction criterion (Ovchinnikov et al., 2021) does not generally hold for the current summaries S, H.

Theorem 4.7 (Joint-summary continuation obstruction). For every $h \geq 2$, $D \geq 3$, finite $\epsilon \geq 0$ and fixed $s \geq 1$ rowwise LayerNorms with unit gains and zero biases, there are strictly positive, nonconstant rows with equal centers, S and H, but different continued IHW. At $h = 2, D = 3$, use

$$G = \begin{pmatrix} 3 & 1 & 1 \\ 1 & 3 & 1 \end{pmatrix}, \qquad \widetilde{G} = \begin{pmatrix} 5/2 & 5/2 & 1 \\ 3/2 & 3/2 & 1 \end{pmatrix}.$$

Both original summaries are $(4, 4, 2)$. With $A_s = \sum_{j=0}^{s-1} \epsilon^j$, the scalar readout $\ell(z) = 2z_1 - z_3$ after continuation differs by $\Delta = 8/[3\sqrt{(8/9)A_s + \epsilon^s}] > 0$. Any predictor of this continued target using only the original shared summaries has worst-case absolute error at least $\Delta/2$.

The missing information is the allocation of coordinates across rows, not merely the sign of a summed action. Appendix D.2 proves the arbitrary-width construction and its realization by an affine generator encoding with a constant center and a residual/LN subpath; attention and FFN contributions can be zero for this parameter choice. A different end-to-end encoder with access to the original input is not subject to this two-state obstruction. Separately, Proposition D.3 characterizes which linear reductions preserve coordinatewise nonlinear updates: for continuously differentiable ϕ with nonconstant derivative, a reduction P admits an update Ψ with $P\phi(x) = \Psi(Px)$ for all x exactly when the nonzero columns of P are linearly independent. In particular, a signed row sum alone does not determine the sum after rowwise GELU.

4.4 NORMALIZED MULTIROW CAPACITY AND FINITE-HEAD ACCESS

Write $P_D = I_D - \mathbf{1}\mathbf{1}^\top/D$ and $N_\epsilon(x) = P_D x/\sqrt{\|P_D x\|^2/D + \epsilon}$. For row encodings $E_i(q)$ of a latent $q \in \mathbb{R}^m$, the readout feature is $\Phi_\epsilon(q) = \sum_{i=1}^{h} |N_\epsilon^{\circ s}(E_i(q))|$. The following results have no informative center side channel or post-IHW normalization. Row variances and coordinates entering absolute value are nonzero at the analyzed points.

Theorem 4.8 (Sharp local dimension). At zero epsilon, the maximum smoothly recoverable latent dimension is

$$m_{\max}(D, h) = \min\{D, h(D-2)\}. \tag{6}$$

Affine encodings attain the bound locally. Each normalized row has $D - 2$ tangent directions; distinct sign patterns can make their images after absolute value complementary. Appendix E.1 gives constructions and full proofs. Thus two rows suffice at $D \geq 4$, whereas one row alone loses two dimensions. This identifies information capacity, not the complexity of a decoder realizing it.

Theorem 4.9 (One row: injective but ill-conditioned). For $\epsilon > 0$, the minimum single-row recovery width is $D = m + 1$ (at zero epsilon, $m + 2$); Appendix E.2 gives the affine encoding, carrier and explicit inverse. At this minimum width $D = m+1$, suppose the differentiable encoding and post-normalization map have Jacobian norms at most $K > 0$ and $G > 0$, and the encoded pre-normalization variance is at least $v_0 > 0$. Then every differentiable local left inverse $\mathcal{D}_\epsilon$ satisfies

$$\|D\mathcal{D}_\epsilon\|_2 \;\geq\; \frac{(A_s v_0 + \epsilon^s)^{3/2}}{GK\epsilon^s}. \tag{7}$$

Why conditioning deteriorates. Write $a^2 = A_s v + \epsilon^s$, with row variance v. Repeated normalization responds with eigenvalue $1/a$ in the $D-2$ centered tangential directions, but only ϵ^s/a^3 radially. At $D = m+1$, the tangent map has rank at most $m-1$, so some input direction can pass only through this weak radial response. Any local inverse must amplify it as in equation 7.

Theorem 4.10 (Uniform two-row access with a finite head). For $D \in 4\mathbb{N}$, form z_1, z_2 by repeating the centered, unit-variance patterns $\sqrt{2/5}(2,1,-1,-2)$ and $\sqrt{2/5}(1,-2,2,-1)$. Let

$$S_i = \operatorname{diag}(\operatorname{sign} z_i), \quad P_i = I_D - \mathbf{1}\mathbf{1}^\top/D - z_i z_i^\top/D, \quad T = [S_1 P_1 \;\; S_2 P_2], \quad B = T^\top (TT^\top)^{-1}.$$

Partition B into blocks $B_1, B_2 \in \mathbb{R}^{D\times D}$. Then $\|B\|_2 = \sqrt{5}$, and $E_i(q) = z_i + B_i q$ has variance at least one on $\|q\| \le 1/10$. For any fixed $s \ge 1$, $\epsilon \in [0,1]$, set $a_0 = \sqrt{\sum_{j=0}^{s} \epsilon^j}$ and $\delta_0 = |z_1| + |z_2|$. The affine decoder $\Psi_\epsilon = a_0 \Phi_\epsilon - \delta_0$ satisfies

$$\|\Psi_\epsilon(q) - q\| \le \frac{5\|q\|^2}{2\sqrt{D}} + \frac{5\sqrt{5}\|q\|^3}{2D}. \tag{8}$$

On this same width-independent ball it is bi-Lipschitz with constants $1/2$ and $3/2$, uniformly in epsilon. For a scalar target $v^\top q + c$, the affine IHW head $\widehat{f}(\delta) = v^\top(a_0\delta - \delta_0) + c$ has error at most $\|v\|_2$ times the right side of equation 8. Two GELU units and an output bias implement this affine map exactly.

Why the construction works. P_i projects onto each row's tangent space; S_i describes the local effect of absolute value. To see why the geometry is uniform, put $s_i = \operatorname{sign} z_i/\sqrt{D}$ and $v_i = |z_i|/\sqrt{D}$. The patterns give $s_1^\top s_2 = 0$, $s_i^\top v_j = 0$ and $v_1^\top v_2 = 4/5$, unchanged by repetition. Hence $TT^\top = 2I_D - \sum_i (s_i s_i^\top + v_i v_i^\top)$ has eigenvalues $1, 1, 1/5, 9/5$ on their four-dimensional span, and two on its orthogonal complement. Thus $\lambda_{\min}(TT^\top) = 1/5$ at every repeated width and $TB = I_D$: the two signed tangent images jointly pass every latent direction. Each perturbation $B_i q$ is orthogonal to both $\mathbf{1}$ and z_i, so its row stays centered and its variance is $1 + \|B_i q\|^2/D$. On the stated ball, $\|B_i q\|_\infty \le \sqrt{5}/10 < \sqrt{2/5}$ keeps every coordinate's sign fixed. Thus the absolute value is locally linear, and normalization scales change only quadratically. The width-independent spectral lower bound controls the encoding norm, while repeating the four-coordinate patterns preserves the geometry. This yields equation 8, while derivative bounds establish the uniform local conditioning (Appendix E.3). Finally, $\mathrm{GELU}(t) - \mathrm{GELU}(-t) = t$ implements the scalar affine head exactly; prediction of the latent linear target still has the stated approximation error. Width uniformity concerns a fixed Euclidean ball, not a growing coordinatewise box.

These encodings use operations already permitted by the generator projection, residual/LN path and head; they do not guarantee that an ISAB/PMA encoder learns the required code.

5 Empirical Evaluation

We first compare training with the exact curves for the analyzed function class, then evaluate complete set-learning architectures.

5.1 Setup

The default MZAttn (MZ-Full) has width 64, four heads, FFN width 128, two ISAB encoder layers, $h = 8$ and $L = 4$. It uses AdamW at 10^{-4}, dropout 0.1 and early-stopping patience 20; task-specific settings are in Appendix H. Comparators include Deep Sets, Set Transformers (ST and ST-Large), ST + FiLM, a low-rank HyperNet, Perceiver (Jaegle et al., 2021), and linear-attention ST (LA-ST; Katharopoulos et al., 2020). Complete models pad variable-size sets without attention masks and step the cosine schedule per batch with an epoch-count period. Tables and captions specify seeds and standard-deviation conventions; we make no significance claims.

5.2 Training the analyzed gated class

We train $\mathcal{G}_L$ itself on the quadratic example of Figure 1. Table 1 reports means over three seeds, selecting among four initialization scales using validation error and evaluating on an independent

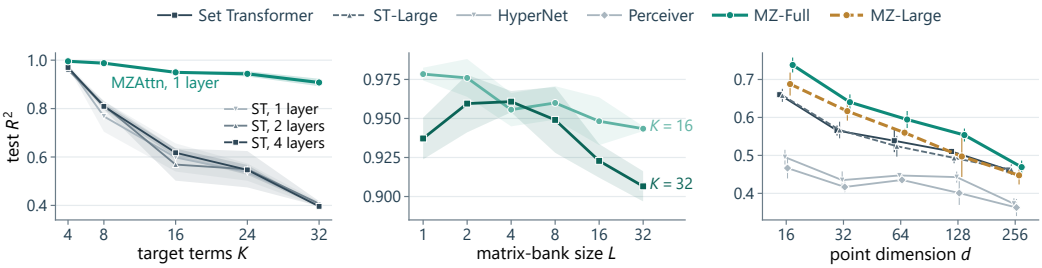

Figure 2: **Complete-model diagnostics. Left:** interaction targets with K terms, comparing one MZAttn layer ($L = K$) with ST at three depths. **Middle:** matrix-bank sweep on two target sizes. Both use three seeds. **Right:** MST scaling with training-split target normalization, five seeds; points are offset horizontally. Lines show means; bands and bars show sample standard deviations.

sample of 2^{19} points. Over sixteen further settings with random spectra in dimension six, no trained error falls below the infimum and the mean excess stays below 0.6% of the target variance (Appendix C.4). The errors approach the exact curves, with near-zero error emerging at different gate budgets under the two input laws. This checks the same function class as Theorems 4.5–4.6 in these settings; it does not establish general optimization convergence or validate the complete architecture's readout constructions.

Table 1: Mean squared error of the native gated class $\mathcal{G}_L$ for f_Q with $Q = \mathrm{diag}(4, 2, 1, -3)$: exact infimum (Theorems 4.5 and 4.6) and trained error, mean over three seeds.

Input law		$L = 0$	$L = 1$	$L = 2$	$L = 3$
$\mathcal{N}(0, I_4)$	infimum	60.0	10.0	2.00	0.000
	trained	60.3	10.3	2.11	0.056
uniform on $2\mathbb{S}^3$	infimum	34.7	0.667	0.000	0.000
	trained	34.7	0.706	0.011	0.004

5.3 Interaction learning and matrix-bank budgets

For complete models, the diagnostic target is $f(S) = w^\top [M_0 + \sum_{l=1}^{K} \mu_l(S) M_l] \mu(S)$, with fixed random w, M_l in dimension 32 and set mean $\mu(S)$. Here K counts construction terms, not the scalar complexity of Section 4.2. Figure 2 (left) shows that additional ST depth does not close the gap to one MZAttn layer. The bank sweep (middle) favors compact banks over matching L to K: a single direction already performs well, while the largest bank has the lowest mean on both targets. These models include FFNs, LayerNorm and the IHW head, so the isolated-core theorems do not by themselves explain the gap. Appendix I.2 retains the detailed comparisons, compact configurations and readout variants.

5.4 Complete-model performance, scope and cost

The seven tasks in Table 2 are (A) summed coordinatewise set maxima, (B) Chamfer distance, (C) ModelNet40 classification, (D) support-conditioned class-fraction prediction, (E) ScanObjectNN classification, (F) post-alignment Chamfer distance under sample-dependent rotation and (G) clustered-set MST weight. Appendix H.2 gives their data and protocols.

MZAttn has the lowest mean MSE on regression, rotation matching and MST, 13.0%, 43.0% and 41.1% below ST-Large. These comparisons use similar budgets (340–372K parameters, depending on the task). FiLM is better on Chamfer and few-shot retrieval with fewer parameters, while ST-Large leads both classification tasks. The full seven-task comparison therefore supports task-dependent benefits rather than a uniform advantage. Table S1 provides the expanded metrics and representative parameter counts; in Table S2, a compact model with $h = 2$ and $L = 1$ stays within

Table 2: **Complete-model set-learning comparisons.** Mean ± population standard deviation over 3–5 seeds; bold marks the best mean. Budgets vary by task (Table S1). Dashes: not reported.

	Regression and matching: MSE ↓			
Model	A: Regression	B: Chamfer	F: Rotation	G: MST
Set Transformer	3.43 ± 0.09	138.0 ± 10.3	487.5 ± 31.1	209.8 ± 1.8
ST + FiLM	3.77 ± 0.26	$\mathbf{48.0} \pm 7.9$	490.5 ± 17.8	232.7 ± 12.5
ST-Large	3.36 ± 0.10	106.5 ± 5.3	423.2 ± 51.8	168.7 ± 23.0
Perceiver	3.47 ± 0.05	136.2 ± 17.5	872.1 ± 163.8	249.0 ± 4.7
HyperNet	3.51 ± 0.01	76.1 ± 5.7	1373.0 ± 141.6	291.2 ± 9.2
MZAttn	$\mathbf{2.92} \pm 0.09$	58.4 ± 11.9	$\mathbf{241.2} \pm 3.0$	$\mathbf{99.4} \pm 1.9$

	Classification: accuracy ↑		Retrieval: MSE ↓
Model	C: ModelNet40	E: ScanObjectNN	D: Few-shot (10^{-2})
Deep Sets	–	$40.6 \pm 1.0\%$	–
Set Transformer	$77.5 \pm 0.6\%$	$71.6 \pm 2.2\%$	4.81 ± 1.85
ST + FiLM	$78.4 \pm 0.8\%$	$68.9 \pm 3.2\%$	$\mathbf{3.06} \pm 0.65$
ST-Large	$\mathbf{79.9} \pm 1.0\%$	$\mathbf{72.9} \pm 2.4\%$	4.28 ± 0.92
Perceiver	$75.0 \pm 0.3\%$	$67.3 \pm 1.2\%$	12.19 ± 5.81
HyperNet	$75.5 \pm 1.7\%$	$66.1 \pm 0.9\%$	11.96 ± 6.96
MZAttn	$78.1 \pm 0.7\%$	$69.4 \pm 3.7\%$	4.71 ± 2.77

0.022 of larger MZ configurations on three further tasks, although this comparison changes both h and L and does not isolate retention.

Scope and cost. The gains on normalized MST shrink with dimension, and target normalization reduces the gaps relative to the raw-target comparison (Figure 2; Table S7). Results on hull volume, diameter and neighbor radius are mixed, with some rankings reversing at higher dimension (Figure 5). EGNN leads the tested invariant targets; its invariant output is mismatched to fixed-frame targets, so those comparisons do not establish an advantage over general equivariant models (Tables S12–S13). The synthetic rotation-matching benefit does not transfer to the tested real ModelNet40 shapes, where MZ-Full has $R^2 = -0.066$ versus 0.690 for Perceiver; other real-feature and transfer tasks show no consistent advantage (Appendix I.5). These complete-model comparisons do not isolate retention or the effect of the shared training protocol. Against a same-width ST, MZAttn uses 1.7–$1.9\times$ the parameters, 2.4–$2.7\times$ the inference time and 2.2–$5.5\times$ the training-step time (Table S14).

6 Discussion and Limitations

MZAttn couples context-dependent transformations with a persistent generator state. Its analysis separates the resources required to generate transformation features, retain them through subsequent computation, and access them at prediction. The central lesson is that preserving a current readout need not preserve later computation, while recoverability alone need not provide stable access. Together with the domain-dependent approximation results, this motivates designing intermediate states around both the prediction task and the computations that follow.

The theoretical characterizations concern specified cores and local readout constructions under explicit input and normalization assumptions. They characterize representation and approximation, rather than guarantee that end-to-end training realizes the constructions. The experiments assess the combined architecture under the reported protocol; further controlled comparisons are needed to separate the effects of retention, gating, capacity, and readout. The richer state also introduces computational overhead, and its empirical value depends on the task. Connecting the constructions to learned representations and reducing the cost of retaining useful features are therefore priorities for further work.

AI USE STATEMENT

Generative AI tools assisted research ideation, conjecture formulation and proof development, experimental implementation and debugging, and execution and analysis of the gated-class experiment in Section 5.2. They also assisted manuscript drafting and editing, figure preparation, literature retrieval and numerical checks. The authors reviewed the AI-assisted material, checked proofs and references, tested code and verified numerical results, and take responsibility for the final content.

ETHICS STATEMENT

This work studies representation and approximation properties of context-dependent set attention. Experiments use synthetic tasks and existing public benchmarks for point clouds, molecular properties, images and tabular prediction.

REPRODUCIBILITY STATEMENT

Appendices A–F state the assumptions and provide constructions and proofs. Appendix C.4 details the direct training checks of the analyzed gated class. Appendices H and I describe data processing, model and training configurations, evaluation protocols, computational costs and variability across runs. The code will be made publicly available after publication.

REFERENCES

Amr Alanwar, Anne Koch, Frank Allgöwer, and Karl Henrik Johansson. Data-driven reachability analysis using matrix zonotopes. In *Conference on Learning for Dynamics and Control*, 2021.

Amr Alanwar, Anne Koch, Frank Allgöwer, and Karl Henrik Johansson. Data-driven reachability analysis from noisy data. *IEEE Transactions on Automatic Control*, 68:3054–3069, 2023.

Matthias Althoff. *Reachability Analysis and its Application to the Safety Assessment of Autonomous Cars*. PhD thesis, Technische Universität München, 2010.

C. Bradford Barber, David P. Dobkin, and Hannu Huhdanpaa. The Quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software*, 22:469–483, 1996.

Peter L. Bartlett, Dylan J. Foster, and Matus J. Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, 2017.

Nora Belrose and Alice Rigg. Converting MLPs into polynomials in closed form. *arXiv preprint arXiv:2502.01032*, 2025.

Gérard Ben Arous, Murat A. Erdogdu, Nuri Mert Vural, and Denny Wu. Learning quadratic neural networks in high dimensions: SGD dynamics and scaling laws. In *Advances in Neural Information Processing Systems*, 2025.

Gregory Bonaert, Dimitar I. Dimitrov, Maximilian Baader, and Martin Vechev. Fast and precise certification of transformers. In *ACM SIGPLAN International Conference on Programming Language Design and Implementation*, 2021.

Shaked Brody, Uri Alon, and Eran Yahav. On the expressivity role of LayerNorm in transformers' attention. In *Findings of the Association for Computational Linguistics: ACL 2023*, 2023.

Joan Bruna, Arthur Szlam, and Yann LeCun. Signal recovery from pooling representations. In *International Conference on Machine Learning*, 2014.

Long Kiu Chung and Shreyas Kousik. Provably-safe neural network training using hybrid zonotope reachability analysis. In *IEEE Conference on Decision and Control*, 2025.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1:211–218, 1936.

Bing Gao, Qiyu Sun, Yang Wang, and Zhiqiang Xu. Phase retrieval from the magnitudes of affine linear measurements. *Advances in Applied Mathematics*, 93:121–141, 2018.

Johannes Gasteiger, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. In *International Conference on Learning Representations*, 2020.

David Ha, Andrew M. Dai, and Quoc V. Le. HyperNetworks. In *International Conference on Learning Representations*, 2017.

Daniel Haider, Martin Ehler, and Peter Balazs. Convex geometry of ReLU-layers, injectivity on the ball and local reconstruction. In *International Conference on Machine Learning*, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Andrew Jaegle, Felix Gimeno, Andrew Brock, Oriol Vinyals, Andrew Zisserman, and João Carreira. Perceiver: General perception with iterative attention. In *International Conference on Machine Learning*, 2021.

Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc Van Gool. Dynamic filter networks. In *Advances in Neural Information Processing Systems*, 2016.

Matt Jordan, Jonathan Hayase, Alexandros G. Dimakis, and Sewoong Oh. Zonotope domains for Lagrangian neural network verification. In *Advances in Neural Information Processing Systems*, 2022.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, 2020.

Hyunjik Kim, Andriy Mnih, Jonathan Schwarz, Marta Garnelo, Ali Eslami, Dan Rosenbaum, Oriol Vinyals, and Yee Whye Teh. Attentive neural processes. In *International Conference on Learning Representations*, 2019.

Niklas Kochdumper and Matthias Althoff. Constrained polynomial zonotopes. *Acta Informatica*, 60:279–316, 2023.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

John Lamperti. On the isometries of certain function-spaces. *Pacific Journal of Mathematics*, 8: 459–466, 1958.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86:2278–2324, 1998.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set Transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning*, 2019.

Moshe Leshno, Vladimir Ya. Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6:861–867, 1993.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with Slot Attention. In *Advances in Neural Information Processing Systems*, 2020.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Brian Mak and Jeffrey Flanigan. Residual matrix transformers: Scaling the size of the residual stream. In *International Conference on Machine Learning*, 2025.

Matthew Mirman, Timon Gehr, and Martin Vechev. Differentiable abstract interpretation for provably robust neural networks. In *International Conference on Machine Learning*, 2018.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT Press, second edition, 2018.

Alexey Ovchinnikov, Isabel Pérez Verona, Gleb Pogudin, and Mirco Tribastone. CLUE: Exact maximal reduction of kinetic models by constrained lumping of differential equations. *Bioinformatics*, 37:1732–1738, 2021.

Michael T. Pearce, Thomas Dooms, Alice Rigg, Jose Oramas, and Lee Sharkey. Bilinear MLPs enable weight-based mechanistic interpretability. In *International Conference on Learning Representations*, 2025.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *AAAI Conference on Artificial Intelligence*, 2018.

Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017a.

Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems*, 2017b.

Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1:140022, 2014.

RDKit Contributors. RDKit: Open-source cheminformatics software. `https://www.rdkit.org`, 2024.

Víctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In *International Conference on Machine Learning*, 2021.

Simon Schug, Seijin Kobayashi, Yassir Akram, João Sacramento, and Razvan Pascanu. Attention as a hypernetwork. In *International Conference on Learning Representations*, 2025.

Kristof T. Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. In *Advances in Neural Information Processing Systems*, 2017.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Duc Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *IEEE/CVF International Conference on Computer Vision*, 2019.

Santosh Vempala and John Wilmes. Gradient descent for one-hidden-layer neural networks: Polynomial convergence and SQ lower bounds. In *Conference on Learning Theory*, 2019.

Edward Wagstaff, Fabian Fuchs, Martin Engelcke, Ingmar Posner, and Michael A. Osborne. On the limitations of representing functions on sets. In *International Conference on Machine Learning*, 2019.

Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3D ShapeNets: A deep representation for volumetric shapes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Brandon Yang, Gabriel Bender, Quoc V. Le, and Jiquan Ngiam. CondConv: Conditionally parameterized convolutions for efficient inference. In *Advances in Neural Information Processing Systems*, 2019.

Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94:103–114, 2017.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R. Salakhutdinov, and Alexander J. Smola. Deep Sets. In *Advances in Neural Information Processing Systems*, 2017.

Defa Zhu, Hongzhi Huang, Zihao Huang, Yutao Zeng, Yunyao Mao, Banggu Wu, Qiyang Min, and Xun Zhou. Hyper-Connections. In *International Conference on Learning Representations*, 2025.

SUPPLEMENTARY MATERIAL: OVERVIEW

The theory appendices follow the main text. Appendix A supports the architecture of Section 3, and Appendices B, C, D and E prove the results of Sections 4.1, 4.2, 4.3 and 4.4. A result stated in the main text keeps its number there. Appendix F collects results the main text does not use. Appendix G extends the related work, Appendix H gives the experimental settings and compute, and Appendix I reports additional experiments.

A ARCHITECTURE AND STRUCTURAL PROPERTIES

This appendix supports Section 3: the zonotope algebra behind the fixed-size representation, permutation structure, the gradient structure of the two streams, and a generalization bound for the full architecture.

Products admit an outer zonotope enclosure with up to $h+L+Lh$ generators, motivating a fixed-size representation.

Slot g receives the conditional action $B_g(\mu)\hat{c}_i$, where $B_g(\mu) = \sum_l W_{\mathrm{mix},gl}\gamma_l(\mu)M_l$.

The distinguishing choice is to keep processing these actions as a separate feature stream rather than reducing them to one vector immediately.

A.1 ZONOTOPE AND MATRIX-ZONOTOPE ALGEBRA

Write $\langle c, G\rangle = \{c + G\alpha : \|\alpha\|_\infty \le 1\}$ for the zonotope with center c and generator columns $g_1, \dots, g_h$; its interval hull $[c-\delta, c+\delta]$, with $\delta = \sum_g |g_g|$, is the tightest axis-aligned box containing it. Multiplying it by the matrix zonotope $\mathcal{M}$ of equation 1 and writing all coefficient dependencies gives

$$\left(M_c + \sum_l \xi_l M_l\right)\left(c + \sum_g \alpha_g g_g\right) = M_c c + \sum_g \alpha_g M_c g_g + \sum_l \xi_l M_l c + \sum_{l,g} \xi_l \alpha_g M_l g_g.$$

The products $\xi_l\alpha_g$ are generally dependent on the original coefficients. Treating each as a new independent element of $[-1,1]$ therefore gives the outer approximation $\mathcal{M}\langle c, G\rangle \subseteq \langle M_c c, [\{M_c g_g\}, \{M_l c\}, \{M_l g_g\}]\rangle$ with $h + L + Lh$ generators. The exact product need not be convex: $M = \left(\begin{smallmatrix} \xi & 0 \\ 1 & \xi \end{smallmatrix}\right)$ and $z = (1, \alpha)^\top$ produce $(\xi, 1 + \xi\alpha)$; $(1, 2)$ and $(-1, 2)$ are attainable but their midpoint $(0, 2)$ is not. For constrained and polynomial variants, see Kochdumper & Althoff (2023) and Althoff (2010).

Proposition A.1 (Generator mixing and a conservative enclosure)**.** *Use column-generator matrices $G \in \mathbb{R}^{d\times h}$, $H \in \mathbb{R}^{d\times L}$ and mixing weights $W \in \mathbb{R}^{h\times L}$. Let $Z = c + G[-1,1]^h + H[-1,1]^L$ and $\tilde{Z} = c + (G + HW^\top)[-1,1]^h$.*

(a) If each column of W has ℓ_1 norm at most one, then $\tilde{Z} \subseteq Z$.

(b) Column norms at least one do not imply the reverse inclusion; norms equal to one do not imply equality. For $G = e_1$, $H = e_2$, $W = [1]$, Z is a square and $\tilde{Z}$ its diagonal segment.

(c) For arbitrary W, $Z \subseteq \tilde{Z} \oplus \mathrm{Box}(r)$, where $r = |H|\mathbf{1} + |HW^\top|\mathbf{1}$ and $\mathrm{Box}(r) = \{z : |z_j| \le r_j\}$.

Proof. For (a), $(G + HW^\top)\zeta = G\zeta + H(W^\top\zeta)$ with $|(W^\top\zeta)_l| \le \sum_k |W_{kl}| \le 1$ for $\zeta \in [-1, 1]^h$, so both coefficient vectors are admissible in the original Minkowski sum. For (b), the point $(1, -1)$ of the square is not on the diagonal $\{(t, t) : |t| \le 1\}$: mixing forces formerly independent directions to share a coefficient. For (c), a source point $x = c + G\zeta + H\eta$ and the mixed point $\tilde{x} = c + (G + HW^\top)\zeta$ satisfy $|x - \tilde{x}| = |H\eta - HW^\top\zeta| \le r$ componentwise. This is a sufficient enclosure, not a minimum-volume claim. □

The box compensation is an available enclosure, not part of the reported trained model. Unconstrained learned mixing is therefore a feature-compression operation without a general set-containment guarantee.

A.2 Permutation structure and initialization

Proposition A.2 (Permutation structure and initialization)**.** *With deterministic shared tokenwise operations and permutation-invariant context gates, MZ-SAB and MZ-ISAB are permutation-equivariant and the full pooled architecture is permutation-invariant. For the core in equation 3, setting $M_l = 0$, $M_c = I$ eliminates the newly injected generator term and gives the usual center aggregation. The reported small, nonzero random initialization is a perturbation of this limit, not exact equality of the full model to standard attention.*

Proof. **Attention.** Let P permute query tokens and R permute key/value tokens. Shared projections transform the score matrix by $QK^\top \mapsto PQK^\top R^\top$. Rowwise softmax satisfies

$$\mathrm{softmax}(PQK^\top R^\top) = P\,\mathrm{softmax}(QK^\top)R^\top.$$

Multiplication by the permuted values cancels $R^\top R$. The context mean and hence its gates are invariant to R. Center and generator outputs consequently transform by the query permutation P.

Blocks and pooling. Shared tokenwise FFNs, normalization, mixing and residual additions commute with token permutations. SAB is equivariant by taking both token sets equal. In ISAB, attention from fixed inducing queries is invariant to input order, and attention back to the input queries is equivariant. Fixed PMA queries give an invariant fixed-order representation; the final output head therefore preserves invariance. These claims concern deterministic inference; independently resampled dropout masks give the corresponding property in distribution.

Initialization and inclusion. Setting all $M_l = 0$ eliminates the injected generator term in equation 3; $M_c = I$ gives the usual center aggregation. Small nonzero random initialization is a perturbation of this limit. Exact equivalence of full networks also requires matching their center operations and ignoring any remaining generator features at the readout. If the center path, with generator contributions disabled, has exactly the operations of a chosen standard Set Transformer, copying its parameters and setting the readout weights on generator features to zero leaves each of its functions unchanged. The extended architecture then contains that center architecture and inherits any approximation property proved for it on the same domain and under the same assumptions, such as that of Lee et al. (2019). This conditional inclusion is not a strict separation between universal model families, and the core initialization alone does not establish equality of the full architectures. □

A.3 Center independence and generator-only gradients

Proposition A.3 (Center independence and generator-only gradients)**.** *In the deterministic default architecture, including all ISAB, PMA and SAB blocks, centers are independent of generator-only parameters. If the prediction head uses centers alone, its prediction and data-loss gradient are independent of every generator-only parameter. Weight decay or a separate regularizer may still update those parameters.*

Proof. The center projection depends only on the input and center parameters. In each attention block, queries, keys, scores, context means, center values and center output are functions of centers and parameters that enter the center path. The generator update receives these quantities but

does not feed back into them. Tokenwise center LayerNorms, FFNs and residuals also contain no generators. Induction through attention to inducing points, attention back to input tokens, pooling and decoder blocks proves center independence. A center-only head therefore defines a function constant in generator-only parameters, and differentiating gives zero. Shared parameters, notably M_c, are excluded from this parameter group; an explicit regularization term is excluded from the data loss. □

The proposition concerns the full implemented data flow. Removing IHW access therefore also removes the generator branch from learning, so that intervention alone cannot identify a benefit of geometric interpretation at equal effective capacity; a suitable comparison retains a trainable path and changes the summary or mixing.

A.4 Generalization of the full architecture

Theorem A.4 (Generalization for the bounded full architecture)**.** *Let $\{f_\theta : \theta \in [-B, B]^P\}$ be a fixed finite MZ architecture with all learned parameters included in θ, and let its loss obey $0 \le \ell_\theta(z) \le M$ and $|\ell_\theta(z) - \ell_{\theta'}(z)| \le K_\ell \|\theta - \theta'\|_\infty$ uniformly over the declared data domain. For n i.i.d. observations, write $R(\theta) = \mathbb{E}\ell_\theta(Z)$ and $R_n(\theta) = n^{-1}\sum_{i=1}^n \ell_\theta(Z_i)$. For any fixed $\eta > 0$ and $0 < \delta < 1$, with probability at least $1 - \delta$, simultaneously for every θ,*

$$R(\theta) - R_n(\theta) \le 2\eta + M\sqrt{\frac{P\log(1 + 2BK_\ell/\eta) + \log(1/\delta)}{2n}}. \tag{A.1}$$

Bounded inputs and targets, bounded set cardinality, positive LayerNorm epsilon and deterministic inference give finite constants of this kind for bounded Lipschitz losses. For squared loss, a uniform output bound F, target bound Y, and parameter Lipschitz constant K_f give $M = (F + Y)^2$ and $K_\ell = 2(F + Y)K_f$.

Proof. **Step 1: a finite parameter net.** The case $M = 0$ is immediate. Otherwise assume $K_\ell > 0$ and set $r = \eta/K_\ell$. Partition each coordinate interval $[-B, B]$ into $\lceil B/r \rceil$ intervals of length at most $2r$ when $r < B$, and into one interval otherwise. Their midpoints give an ℓ_∞ r-net of the cube, with cardinality

$$N_r \le (1 + 2B/r)^P = (1 + 2BK_\ell/\eta)^P.$$

The slightly looser constant also covers all endpoint cases.

Step 2: concentration on the net. For each fixed net point, Hoeffding's inequality for losses in $[0, M]$ gives $\Pr\{R - R_n > t\} \le \exp(-2nt^2/M^2)$. A union bound over the net, with $t = M\sqrt{[\log N_r + \log(1/\delta)]/(2n)}$, makes the inequality hold at every net point with probability at least $1 - \delta$.

Step 3: extend to every parameter. For any θ choose a net point $\bar\theta$ with $\|\theta - \bar\theta\|_\infty \le r$. Uniform parameter Lipschitzness implies both $|R(\theta) - R(\bar\theta)| \le \eta$ and $|R_n(\theta) - R_n(\bar\theta)| \le \eta$. Hence

$$R(\theta) - R_n(\theta) \le R(\bar\theta) - R_n(\bar\theta) + 2\eta,$$

which proves equation A.1. If $K_\ell = 0$, all losses agree across parameters, and the one-function concentration bound suffices. Since the event is uniform, the result applies to a data-selected parameter as well.

Step 4: account for the complete MZ computation. With a fixed maximum cardinality there are finitely many input shapes. For each shape, bounded inputs and a compact parameter cube yield bounded intermediate activations in the finite computation graph. Affine operations, products, softmax, tanh and GELU have bounded derivatives on the resulting compact sets; ReLU and absolute value are Lipschitz. LayerNorm with fixed positive epsilon has a denominator bounded away from zero, so it is Lipschitz on those sets as well. Composing these operations gives a finite uniform parameter constant K_f, taking the maximum over the finitely many input shapes. This argument includes gate weights and biases, Q/K routing, value projections, mixing, residuals, both normalization paths, FFNs and the final readout. For squared loss, the identity

$$\big|\|f - y\|_2^2 - \|f' - y\|_2^2\big| \le (\|f - y\|_2 + \|f' - y\|_2)\|f - f'\|_2$$

gives the stated M and K_ℓ. □

The argument is a classical covering and concentration argument, of the type developed in Mohri et al. (2018), rather than a new statistical rate specific to MZ; its constants can be large. The module estimate of Proposition F.2 controls one part of the computation only. Replacing all network factors by $\|M_c\|_2 + \sum_l \|M_l\|_2$ would omit the changing gates and routing and would not justify a spectral-margin bound such as that of Bartlett et al. (2017).

B One Gated Core versus Attention

This appendix proves the results of Section 4.1 in the order stated there, after a definition and the native-core representation on which they rely.

B.1 Transformation degrees of freedom

Definition B.1 (Transformation degrees of freedom of an operator)**.** *For a specified matrix-valued map T on a declared context domain $\mathcal{C}$, define*

$$\mathrm{TDOF}(T) = \dim \operatorname{span}\{T(c) - T(c') : c, c' \in \mathcal{C}\},$$

the affine dimension of the operator image.

TDOF does not depend on a choice of matrix basis, but it does depend on the choice of T, the context and the domain. It is not an intrinsic complexity of every scalar prediction formed from T: a single probe need not identify the operator (Lemma C.1), so prediction alone does not determine its TDOF.

B.2 Native-core representation

Theorem B.2 (Native-core representation)**.** *(a) On constant-token sets, standard attention with affine projections and affine output, no FFN and no normalization, is affine in the token for any head count. (b) With uniform aggregation and zero incoming generators, an isolated signed MZ core with one stored slot realizes $[M_0 + \sum_{l=1}^{L} \tanh(w_l^\top \mu + b_l) M_l]\mu$ exactly. (c) Its square bank and gates contain $(L+1)d_k^2 + L(d_k+1)$ parameters, or L fewer without gate biases, excluding routing, other learned projections, mixing, normalization, FFNs and the full prediction head. Its operator image has affine dimension at most $\min\{L, d_k^2\}$, so a fixed bank representing an operator of TDOF r (Definition B.1) exactly needs at least r directions.*

Proof. **Part (a): fixed value maps on constant tokens.** For an H-head standard attention block without FFN or normalization, write its bias-free output as

$$o(S) = \sum_{h=1}^{H} W_o^{(h)} W_V^{(h)} \left(\sum_{j=1}^{n} \alpha_j^{(h)}(S) x_j \right).$$

Each value transformation is fixed, but the weighted averages can depend nonlinearly on S. We therefore restrict the argument to $S_x = \{x, \dots, x\}$. All keys in a head are identical, so their softmax scores are equal and $\alpha_j^{(h)} = 1/n$. This holds for either a shared token query or a fixed pooling query. Consequently

$$o(S_x) = Wx, \qquad W = \sum_h W_o^{(h)} W_V^{(h)}.$$

With affine value and output projections the same calculation gives $o(S_x) = Wx + b$. Affine query/key biases do not change equality of keys. An affine residual also preserves this conclusion. Thus the restriction is affine for every head count; the argument does not claim that attention is linear on nonconstant sets.

Part (b): exact native construction. Use $W_Q = 0$ and identity value projection, giving uniform aggregation $\hat{c}_i = \mu(S)$. Set incoming generators to zero. For each target term choose its matrix M_l and the native gate $\gamma_l(\mu) = \tanh(w_l^\top \mu + b_l)$, and let $M_c = M_0$. Use one generator slot with the

fixed mixing row $W_{\mathrm{mix}} = (1, \ldots, 1)$. Equation 3 then gives

$$o_i^c = M_0\mu, \qquad o_i^G[1,:] = \left(\sum_{l=1}^{L} \gamma_l(\mu) M_l \mu\right)^\top .$$

All output tokens are identical. Arithmetic-mean pooling preserves these features, and the signed readout $(c, g) \mapsto c + g$ returns the target exactly. No additional adaptive PMA layer, normalization or FFN is part of this isolated-core construction.

Part (c): parameter count and operator span. The $L + 1$ square matrices supply $(L + 1)d_k^2$ parameters and the L affine gates supply $L(d_k + 1)$, or L fewer without gate biases. The chosen fixed projections, mixing and readout require no learned parameters in this realization. For two contexts, $M(c) - M(c') = \sum_l [\gamma_l(c) - \gamma_l(c')] M_l$ lies in a subspace of dimension at most $\min\{L, d_k^2\}$. An operator image of affine dimension r therefore requires at least r directions in any such fixed bank. □

B.3 Proof of Theorem 4.1 (separation for every nonzero matrix)

Proof. **Part (a): native realization.** Theorem B.2(b) with $L = 1$, $M_0 = I$, $M_1 = A$, $w_1 = e_1$ and $b_1 = 0$ gives f_A exactly. This uses a native gate, not an approximation to multiplication by a finite-width FFN.

Part (b): exclusion of the standard block. The comparator has affine query/key/value projections, softmax aggregation and an affine output map, without nonlinear input embeddings, FFNs or normalization. By Theorem B.2(a), the standard block is $Wx + b$ on S_x. Its restriction to every affine line in x is affine. We show that f_A has a nonaffine line restriction for each $A \neq 0$.

If $Ae_1 \neq 0$, take $x(t) = te_1$. The target becomes

$$f_A(S_{x(t)}) = te_1 + t\tanh(t)Ae_1.$$

The second derivative at $t = 0$ is $2Ae_1 \neq 0$, so this restriction cannot be affine.

If $Ae_1 = 0$, choose $j \geq 2$ with $Ae_j \neq 0$, which exists because $A \neq 0$. Crucially, vary the gate coordinate and keep the nonzero value direction fixed: set $x(t) = te_1 + e_j$. Then

$$Ax(t) = Ae_j, \qquad f_A(S_{x(t)}) = te_1 + e_j + \tanh(t)Ae_j.$$

Its third derivative at zero is $-2Ae_j \neq 0$, again excluding an affine restriction. The alternative line $te_j + \varepsilon e_1$ would keep the gate fixed and cannot supply this contradiction. This second case completes the result for every nonzero A, without requiring its first column to be nonzero. For $d = 1$ only the first case is possible. Projection biases on the comparator merely change its affine offset and do not affect either contradiction. □

B.4 Proof of Theorem 4.2 (ReLU width and depth cost)

Proof. **Step 1: a strictly curved scalar restriction.** For $u = Ae_1/\|Ae_1\|_2$, $g(t) = t\,u^\top e_1 + \|Ae_1\|_2 t\tanh t$. Direct differentiation gives

$$g''(t) = 2\|Ae_1\|_2 \operatorname{sech}^2 t(1 - t\tanh t) \geq \kappa > 0 \quad (|t| \leq a).$$

The inequality uses $t\tanh t \leq a\tanh a$ and $\operatorname{sech}^2 t \geq \operatorname{sech}^2 a$.

Step 2: an affine piece has a minimum approximation error. On an interval $[r, s]$ of length ℓ, convexity of $g(t) - \kappa t^2/2$ implies

$$g(r) + g(s) - 2g((r + s)/2) \geq \kappa\ell^2/4.$$

An affine function has zero midpoint difference. If it approximates g with uniform error ϵ on this interval, the difference of the three errors has absolute value at most 4ϵ. Hence $\epsilon \geq \kappa\ell^2/16$. For a continuous piecewise-affine approximant with at most K pieces on $[-a, a]$, one interval has length at least $2a/K$, and therefore $\epsilon \geq \kappa a^2/(4K^2)$.

Step 3: count the pieces of the declared comparator. Identical input tokens remain identical under every shared operation. Every self-attention row therefore averages identical affine values; its result

is affine in the current token even when query/key scores depend on that token. On each existing linear interval, the w_t preactivations of the next ReLU FFN are affine in t. Each can add at most one breakpoint, giving at most $w_t + 1$ subintervals. Residual addition does not increase this partition. Induction yields $K \le \prod_t (w_t + 1)$. Arithmetic-mean pooling and affine readout add no breakpoints. Substitution into Step 2 proves the bound.

Step 4: a matching univariate order. Let $M = \max_{[-a,a]} |g''(t)| < \infty$. Linear interpolation on K equal intervals has error at most $M(2a/K)^2/8$. Every continuous linear spline has a representation as an affine function plus a sum of $K-1$ terms $b_j \,\mathrm{ReLU}(t - t_j)$. Thus $O(\epsilon^{-1/2})$ units suffice for this fixed restriction, with an affine residual or at most two additional units for the affine part. This proves the claimed order without adding neuron costs across unrelated target coordinates. Nonidentical inducing seeds, LayerNorm or GELU would invalidate Step 3 and are explicitly excluded. More general ReLU approximation results are studied by Yarotsky (2017). □

B.5 Proof of Theorem 4.3 (exact depth–direction tradeoff)

Proof. Here $1 \le k \le d-1$, and $\mathcal{N} = \{u e_d^\top : u \perp e_d\}$. The residual operator stack is $T_D(c) = \prod_{t=1}^{D} [I + A_t(c)]$, with $A_t(c) = \sum_{j=1}^{q_t} a_{tj}(c) B_{tj}$, fixed $B_{tj} \in \mathcal{N}$, and square-integrable coefficients. No intervening maps outside this algebra are permitted, and $Q = \sum_t q_t$.

Step 1: closure of the entire comparator algebra. For any $u e_d^\top, v e_d^\top \in \mathcal{N}$, $(u e_d^\top)(v e_d^\top) = u(e_d^\top v) e_d^\top = 0$. This holds for every allowed layer matrix, not only the target bank. Hence expansion of the residual product gives exactly

$$T_D(c) = I + \sum_{t=1}^{D} A_t(c) = I + \sum_{t,j} a_{tj}(c) b_{tj} e_d^\top, \quad B_{tj} = b_{tj} e_d^\top.$$

All vectors b_{tj} lie in $e_d^\perp$ and span a subspace U of dimension at most Q.

Step 2: expose the operator difference in prediction error. Set $h(c) = \sum_{l=1}^{k} \tanh(c_l) e_l$ and $\hat{h}(c) = \sum_{t,j} a_{tj}(c) b_{tj} \in U$. Since $e_d^\top \mu = z$ and z is independent of c with unit variance,

$$\mathbb{E}\|[T_*(c) - T_D(c)]\mu\|_2^2 = \mathbb{E}\{z^2 \|h(c) - \hat{h}(c)\|_2^2\} = \mathbb{E}\|h(c) - \hat{h}(c)\|_2^2.$$

For fixed U the best choice is the orthogonal projection $P_U h(c)$. The independent odd tanh coordinates have mean zero and common variance $\nu > 0$, so $\mathbb{E}[hh^\top] = \nu P_E$ for $E = \mathrm{span}\{e_1, \ldots, e_k\}$. Therefore

$$\mathbb{E}\|(I - P_U)h\|_2^2 = \nu\{k - \mathrm{tr}(P_U P_E)\} \ge \nu\{k - \min(k, Q)\}.$$

Step 3: attain the lower bound. Let $r = \min(k, Q)$. Assign r distinct directions $e_l e_d^\top$ to the available layer slots, with coefficients $\tanh(c_l)$; fill any remaining slots with zero matrices. Then $T_D = I + \sum_{l=1}^{r} \tanh(c_l) e_l e_d^\top$, and the omitted coordinates have squared error exactly $(k - r)\nu$. This proves equality for every integer budget Q.

Step 4: one native block and the depth consequence. The target coefficient vector ranges over $(-1, 1)^k$, and the $e_l e_d^\top$ are linearly independent, so $\mathrm{TDOF}(T_*) = k$. Use $M_c = I$, $M_l = e_l e_d^\top$, native gates $\tanh(\mu_l)$, uniform aggregation and one signed slot, as in Theorem B.2(b). This realizes $T_* \mu$ in one core. Since zero error in the comparator requires $Q \ge k$, the per-layer bound $q_t \le q$ implies $D \ge \lceil k/q \rceil$. No FFN-width identity is used; q is a matrix-direction budget, not d_{ff}/d_k. An intervening matrix $A = e_d e_j^\top$ gives $M_i A M_j = M_i \ne 0$, showing precisely why arbitrary learned maps fall outside this proof. □

Corollary B.3 (Stability under nonzero cross-layer products)**.** *Use the target, Gaussian input and direction budget Q of Theorem 4.3. Let $A_t(c)$ be an allowed square-zero layer with $\sup_c \|A_t(c)\|_2 \le a_t$, and let $E_t(c)$ be arbitrary matrix-valued perturbations with $\sup_c \|E_t(c)\|_2 \le \varepsilon_t$. Define the perturbed residual stack, in the displayed order, by*

$$\widetilde{T}_D(c) = [I + A_D(c) + E_D(c)] \cdots [I + A_1(c) + E_1(c)],$$

and set

$$\rho_D = \sum_{t=1}^{D} \varepsilon_t \left(1 + \sum_{j<t} a_j\right) \prod_{j>t} (1 + a_j + \varepsilon_j).$$

Allow also a perturbed target $\widetilde{T}_*(c) = T_*(c) + F(c)$*, with* $\sup_c \|F(c)\|_2 \le \varepsilon_*$*. Then*

$$\mathbb{E}\|[\widetilde{T}_*(c) - \widetilde{T}_D(c)]\mu\|_2^2 \ge \left[\sqrt{\nu(k - \min\{k, Q\})} - \sqrt{d}\,(\rho_D + \varepsilon_*)\right]_+^2, \tag{B.1}$$

where $[x]_+ = \max(x, 0)$*. No zero-product condition is imposed on the perturbations or their cross products with the original layers.*

Proof. Write $B_t = I + A_t$ and $\widetilde{B}_t = B_t + E_t$. The unperturbed stack $T_D = B_D \cdots B_1$ belongs to Theorem 4.3. An exact telescoping expansion gives

$$\widetilde{T}_D - T_D = \sum_{t=1}^{D} \widetilde{B}_D \cdots \widetilde{B}_{t+1} E_t B_{t-1} \cdots B_1.$$

Every unperturbed partial product is $B_{t-1} \cdots B_1 = I + \sum_{j<t} A_j$, so its norm is at most $1 + \sum_{j<t} a_j$. Each perturbed factor has norm at most $1 + a_j + \varepsilon_j$. Submultiplicativity therefore gives the uniform bound $\|\widetilde{T}_D - T_D\|_2 \le \rho_D$. Because $\mathbb{E}\|\mu\|_2^2 = d$, it follows that

$$\|(\widetilde{T}_D - T_D)\mu\|_{L^2} \le \sqrt{d}\,\rho_D, \qquad \|F\mu\|_{L^2} \le \sqrt{d}\,\varepsilon_*.$$

The reverse triangle inequality in L^2, together with Theorem 4.3, gives the square root of equation B.1; taking the nonnegative part and squaring proves the statement. □

For example, $E_t = \varepsilon I$ permits $(A_i + \varepsilon I)(A_j + \varepsilon I) = \varepsilon(A_i + A_j) + \varepsilon^2 I$, which is nonzero for $\varepsilon > 0$. For zero perturbations the bound recovers the original one; it does not assert that a trained Transformer or a nonlinear geometric benchmark belongs to this neighborhood.

B.6 Proof of Theorem B.4 (parameter efficiency)

For Theorem B.4, take $a > 0$, $B > 0$, integers $m, d \ge 1$ and $L \ge 0$. The context C and probe V are independent and uniform on $\{-a, a\}^L$ and $\{\pm\sqrt{d}\,e_j : 1 \le j \le d\}$, respectively; all L^2 errors refer to this joint law.

Theorem B.4 (Parameter efficiency)**.** *Let* $C \in \{-a, a\}^L$*,* $V \in \{\pm\sqrt{d}\,e_j\}$*,* $\sigma_0 = 1$*,* $\sigma_l(C) = \tanh(C_l)/\tanh(a)$*, and* $f_{\mathcal{A}}(C, V) = \sum_{l=0}^{L} \sigma_l(C) A_l V$ *with* $A_l \in \mathbb{R}^{m \times d}$ *and* $\sum_l \|A_l\|_F^2 \le B^2$*. A parameterization with outputs locally Lipschitz in its* P *parameters that represents every member exactly needs* $P \ge (L+1)md$*, which the matrix bank with native gates attains; with* b *bits per parameter, approximating every member within* $0 < \epsilon < B$ *in* L^2 *needs* $Pb \ge (L+1)md \log_2(B/\epsilon)$*.*

Proof. Write $p = (L+1)md$. The context C and probe V are sampled independently and uniformly from their grids, which defines the $L^2(C, V)$ error. The normalization of σ_l is absorbed into $A_l/\tanh(a)$, so the bank uses native tanh gates. Separate context and value channels make all coefficients identifiable: a constant-token set with token (C, V) supplies both, with the gate selecting C and the value projection selecting V. Rectangular maps can be padded into a square core. P counts the free matrix coefficients for fixed gates and fixed projections; learned gates, mixing, FFNs and output heads add parameters.

Step 1: identify all matrix coefficients. The random variables $\sigma_1(C), \ldots, \sigma_L(C)$ are independent symmetric signs. Together with $\sigma_0 = 1$ they satisfy $\mathbb{E}[\sigma_l \sigma_r] = \mathbf{1}_{l=r}$. Also $\mathbb{E}[VV^\top] = I_d$, independently of C. For any two banks, expansion of the squared prediction difference gives

$$\|f_{\mathcal{A}} - f_{\mathcal{A}'}\|_{L^2(C,V)}^2 = \mathbb{E}\left\|\sum_{l=0}^{L} \sigma_l(C)(A_l - A_l')V\right\|_2^2 \tag{B.2}$$

$$= \sum_{l=0}^{L} \|A_l - A_l'\|_F^2. \tag{B.3}$$

Thus the family is an isometric copy of the radius-B Euclidean ball in $\mathbb{R}^p$, not merely a list of matrices that might act identically on the observed input. In fact the coefficients are recovered by

$$A_l = \mathbb{E}[f_{\mathcal{A}}(C, V) V^\top \sigma_l(C)].$$

The expectation is a finite sum over the declared context/probe grid.

Step 2: a real-parameter lower bound with explicit regularity. For a predictor f_θ, define the coefficient extraction map

$$\Psi(\theta) = \left(\operatorname{vec} \mathbb{E}[f_\theta(C,V)V^\top \sigma_l(C)]\right)_{l=0}^{L} \in \mathbb{R}^p.$$

Since each grid output is locally Lipschitz in θ and the grid is finite, Ψ is locally Lipschitz. Exact representation of every bank requires the image of Ψ to contain the entire radius-B ball. If $P < p$, this is impossible. To see this without a parameter-counting heuristic, restrict to a bounded cube on which Ψ is Lipschitz with constant K. Cover the cube by $O(\eta^{-P})$ cubes of side η. Their images lie in p-dimensional balls of radius at most $K\sqrt{P}\eta$, of total outer volume $O(\eta^{p-P})$, which tends to zero. Local Lipschitz neighborhoods admit a countable cover by such bounded cubes, so the full image has zero p-dimensional volume. It cannot contain a positive-radius p-ball. Hence $P \geq p$. The bank coefficients themselves supply p real parameters and realize every member, attaining the bound. The $P = 0$ case has a singleton image and the same conclusion immediately.

Step 3: separate finite precision from real dimension. Let $\mathcal{H}$ be the finite-dimensional subspace spanned by the coordinate functions $\sigma_l(C)V_j$. Orthogonal projection of an arbitrary predictor onto $\mathcal{H}$ cannot increase its distance to any target in this family. By equation B.3, an ϵ-cover of the family therefore induces an ϵ-cover of the radius-B ball in $\mathbb{R}^p$. Comparing volumes, any such cover needs at least $(B/\epsilon)^p$ centers. A fixed model with P parameters and at most b bits per parameter has at most 2^{Pb} distinct predictions. Therefore

$$2^{Pb} \geq (B/\epsilon)^p, \qquad Pb \geq p \log_2(B/\epsilon).$$

This is the precision-dependent information cost. It does not add a $\log(1/\epsilon)$ factor to the real-parameter lower bound in Step 2. □

B.7 Proof of Corollary 4.4 (local approximation of any smooth operator)

Proof. **Step 1: linearize the specified operator.** Let $h = \mu - \mu_0$ and vectorize the matrix output. Taylor's formula with integral remainder gives

$$\operatorname{vec} T(\mu) = \operatorname{vec} T(\mu_0) + Jh + \int_0^1 (1-t) D^2 \operatorname{vec} T(\mu_0 + th)[h,h]\, dt, \qquad J = \sum_{l=1}^{r} s_l u_l v_l^\top,$$

with unit singular vectors. The segment lies in K, so the Euclidean norm of the remainder is at most $C_T\|h\|_2^2/2$. Let $U_l = \operatorname{mat}(u_l)$; then $\|U_l\|_F = 1$ and $\operatorname{mat}(Jh) = \sum_l s_l(v_l^\top h)U_l$. Vectorization preserves the Frobenius norm, so the linearized bank $\tilde{T}(\mu) = T(\mu_0) + \sum_l s_l(v_l^\top h)U_l$ satisfies $\|T(\mu) - \tilde{T}(\mu)\|_F \leq \frac{1}{2}C_T\|h\|_2^2$.

Step 2: realize it with native gates. The bank T_τ has generators $s_l U_l/\tau$ and gates with weight τv_l and bias $-\tau v_l^\top \mu_0$. At μ_0 every gate is zero and its derivative is τv_l, so the center and Jacobian agree exactly. For $t \geq 0$, $t - \tanh t = \int_0^t \tanh^2 s\, ds \leq t^3/3$; oddness gives the same absolute bound for negative t. Applying this with $t = \tau v_l^\top h$ bounds the discrepancy from the linearized bank by $\tau^2 \sum_l s_l |v_l^\top h|^3/3$ in Frobenius norm. Combining with the Taylor remainder and using $\|Ax\|_2 \leq \|A\|_F\|x\|_2$ proves the error bound. Smaller τ reduces the gate approximation term but increases the generator norms to s_l/τ.

Step 3: the direction count. For a fixed bank $B_0 + \sum_{l=1}^{L} a_l(\mu)B_l$ with differentiable coefficients, every operator difference lies in $\operatorname{span}\{B_l\}$, and the Jacobian $\sum_l \operatorname{vec}(B_l)\nabla a_l(\mu_0)^\top$ has rank at most L. Any bank with L directions has Jacobian rank at most L, so matching this one needs $L \geq r$. □

C Scalar Prediction and the Input Domain

This appendix proves the results of Section 4.2. Three resources must be distinguished: the number L of matrices in a context-conditioned bank, the number h of stored vector-generator slots, and the feature dimension D exposed to the final head. Appendices C–E give different requirements for these resources, in this order. In particular, a lower bound for recovering an entire operator or an unsigned collection of actions is not a lower bound for predicting one scalar label.

C.1 OBSERVABLE OPERATOR ERROR

How many matrix directions a model needs depends on what is observed. When full-rank value probes V with $\mathbb{E}[VV^\top \mid C] = \Sigma$ expose an operator $T(C)$, the best fixed L-direction bank with unrestricted square-integrable context coefficients has error $\sum_{j>L} \lambda_j$, a one-sided tail of the eigenvalues of $\mathrm{Cov}[\mathrm{vec}(T(C)\Sigma^{1/2})]$ that lower-bounds native tanh-gated banks (Theorem C.2). A scalar prediction exposes different quantities, which we now characterize (Figure 1, left). Let C be a random context, $V \in \mathbb{R}^d$ a value probe, and $T(C) \in \mathbb{R}^{m\times d}$ a measurable matrix with finite second moment. The target in this subsection is the vector $Y = T(C)V$. An admissible prediction has the form $\widehat{T}(C)V$, where

$$\widehat{T}(c) = B_0 + \sum_{j=1}^{L} a_j(c) B_j$$

and the B_j are fixed matrices. In the unrestricted coefficient class, the scalar functions a_j are measurable and square-integrable. They cannot depend on the probe V. The loss observes the full vector output.

Lemma C.1 (Probe identifiability). *For any square-integrable matrix error $A(C)$, put $Q(C) = \mathbb{E}[VV^\top \mid C]$. Then*

$$\mathbb{E}\|A(C)V\|_2^2 = \mathbb{E}\,\mathrm{tr}\big(A(C)Q(C)A(C)^\top\big). \tag{C.1}$$

If $\alpha I \preceq Q(C) \preceq \beta I$ almost surely, with $0 < \alpha \le \beta < \infty$, this loss lies between $\alpha\mathbb{E}\|A(C)\|_F^2$ and $\beta\mathbb{E}\|A(C)\|_F^2$. If $Q(C)$ is singular, the observation can fail to identify an operator even when its affine dimension is large.

Proof. Condition on C and expand $\|AV\|_2^2 = \mathrm{tr}(AVV^\top A^\top)$. The conditional expectation gives the identity. The two Loewner inequalities give the bounds by taking traces. For failure of identifiability, take $V = C \neq 0$ deterministically and $A(C) = I - CC^\top/\|C\|_2^2$ for $d \ge 2$. Then $A(C)C = 0$ although A is nonzero and varies with the direction of C. □

Theorem C.2 (Observable fixed-bank error). *Let a measurable $T(C) \in \mathbb{R}^{m\times d}$ have finite second moment, and suppose $\mathbb{E}[VV^\top \mid C] = \Sigma \succ 0$ almost surely. Let λ_j be the decreasing eigenvalues of $\mathrm{Cov}[\mathrm{vec}(T(C)\Sigma^{1/2})]$. Over fixed L-direction banks $\widehat{T}(c) = B_0 + \sum_{l=1}^{L} a_l(c)B_l$ with unrestricted square-integrable coefficients depending only on c,*

$$\inf_{\widehat{T}} \mathbb{E}\|[T(C) - \widehat{T}(C)]V\|_2^2 = \sum_{j>L} \lambda_j.$$

Proof. Suppose $Q(C) = \Sigma \succ 0$ almost surely; independence with second moment Σ is one sufficient way to satisfy this assumption. Let $x(C) = \mathrm{vec}(T(C)\Sigma^{1/2})$ and $K = \mathrm{Cov}(x(C))$, whose eigenvalues are $\lambda_1 \ge \cdots \ge \lambda_{md} \ge 0$. By Lemma C.1, prediction risk equals the mean squared Euclidean error in x. Any L-direction bank maps, after right multiplication by $\Sigma^{1/2}$, into an affine subspace of dimension at most L. For a direction subspace U with orthogonal projector P_U, pointwise orthogonal projection is optimal. The best affine translate can be chosen to pass through $\bar{x} = \mathbb{E}x$. To see this directly, the residual for a translate $b + U$ has mean-square norm

$$\mathbb{E}\|(I - P_U)(x - b)\|_2^2 = \mathrm{tr}((I - P_U)K) + \|(I - P_U)(\bar{x} - b)\|_2^2.$$

The second term is minimized at $b = \bar{x}$. Writing an orthonormal eigenbasis as u_i, one has

$$\mathrm{tr}(P_U K) = \sum_i \lambda_i \|P_U u_i\|_2^2 \le \sum_{i=1}^{L} \lambda_i,$$

because each coefficient is in $[0, 1]$ and their sum is at most L. Equality is attained by the span of the leading eigenvectors. Thus the optimal loss is exactly $\sum_{i>L} \lambda_i$. Reshaping the leading eigenvectors and right multiplying by $\Sigma^{-1/2}$ supplies the B_j; the coefficients are $a_j(c) = u_j^\top(x(c) - \bar{x})$. They are square-integrable by the moment assumption. This proves both the lower bound and attainability in the declared coefficient class. □

This is the classical least-squares projection principle underlying low-rank approximation (Eckart & Young, 1936), applied to a random *observable operator*; K can have large rank even when every $T(c)$ has rank one. If T is continuous, C has full support on the declared context domain and $\Sigma \succ 0$, then $\mathrm{rank}(K)$ equals $\mathrm{TDOF}(T)$ (Definition B.1), because $u^\top K u = 0$ exactly when $u^\top x(C)$ is almost surely constant, continuity and full support extend that equality to the whole domain, and taking orthogonal complements gives the affine-span equality. Without these assumptions K measures only almost-sure variation.

Corollary C.3 (Native gates attain the curve on an orthogonal family)**.** *Let $T(C) = B_0 + \sum_{j=1}^{r} \tanh(C_j) B_j$, where the C_j are independent, symmetrically distributed, and have $v_j = \mathbb{E}\tanh^2(C_j) > 0$. Suppose the nonzero matrices $B_j\Sigma^{1/2}$ are pairwise Frobenius-orthogonal. The nonzero observable eigenvalues are $v_j\|B_j\Sigma^{1/2}\|_F^2$. Retaining the L largest terms attains the optimal L-direction risk using native tanh gates themselves.*

Proof. Symmetry centers every gate and independence eliminates cross covariance. The covariance is the sum of the mutually orthogonal rank-one matrices $v_j \,\mathrm{vec}(B_j\Sigma^{1/2})\,\mathrm{vec}(B_j\Sigma^{1/2})^\top$. Deleting the remaining terms implements exactly the leading-eigenspace projection, and requires no approximation of a gate. □

For a general family, principal-component coefficients need not be single affine-tanh gates. Consequently, the spectral expression is always a lower bound for that restricted gate class, but its upper bound requires either the special structure above or a separate coefficient approximation.

With a readout direction u and $M_l = u b_l^\top / \|u\|^2$, so that $u^\top M_l x = b_l^\top x$, a native isolated signed MZ core with an affine scalar readout realizes every member of $\mathcal{G}_L$; suitable GLU and conditional-weight models realize it too. Parameters are finite for each function, with no uniform norm budget. Write $\mathcal{E}_L^P(Q) = \inf_{g\in\mathcal{G}_L} \mathbb{E}_P[(f_Q - g)^2]$.

C.2 A SHARP SCALAR PREDICTION BUDGET FOR NATIVE GATED CORES

We use f_Q, $\mathcal{G}_L$ (with $c, \theta_l \in \mathbb{R}$ and $a, b_l, w_l \in \mathbb{R}^d$), $\mathcal{E}_L^P$ and the zero-padded spectral lists α_j, β_j of Section 4.2, and write $n_+(Q), n_-(Q)$ for the numbers of positive and negative eigenvalues of Q. In the realization of $\mathcal{G}_L$ by an MZ core described there, the center matrix supplies the linear component and the scalar readout bias supplies the constant. Polynomial projections of fixed GLUs under Gaussian inputs are studied by Belrose & Rigg (2025, Sections 3.2 and 3.5), and spectral learning analyses of quadratic-activation networks by Ben Arous et al. (2025); the result here optimizes over the whole affine–tanh gated class, allows gate biases, and separates the positive and negative spectra.

Theorem 4.5 (Gaussian curve, full statement). *For $X \sim \mathcal{N}(0, I_d)$, $\mathcal{E}_L^{\mathrm{Gaussian}}(Q)$ is given by equation 4, and it is zero precisely when $L \geq \max\{n_+(Q), n_-(Q)\}$. It is an infimum, not in general an attained minimum: for $Q \neq 0$, no finite-parameter member of $\mathcal{G}_L$ equals f_Q on all of $\mathbb{R}^d$. A constructive sequence approaches the infimum using L native gates, with gate weights tending to zero and value coefficients potentially diverging.*

For a symmetric B, set $q_B(x) = x^\top Bx - \mathrm{tr}\, B$. Independence of the standard normal coordinates, their zero odd moments, and the Gaussian fourth moments $\mathbb{E}[X_iX_jX_kX_l] = \delta_{ij}\delta_{kl} + \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}$ give

$$\mathbb{E}[q_B(X)q_C(X)] = 2\,\mathrm{tr}(BC), \qquad \mathbb{E}[q_B(X)] = 0, \qquad \mathbb{E}[Xq_B(X)] = 0 : \tag{C.2}$$

in the product of the two quadratic forms, the first pairing cancels the product of traces and the other two give $2\,\mathrm{tr}(BC)$. The centered quadratics form a finite-dimensional closed subspace of $L^2(\mathcal{N}(0, I_d))$. Its orthogonal projection of a square-integrable function g is q_{B_g}, where

$$B_g = \tfrac{1}{2}\mathbb{E}[g(X)(XX^\top - I_d)]. \tag{C.3}$$

Indeed, testing this formula against every symmetric C and using equation C.2 gives the defining orthogonality identity. All expectations exist by Cauchy–Schwarz and the Gaussian fourth moment.

Lemma C.4 (One gate contributes at most one direction of each sign)**.** *The centered-quadratic projection matrix of $(b^\top X)\tanh(w^\top X + \gamma)$ has at most one positive and at most one negative eigenvalue. The statement also holds for an affine value factor $b^\top X + e$.*

Proof. If $w = 0$, the function is affine and its projection is zero. Otherwise write $w = tu$, where $t = \|w\| > 0$ and $\|u\| = 1$, and decompose $X = Zu + Y$, $b = \rho u + b_\perp$. Then $Z \sim \mathcal{N}(0,1)$ is independent of $Y \sim \mathcal{N}(0, I - uu^\top)$, and $b_\perp \perp u$. Put $k(Z) = \tanh(tZ + \gamma)$. In an orthonormal basis whose first vector is u, equation C.3 has the block form

$$B_g = \kappa u u^\top + \frac{\eta}{2}(u b_\perp^\top + b_\perp u^\top), \quad \kappa = \tfrac{1}{2}\mathbb{E}[(Z^2-1)(\rho Z + e)k(Z)], \quad \eta = \mathbb{E}[Zk(Z)].$$

To check the blocks, the u, u entry is the displayed one-dimensional expectation. The u, v entry for $v \perp u$ is $\mathbb{E}[Zk(Z)]\langle b_\perp, v\rangle/2$. Every $u^\perp, u^\perp$ entry is zero: the term $(\rho Z + e)k(Z)$ is independent of Y and its centered quadratic moment is zero; the remaining term is linear in Y and has zero Gaussian third moments. Thus $z^\top B_g z = 0$ for every $z \perp u$. A strictly positive subspace of dimension at least two would intersect $u^\perp$ in a nonzero vector, which is impossible; the same argument applies to a strictly negative subspace. □

Lemma C.5 (Inertia and optimal spectral truncation)**.** *If B is a sum of L symmetric matrices each having at most one positive and one negative eigenvalue, then $n_+(B), n_-(B) \le L$. Moreover, for every symmetric Q with spectral lists α_j, β_j,*

$$\min_{B = B^\top : n_+(B), n_-(B) \le L} \|Q - B\|_F^2 = \sum_{j>L} \alpha_j^2 + \sum_{j>L} \beta_j^2.$$

An optimizer Q_L keeps the L largest positive eigenvalues and the L largest negative magnitudes of Q, or all of a sign when there are fewer than L.

Proof. For each summand, its nonpositive eigenspace has codimension at most one. The intersection of these spaces has codimension at most L, and the sum has a nonpositive quadratic form on this intersection. A positive eigenspace of dimension greater than L would meet it nontrivially. This proves $n_+(B) \le L$; reverse signs for the negative bound.

For the spectral minimization, let λ_i and μ_i be the decreasing eigenvalues of Q and B, with orthonormal eigenvectors u_i, v_j. The matrix $P_{ij} = |u_i^\top v_j|^2$ is doubly stochastic. If $z_i = \sum_j P_{ij}\mu_j$, then $\sum_{i \le k} z_i \le \sum_{j \le k} \mu_j$: the coefficients $\sum_{i \le k} P_{ij}$ lie in $[0,1]$ and sum to k, so their weighted sum is maximized by selecting the k largest μ_j. Summation by parts, using the nonnegative gaps $\lambda_k - \lambda_{k+1}$, implies $\operatorname{tr}(QB) = \sum_i \lambda_i z_i \le \sum_i \lambda_i \mu_i$. Consequently $\|Q - B\|_F^2 \ge \sum_i (\lambda_i - \mu_i)^2$. Only the first L ordered μ_i can be positive and only the last L can be negative. A positive λ_i beyond the first L therefore contributes at least λ_i^2; a negative one outside the last L contributes at least λ_i^2. These are exactly the displayed two tails. The stated truncation has both inertia bounds and attains their sum, including when either tail is empty. □

Proof of Theorem 4.5. Each $g \in \mathcal{G}_L$ is square-integrable because its gates are bounded and its growth is at most affine. Its centered-quadratic projection is q_{B_g}, whose matrix is the sum of the one-gate matrices; the affine part projects to zero. The two lemmas and the contraction property of orthogonal projection give

$$\mathbb{E}[(f_Q - g)^2] \ge \|q_Q - q_{B_g}\|_{L^2}^2 = 2\|Q - B_g\|_F^2 \ge 2\left(\sum_{j>L} \alpha_j^2 + \sum_{j>L} \beta_j^2\right).$$

For attainability as a limit, let Q_L be the truncation above. Each positive/negative pair is one product:

$$\alpha u u^\top - \beta v v^\top = \operatorname{sym}\big((\sqrt{\alpha}u + \sqrt{\beta}v)(\sqrt{\alpha}u - \sqrt{\beta}v)^\top\big), \quad \operatorname{sym}(ab^\top) = \tfrac{1}{2}(ab^\top + ba^\top).$$

Unpaired positive eigenvalues use $a = b = \sqrt{\alpha}u$; unpaired negative ones use $a = \sqrt{\beta}v, b = -\sqrt{\beta}v$. Hence $Q_L = \sum_{l=1}^r \operatorname{sym}(a_l b_l^\top)$ with $r = \max\{n_+(Q_L), n_-(Q_L)\} \le L$. For $\tau > 0$ define

$$g_\tau(x) = c_* + \operatorname{tr}(Q - Q_L) + a_*^\top x + \sum_{l=1}^r \frac{\tanh(\tau a_l^\top x)}{\tau}(b_l^\top x).$$

It belongs to $\mathcal{G}_L$, padding unused gates by zero. The inequality $|\tanh(\tau s)/\tau| \le |s|$ bounds every summand by a quadratic polynomial. Such polynomials have finite Gaussian second moment, so dominated convergence in L^2 gives

$$g_\tau \longrightarrow c_* + \operatorname{tr}(Q - Q_L) + a_*^\top x + x^\top Q_L x.$$

The limiting residual is q_{Q-Q_L} and has the claimed squared norm by equation C.2. This proves equality of the infimum. Its zero criterion follows because both spectral tails vanish exactly when $L \geq \max\{n_+(Q), n_-(Q)\}$. Finally, a finite member has $|g(x)| \leq C(1 + \|x\|)$, whereas a nonzero quadratic grows as t^2 on some line $x = tu$. They cannot coincide everywhere. If they agreed Gaussian-almost surely, continuity and the strictly positive Gaussian density on open balls would imply equality everywhere, the same contradiction. This proves the non-attainment statement. □

Proposition C.6 (Rate and local curvature budget)**.** *The constructed g_τ satisfies*

$$\|f_Q - g_\tau\|_{L^2} \leq \sqrt{2\sum_{j>L}(\alpha_j^2 + \beta_j^2)} + \frac{\sqrt{105}}{3}\tau^2 \sum_{l=1}^{r} \|a_l\|^3 \|b_l\|.$$

Without a distributional assumption, the smallest number of gates in $\mathcal{G}_L$ needed to match the Hessian $2Q$ at $x = 0$ is $\max\{n_+(Q), n_-(Q)\}$; value and gradient can also be matched using the affine term.

Proof. For real s, $|s - \tanh s| \leq |s|^3/3$, by integrating $\tanh^2 s \leq s^2$. Thus the lth product error is bounded by $\tau^2 |a_l^\top X|^3 |b_l^\top X|/3$. Gaussian moment expansion or a decomposition of b_l parallel and perpendicular to a_l gives

$$\mathbb{E}[(a^\top X)^6 (b^\top X)^2] = 15\|a\|^6\|b\|^2 + 90\|a\|^4(a^\top b)^2 \leq 105\|a\|^6\|b\|^2.$$

Minkowski's inequality and the exact quadratic residual give the bound. For local curvature, the Hessian of $(b^\top x)\tanh(w^\top x + \gamma)$ at zero is $\operatorname{sech}^2(\gamma)(bw^\top + wb^\top)$, with at most one eigenvalue of either sign, so Lemma C.5 proves necessity. The paired construction with zero gate biases has Hessian $2Q$ for every $\tau > 0$; its product terms have value and gradient zero at zero, so the affine term independently sets those two quantities. □

Correlated Gaussian inputs. For $X \sim \mathcal{N}(0, \Sigma)$, $\Sigma \succ 0$, the change of variables $X = \Sigma^{1/2} Z$ is bijective on all gate and value coefficients, so the curve holds with $\Sigma^{1/2} Q \Sigma^{1/2}$ in place of Q. If Σ is singular, write $X = AZ$ with A of full column rank and Z standard normal of dimension $\operatorname{rank}\Sigma$, and use $A^\top Q A$; rank zero gives a constant input and zero best risk.

Quadratic diagnostic. For the interaction target $w^\top[M_0 + \sum_{l=1}^{K} x_l M_l]x$ with $K \leq d$ (Figure 2, left), the scalar quadratic matrix is $Q = \operatorname{sym}\sum_{l=1}^{K} e_l (M_l^\top w)^\top$. Its inertia can be smaller than K and is not identified by the reported aggregate scores.

Projection argument. The quadratic Gaussian projection of one gated-linear term has at most one eigenvalue of either sign, even with gate bias: it vanishes on the hyperplane orthogonal to the gate direction, which every two-dimensional definite subspace would meet. Summing L terms and projecting yields the lower bound; pairing positive and negative modes and taking $\tau^{-1}\tanh(\tau a^\top x) b^\top x \to (a^\top x)(b^\top x)$ attains it in the limit.

C.3 Prediction on a normalized input domain

Let X be uniform on $r\mathbb{S}^{d-1}$ with $d \geq 2$ and $r > 0$. We use $\mathcal{G}_L$, f_Q, $\mathcal{E}_L^{\mathrm{sphere}}$ and the ordered eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d$ of Q from Section 4.2, and put $\kappa_{d,r} = 2r^4/[d(d+2)]$.

Theorem 4.6 (Spherical curve, full statement). *$\mathcal{E}_L^{\mathrm{sphere}}(Q)$ is given by equation 5 when $2L < d$ and is zero when $2L \geq d$. In either case it equals*

$$\kappa_{d,r} \min_{\substack{B = B^\top,\ n_+(B) \leq L,\ n_-(B) \leq L \\ t \in \mathbb{R}}} \|Q - B - tI_d\|_F^2. \tag{C.4}$$

The least gate count for zero approximation infimum is $\min_{t\in\mathbb{R}} \max\{n_+(Q - tI_d), n_-(Q - tI_d)\}$. For a simple spectrum this count is $\lfloor d/2 \rfloor$; repeated eigenvalues can reduce it.

C.3.1 SPHERICAL QUADRATIC PROJECTION

For a matrix $C = C^\top$ with $\operatorname{tr} C = 0$, set $q_C(x) = x^\top Cx$. Rotational and reflection symmetry imply

$$\mathbb{E}[X_iX_j] = \frac{r^2}{d}\delta_{ij}, \qquad \mathbb{E}[X_iX_jX_kX_l] = \frac{r^4}{d(d+2)}(\delta_{ij}\delta_{kl} + \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}).$$

For the fourth-moment constants, let $a = \mathbb{E}X_1^4$ and $b = \mathbb{E}X_1^2X_2^2$. Applying the rotation $(X_1, X_2) \mapsto ((X_1+X_2)/\sqrt{2}, (X_1-X_2)/\sqrt{2})$ and cancelling odd moments gives $a = (2a+6b)/4$, hence $a = 3b$. Squaring $\sum_i X_i^2 = r^2$ gives $da+d(d-1)b = r^4$. These two equations yield the displayed constants. Other fourth moments vanish unless indices pair, by reflection symmetry.

It follows by expansion that, for traceless symmetric C, D,

$$\mathbb{E}[q_C(X)q_D(X)] = \kappa_{d,r}\operatorname{tr}(CD), \qquad \mathbb{E}q_C(X) = 0, \quad \mathbb{E}[Xq_C(X)] = 0. \tag{C.5}$$

Thus traceless quadratic forms constitute a finite-dimensional closed subspace, orthogonal to all affine functions. Its orthogonal projection of any square-integrable g has coefficient matrix

$$C_g = \kappa_{d,r}^{-1}\mathbb{E}\left[g(X)\left(XX^\top - \frac{r^2}{d}I_d\right)\right]. \tag{C.6}$$

This matrix is symmetric and traceless. Taking its inner product with any traceless C and using equation C.5 proves the projection identity. All expectations exist because X is bounded and g is square-integrable.

Lemma C.7 (One gated term has an inertia-constrained representative)**.** *For $g(x) = (b^\top x + e)\tanh(w^\top x + \gamma)$ there is a symmetric matrix M with $n_+(M), n_-(M) \le 1$ such that $C_g = M - \operatorname{tr}(M)I_d/d$.*

Proof. If $w = 0$, g is affine and $C_g = 0$. Otherwise put $u = w/\|w\|$, $Z = u^\top X$, $Y = X - Zu$, and $b = \rho u + b_\perp$. Conditional on Z, the distribution of Y is reflection-symmetric and rotationally invariant in $u^\perp$, with

$$\mathbb{E}[Y \mid Z] = 0, \qquad \mathbb{E}[YY^\top \mid Z] = \frac{r^2 - Z^2}{d-1}(I_d - uu^\top).$$

The possible poles $|Z| = r$ have probability zero for $d \ge 2$; the formulas remain valid there as zero covariance. All conditional third moments of Y vanish by symmetry. Substitution into equation C.6 shows that the $u^\perp$ block of C_g is a scalar matrix, its u–$u^\perp$ block is a multiple of $b_\perp$, and its u, u entry is some a. Since its trace is zero, it has the form

$$C_g = auu^\top - \frac{a}{d-1}(I_d - uu^\top) + \eta(ub_\perp^\top + b_\perp u^\top).$$

For example, the off-diagonal coefficient is $\eta = \kappa_{d,r}^{-1}\mathbb{E}[Z\tanh(\|w\|Z + \gamma)(r^2 - Z^2)/(d-1)]$. Set $M = da\,uu^\top/(d-1) + \eta(ub_\perp^\top + b_\perp u^\top)$. Then $C_g = M - \operatorname{tr}(M)I_d/d$. Moreover $v^\top Mv = 0$ on $u^\perp$, so, as in Lemma C.4, M has at most one eigenvalue of each sign. □

C.3.2 LOWER BOUND AND EXACT SPECTRAL MINIMIZATION

By this lemma and the first part of Lemma C.5, the quadratic projection of any $g \in \mathcal{G}_L$ is represented by $B_0 = B - \operatorname{tr}(B)I_d/d$ for a B with $n_+(B), n_-(B) \le L$. With $Q_0 = Q - \operatorname{tr}(Q)I_d/d$, contraction of orthogonal projection gives

$$\mathbb{E}[(f_Q - g)^2] \ge \kappa_{d,r}\|Q_0 - B_0\|_F^2 \tag{C.7}$$

$$= \kappa_{d,r}\min_{t\in\mathbb{R}}\|Q - B - tI_d\|_F^2. \tag{C.8}$$

The second equality follows by minimizing a quadratic polynomial in t; its minimizing value is $\operatorname{tr}(Q - B)/d$.

If $2L < d$, the second part of Lemma C.5, applied to $Q - tI_d$ in place of Q, bounds $\|Q - tI_d - B\|_F^2$ below by the two spectral tails of $Q - tI_d$. Every index $L < i \le d - L$ with $\lambda_i \ne t$ lies in one of these tails. Hence

$$\|Q - tI_d - B\|_F^2 \ge \sum_{i=L+1}^{d-L}(\lambda_i - t)^2 \ge \sum_{i\in I_L}(\lambda_i - \bar{\lambda}_L)^2.$$

This is attained by $t = \bar{\lambda}_L$ and by taking B diagonal in a Q eigenbasis, with entries $\lambda_i - t$ for $i \le L$ and $i > d - L$, and zero elsewhere. The first group is nonnegative and the last nonpositive because t lies between the extreme central eigenvalues. The residual $Q - tI_d - B$ is traceless. If $2L \ge d$, choose t to be a median eigenvalue (or between the two middle eigenvalues). Then $B = Q - tI_d$ has both inertia counts at most L and the residual vanishes. This proves the minimization in equation C.4.

C.3.3 Attaining sequence

Let t, B be the optimizer just constructed. Pairing its positive and negative eigenmodes as in the proof of Theorem 4.5 gives $B = \sum_{\ell=1}^{k} \mathrm{sym}(a_\ell b_\ell^\top)$ with $k \le L$. Define

$$g_\tau(x) = c_* + r^2 t + a_*^\top x + \sum_{\ell=1}^{k} \frac{\tanh(\tau a_\ell^\top x)}{\tau} (b_\ell^\top x), \quad \tau > 0.$$

Unused gates can be zeroed. The inequality $|u - \tanh u| \le |u|^3/3$ from the proof of Proposition C.6 implies the uniform bound

$$\sup_{\|x\|=r} |g_\tau(x) - (c_* + r^2 t + a_*^\top x + x^\top B x)| \le \frac{\tau^2 r^4}{3} \sum_\ell \|a_\ell\|^3 \|b_\ell\|. \tag{C.9}$$

Consequently the sequence converges uniformly and in L^2 to the displayed polynomial. Its residual is $x^\top (Q - tI_d - B)x$, with zero trace, so equation C.5 gives exactly the optimum in equation C.4. This closes the upper and lower bounds. All parameter values are finite at any fixed τ. The zero-risk gate criterion follows from $Q = tI_d + B$ and the inertia constraints. For a simple spectrum a shift at a middle eigenvalue leaves $\lfloor d/2 \rfloor$ eigenvalues on the larger side, and no shift can reduce that maximum; this proves the stated generic count. With per-pair scales $\tau_\ell = \sqrt{s_\ell}/B_v$ for a value budget $B_v > 0$, where $s_\ell = \|a_\ell\|^2 = \|b_\ell\|^2$, each value coefficient has norm B_v and each gate weight norm s_ℓ/B_v, and the proof of equation C.9 gives $\|f_Q - g\|_{L^2} \le \sqrt{\mathcal{E}_L^{\mathrm{sphere}}(Q) + r^4 \sum_\ell s_\ell^3/(3B_v^2)}$.

LayerNorm correspondence. Let $D \ge 3$, $Z \sim \mathcal{N}(0, I_D)$, let N_0 be the zero-epsilon row normalization of Section 4.4, and let $U \in \mathbb{R}^{D \times (D-1)}$ have orthonormal columns spanning $\mathbf{1}^\perp$. Then

$$U^\top N_0(Z) = \sqrt{D}\, \frac{U^\top Z}{\|U^\top Z\|}$$

is uniform on the radius-$\sqrt{D}$ sphere in $\mathbb{R}^{D-1}$. This follows from the rotationally invariant Gaussian density and the independence of direction from radial integration; the event of zero denominator has probability zero. Thus the theorem applies, in these intrinsic coordinates, to $U^\top Q U$ and any gates and values restricted to this subspace. Restriction and lifting by U preserve the native function class.

C.3.4 Which part depends on the activation?

Corollary C.8 (The curves extend beyond tanh)**.** *Let $\phi : \mathbb{R} \to \mathbb{R}$ be globally Lipschitz and differentiable at some θ_0 with $\phi'(\theta_0) \ne 0$. Replace tanh by ϕ in $\mathcal{G}_L$, allowing gate biases and a free affine output term. The Gaussian curve equation 4 and spherical curve equation 5 are unchanged. When $\theta_0 = 0$, the upper constructions need no gate bias.*

Proof. Lipschitzness implies $|\phi(t)| \le |\phi(0)| + K|t|$ for a finite K. Each gated-linear term consequently grows at most quadratically and is square-integrable under the Gaussian measure, as well as on the sphere. In the Gaussian projection proof in Appendix C.2, replace $k(Z)$ by $\phi(\|w\| Z + \gamma)$. Conditional Gaussian symmetry still makes the $u^\perp, u^\perp$ block vanish, and the projection matrix has form $auu^\top + \eta(ub_\perp^\top + b_\perp u^\top)$. It therefore has at most one eigenvalue of either sign. In the spherical proof above, the same substitution leaves the $u^\perp$ block isotropic, giving the same inertia-constrained representative modulo an identity matrix. Thus both lower bounds remain valid for the entire new class.

For each paired bilinear factor from the corresponding spectral optimizer, use

$$h_\tau(x) = \frac{\phi(\theta_0 + \tau a^\top x) - \phi(\theta_0)}{\tau \phi'(\theta_0)} (b^\top x).$$

This uses one gate and an affine correction, since the subtracted term is a fixed scalar times $b^\top x$. It tends pointwise to $(a^\top x)(b^\top x)$ by differentiability at θ_0, and

$$|h_\tau(x)| \le \frac{K}{|\phi'(\theta_0)|}|a^\top x|\,|b^\top x|.$$

The squared dominating polynomial is integrable for the Gaussian law and bounded on the sphere. Dominated convergence in L^2, applied to the finite sum, therefore supplies the same attaining sequence. The constant and linear corrections are absorbed into the free affine term for every finite τ. □

Tanh, sigmoid and exact GELU satisfy these assumptions; for GELU, $\phi'(t) = \Phi_{\mathrm{normal}}(t) + t\varphi_{\mathrm{normal}}(t)$ is bounded and $\phi'(0) = 1/2$. Biased ReLU also qualifies by choosing $\theta_0 > 0$; on the sphere its construction is exactly bilinear for a sufficiently small fixed τ, since every preactivation remains in its positive linear region, so its approximation optimum is attained. The tanh-specific finite-τ rate does not transfer under these assumptions alone: a nonzero second derivative at the expansion point generally leaves an $O(\tau)$ product error.

C.3.5 Boundary cases of the input law

The curves depend on the input law, not only on its symmetry or radius. The Gaussian assumption of Theorem 4.5 cannot simply be dropped: for independent Rademacher coordinates and $Q = I$, $X^\top QX = d$ is constant, so an affine predictor has zero error although the Gaussian $L = 0$ expression is $2d$. A fixed radius alone does not suffice for Theorem 4.6: for a uniform distribution on $\{\pm re_i\}_{i=1}^d$ and $Q = e_1e_2^\top + e_2e_1^\top$, the target is identically zero, whereas equation 5 at $L = 0$ is positive. Rotational invariance alone does not preserve the one-gate inertia bound either.

Proposition C.9 (A two-shell boundary)**.** *Let U be uniform on $\mathbb{S}^2$ and independent of R, where $\Pr(R = 1) = \Pr(R = 2) = 1/2$, and put $X = RU$. For all sufficiently large finite $t > 0$, the orthogonal projection of the single native term $g_t(X) = X_1 \tanh(tX_1)$ onto the space of centered quadratic forms has a positive-definite coefficient matrix. Thus the Gaussian one-positive/one-negative projection lemma need not hold for a rotationally invariant input law.*

Proof. Write $m_j = \mathbb{E}R^j$. Here $(m_1, m_2, m_3, m_4) = (3/2, 5/2, 9/2, 17/2)$. Rotational symmetry and the spherical moments above give, for symmetric B, C,

$$\mathrm{Cov}(X^\top BX, X^\top CX) = \frac{17}{15}\operatorname{tr}(BC) - \frac{23}{180}\operatorname{tr}(B)\operatorname{tr}(C).$$

This bilinear form is positive definite on symmetric matrices: its traceless part has coefficient $17/15$, while on diagonal trace coordinates the corresponding eigenvalue is $3/4$. Quadratic projection therefore has a unique coefficient matrix.

The bounded functions g_t tend in L^2 to $g_\infty(X) = |X_1|$. Conditional on a shell of radius R, X_1 is uniform on $[-R, R]$, and $\mathbb{E}[X_2^2 \mid X_1, R] = (R^2 - X_1^2)/2$. Direct integration gives

$$\mathbb{E}g_\infty = m_1/2, \qquad \mathbb{E}[g_\infty X_1^2] = m_3/4, \qquad \mathbb{E}[g_\infty X_2^2] = \mathbb{E}[g_\infty X_3^2] = m_3/8.$$

Subtracting $\mathbb{E}g_\infty\,\mathbb{E}X_j^2$, where $\mathbb{E}X_j^2 = m_2/3$, gives covariance vector $(1/2, -1/16, -1/16)^\top$. Reflection and interchange symmetry make the projection matrix diagonal with its last two entries equal. Its normal equations are

$$\left(\frac{17}{15}I_3 - \frac{23}{180}\mathbf{1}\mathbf{1}^\top\right)\beta = \begin{pmatrix}1/2\\ -1/16\\ -1/16\end{pmatrix}, \qquad \beta = \begin{pmatrix}203/408\\ 1/816\\ 1/816\end{pmatrix}.$$

All three entries are strictly positive. Orthogonal projection is continuous in L^2; since the coefficient Gram matrix is nonsingular, the finite-t coefficient matrices converge to $\operatorname{diag}\beta$. Eigenvalues are continuous, so the finite-t projections are positive definite for all sufficiently large t. Bounded convergence is justified by $|g_t| \le |X_1| \le 2$. Each such t is one finite native gate with zero bias. □

Why the answer changes. On the sphere, $x^\top(tI)x = tr^2$ is constant. Quadratic information is therefore represented modulo the identity. The spherical quadratic projection of a gated-linear term is an

inertia-$(1,1)$ matrix modulo an identity shift. The best L terms retain the outer L eigenvalues on each side; the middle eigenvalues are approximated by their common mean. For $Q = \mathrm{diag}(4,2,1,-3)$ and $L = 0,1,2,3$, the Gaussian errors are $60, 10, 2, 0$, whereas on $2\mathbb{S}^3$ they are $34.7, 0.67, 0, 0$: zero error needs three gates in one domain and two in the other.

For isotropic Gaussian features, zero-epsilon LayerNorm produces a uniform sphere in $\mathbf{1}^\perp$: in an orthonormal basis U of that subspace, $U^\top N_0(Z)$ has radius $\sqrt{D}$ and dimension $D-1$. Applying equation 5 to $U^\top QU$ is exact in this case. Both curves extend to globally Lipschitz gates with a nonzero derivative at some point, including suitable GLU variants (Corollary C.8); they characterize the multiplicative form and the input law, not a tanh-specific advantage.

C.4 Training check of the exact approximation curves

Theorems 4.5 and 4.6 give infima over $\mathcal{G}_L$ and do not assert that training attains them. We trained $\mathcal{G}_L$ in the parameterization of Section 4.2 with Adam, drawing fresh inputs at every step (20,000 steps of 4,096 points, learning rate 3×10^{-3} with cosine decay). For each of three seeds, four initialization scales were trained and the one with the lowest error on a validation sample was kept; errors are measured on an independent sample of 2^{19} points. Table 1 reports the example of Figure 1. Over these eight settings and sixteen more with four random symmetric $Q \in \mathbb{R}^{6\times 6}$ (eigenvalues drawn from $\mathcal{N}(0,4)$, $L \in \{1,2\}$), no trained error falls below the infimum, and the mean excess stays below 0.6% of the target variance. The trained error collapses at the predicted gate counts: three under the Gaussian law and two on the sphere.

The complete gate-budget comparison is shown in Table 1 in Section 5.2.

D Summaries and Continued Computation

This appendix proves the results of Section 4.3.

D.1 What fixed generator mixing preserves

Fix one feature coordinate of the L injected actions in equation 3, and write their signed amplitudes as $a \in \mathbb{R}^L$. With zero inherited generators, mixing gives Wa, where $W \in \mathbb{R}^{h\times L}$ is fixed after training. The uncompressed IHW coordinate is $t(a) = \|a\|_1$, while the compressed IHW is $\|Wa\|_1$. An additional center feature is allowed below only if it is constant on the compared amplitude domain. Equivalently, the claims can be applied on a fiber where that side feature agrees and the stated amplitude domain is attainable. This requirement cannot be dropped for a complete network.

Theorem D.1 (Signed and magnitude preservation)**.** *Fix $W \in \mathbb{R}^{h\times L}$ and constant center-side information on the compared amplitude domain. The equality $\|Wa\|_1 = \|a\|_1$ for every $a \in \mathbb{R}^L$ holds exactly when the columns of W have disjoint supports and unit ℓ_1 norms, which requires $h \geq L$. If $h < L$ and $R > 0$, the best decoder of Wa for $\|a\|_1$ on $\|a\|_1 \leq R$ has worst-case absolute error $R/2$. One slot instead preserves the signed sum exactly.*

Proof. Suppose $\|Wa\|_1 = \|a\|_1$ for every $a \in \mathbb{R}^L$. Testing $a = e_j$ shows that every column w_j has unit ℓ_1 norm. For distinct i, j, testing $e_i + e_j$ and $e_i - e_j$ requires equality in both triangle inequalities. But in each row the identity

$$|s+t| + |s-t| = 2\max\{|s|,|t|\}$$

implies that their two norms sum to $4 - 2\sum_g \min\{|W_{gi}|, |W_{gj}|\}$. Both norms can equal two only if no row is nonzero in both columns. Conversely, disjoint column supports and unit column norms give $\|Wa\|_1 = \sum_j |a_j| \|w_j\|_1 = \|a\|_1$. There must be at least L rows to hold L nonempty disjoint supports. This finite-dimensional characterization is an instance of the classical disjoint-support structure of ℓ_p isometries (Lamperti, 1958).

For the stronger statement about arbitrary decoders, suppose $h < L$ and fix $R > 0$. Choose a nonzero $z \in \ker W$ and rescale it so that $\|z\|_1 = R$. The inputs 0 and z have the same full compressed vector $Wa = 0$, but their targets are 0 and R. Therefore any decoder D has worst-case

absolute error at least $R/2$. The constant decoder $D \equiv R/2$ attains that error on the whole ball $\|a\|_1 \le R$. Thus

$$\inf_D \sup_{\|a\|_1 \le R} |D(Wa) - \|a\|_1| = R/2, \qquad h < L. \tag{D.1}$$

The same conclusion holds if D sees only compressed IHW, which is a function of Wa. On the equiprobable pair $\{0, z\}$, the minimum squared risk is $R^2/4$, by completing the square around $R/2$. In contrast, the signed target $\sum_l a_l$ is recovered exactly with $h = 1$ and $W = (1, \dots, 1)$. Applying this construction in every output coordinate recovers $\sum_l \gamma_l(c) M_l v$ in one vector slot. □

Proposition D.2 (Restricted domains and quantitative cancellation)**.** *Let $U \subseteq \mathbb{R}^L$ be an r-dimensional linear subspace. Some decoder recovers $\|a\|_1$ from Wa for every $a \in U$ if and only if* $\ker W \cap U = \{0\}$. *Consequently $h \ge r$ is necessary, and $h = r$ is sufficient with a suitable W and an unrestricted decoder. For unit-ℓ_1 columns of W, define $\kappa_{ij} = \sum_g \min(|W_{gi}|, |W_{gj}|)$. At least one of $a = e_i + e_j$ and $a = e_i - e_j$ has IHW underestimation $\|a\|_1 - \|Wa\|_1 \ge \kappa_{ij}$.*

Proof. A nonzero kernel vector and zero are indistinguishable and have different norms. Conversely, injectivity on U provides a linear inverse on $W(U)$; composing it with the norm gives a decoder. An isomorphism from U to $\mathbb{R}^r$ realizes sufficiency. The rowwise identity used above shows that the two deficits sum to $2\kappa_{ij}$, so one is at least κ_{ij}. Unit column norms make both deficits nonnegative. □

The distinction between recovering the norm with a decoder and preserving it *as IHW* is essential: injectivity alone need not give an ℓ_1 isometry. The dimension obstruction also needs sign-rich domains. On $a \ge 0$, $W = (1, \dots, 1)$ preserves the norm with one slot. If the mixing itself may depend on a, the choice $W(a) = \operatorname{sign}(a)^\top$ does so on all of $\mathbb{R}^L$. If a center branch already supplies $\|a\|_1$, there is no prediction obstruction. These are explicit boundary cases, not exceptions to a purported lower bound on all Transformers.

D.2 Persistent state versus an immediately readable summary

Let $\mathcal{S}$ be a state space, $R : \mathcal{S} \to \mathcal{Z}$ a reduction, and $F : \mathcal{S} \to \mathcal{S}$ a fixed continuation. A reduced update $\bar{F}$ obeying $R \circ F = \bar{F} \circ R$ exists exactly when

$$R(s) = R(t) \quad \Longrightarrow \quad R(F(s)) = R(F(t)). \tag{D.2}$$

Necessity follows by applying $\bar{F}$; for sufficiency define $\bar{F}$ on each attained value $R(s)$ by $R(F(s))$, which is well-defined by the implication. This is the elementary fiber criterion of exact lumping (Ovchinnikov et al., 2021, Section 2); general set-embedding limits (Wagstaff et al., 2019) provide related background.

Proposition D.3 (No nontrivial linear merging through a coordinate nonlinearity)**.** *Let $\phi : \mathbb{R} \to \mathbb{R}$ be continuously differentiable with nonconstant derivative, and let $P \in \mathbb{R}^{r \times n}$. Apply ϕ coordinatewise. There exists a function Ψ on $P\mathbb{R}^n$ such that $P\phi(x) = \Psi(Px)$ for every $x \in \mathbb{R}^n$ if and only if every vector in* $\ker P$ *is supported only on zero columns of P. Equivalently, the nonzero columns of P are linearly independent. Thus a linear reduction can discard unused coordinates, but cannot nontrivially merge used coordinates while preserving all such updates.*

Proof. Suppose a reduction exists. For $z \in \ker P$, the vector $P\phi(x + tz)$ is constant in t. Differentiating at zero gives $P \operatorname{diag}(\phi'(x))z = 0$ for every x. Change only x_j between two values at which ϕ' differs, and subtract the two identities. It follows that $z_j P e_j = 0$. Whenever column j is nonzero, every $z \in \ker P$ therefore has $z_j = 0$. This is the claimed support condition. Conversely, under that condition $Px = Py$ implies $x_j = y_j$ on every nonzero column. Zero columns do not affect $P\phi$, so $P\phi(x) = P\phi(y)$; apply equation D.2. A dependence among the nonzero columns would supply a forbidden kernel vector and vice versa, proving the equivalent formulation. This covers $P = 0$ and arbitrary rank. □

The obstruction need not be numerically large when the state spread is small: if $\phi : \mathbb{R}^d \to \mathbb{R}^k$ has Hessian operator norm at most C on a convex region containing rows $g_1, \dots, g_h$ and their mean $\bar{g}$, Taylor's integral remainder, whose first-order terms cancel, gives $\| \sum_i \phi(g_i) - h\phi(\bar{g}) \| \le (C/2) \sum_i \|g_i - \bar{g}\|^2$.

With S, H as in Section 4.3, P_D, N_ϵ as in Section 4.4 (nonconstant inputs when $\epsilon = 0$) and $A_s = \sum_{j=0}^{s-1} \epsilon^j$, let $F_{\epsilon,s}$ apply $s \geq 1$ unit-gain, zero-bias normalizations with the same epsilon to every row of $G \in \mathbb{R}^{h \times D}$.

Theorem 4.7 (Joint-summary continuation obstruction, full statement). *For every $D \geq 3$, $h \geq 2$, $s \geq 1$ and finite $\epsilon \geq 0$, there are generator states $G, \widetilde{G}$ with strictly positive rows of nonzero variance, identical center-side state, $S(G) = S(\widetilde{G})$ and $H(G) = H(\widetilde{G})$, whose continued IHW the scalar readout $\ell(z) = 2z_1 - z_3$ separates by*

$$\Delta_{D,\epsilon,s} = \frac{8}{3\sqrt{A_s\, 8/(3D) + \epsilon^s}} > 0. \tag{D.3}$$

Every predictor using only the identical center, S and H of the original state has worst-case absolute error at least $\Delta_{D,\epsilon,s}/2$ on these two continued-readout targets. Its minimum squared error on the equiprobable pair is $\Delta^2_{D,\epsilon,s}/4$.

Proof. First use $h = 2, D = 3$ and

$$G = \begin{pmatrix} 3 & 1 & 1 \\ 1 & 3 & 1 \end{pmatrix}, \qquad \widetilde{G} = \begin{pmatrix} 5/2 & 5/2 & 1 \\ 3/2 & 3/2 & 1 \end{pmatrix}.$$

All entries are positive, so both reductions equal the column sum $(4, 4, 2)$. The centered rows of G are $(4/3, -2/3, -2/3)$ and $(-2/3, 4/3, -2/3)$, each with squared norm $8/3$. Those of $\widetilde{G}$ are $(1/2, 1/2, -1)$ and $(1/6, 1/6, -1/3)$, each nonzero. For $D > 3$, append to each row its own first-three-coordinate mean, repeated $D-3$ times. The sums of these appended means agree: $5/3+5/3 = 2 + 4/3$. Positivity ensures IHW agrees as well. The added centered coordinates are zero, so the centered squared norms remain unchanged. Thus, by equation E.1, each normalized G row has denominator $a_G = \sqrt{A_s\, 8/(3D) + \epsilon^s}$ after s normalizations.

The first three coordinates of $H(F_{\epsilon,s}(G))$ are $(2, 2, 4/3)/a_G$. For $\widetilde{G}$, they are $(t, t, 2t)$ for a positive t: both centered rows are positive multiples of $(1, 1, -2)$ and normalization multiplies each by a positive scalar. Therefore ℓ is $8/(3a_G)$ on the first continued IHW and zero on the second. This proves equation D.3 for every finite epsilon. For $h > 2$, append identical positive nonconstant rows to the two states. Their contributions to S, H and to the continued IHW agree, so the difference remains Δ; no zero-variance row is needed.

Any reduced-state predictor must assign the same scalar z to the two inputs. For two target values differing by Δ, at least one error is $\Delta/2$. Completing the square shows $\frac{1}{2}[(z - y_0)^2 + (z - y_1)^2] = (z-(y_0+y_1)/2)^2+\Delta^2/4$. The midpoint attains both bounds on the pair. This proves the quantitative impossibility statement. □

Connection to the actual generator path. As in Appendix E.3.4, zero attention and generator-FFN contributions with the generator residual retained leave unit-gain, zero-bias rowwise LayerNorms as the continuation, two per block ($s = 2$). A shared affine generator input projection can map a scalar t to $G + t(\widetilde{G} - G)$, absorbing the input scaling factor into its weights and bias. Set the center projection to a constant. Then inputs $t = 0, 1$ realize the two states with identical centers, signed sums and IHW. For $D > 3$ the mean-padding is affine as well. A fixed linear readout of continued IHW yields the gap, and two GELU hidden units implement it exactly (Appendix E.3.4).

The obstruction is not specific to LayerNorm. For any continuously differentiable coordinatewise ϕ with nonconstant derivative, a linear reduction P admits an update Ψ with $P\phi(x) = \Psi(Px)$ for every x exactly when the nonzero columns of P are linearly independent (Proposition D.3). The row sum S merges coordinates, so $\sum_i \mathrm{GELU}(g_i)$ is not a function of $\sum_i g_i$: a rowwise GELU cannot be continued from the summed state either.

E NORMALIZED READOUT CAPACITY AND ACCESS

This appendix proves the results of Section 4.4. Fix $s \geq 1$ and $D \geq 2$, with P_D and N_ϵ as in Section 4.4, and set $A_s(\epsilon) = \sum_{j=0}^{s-1} \epsilon^j$. Direct induction gives

$$N_\epsilon^{\circ s}(x) = \frac{P_D x}{\sqrt{A_s(\epsilon)\|P_D x\|^2/D + \epsilon^s}}. \tag{E.1}$$

For the induction, a centered numerator $y = P_D x$ with denominator a_k is normalized with new denominator squared $\|y\|^2/D + \epsilon a_k^2$; this solves the recurrence. At zero epsilon the formula is used only for $y \neq 0$.

E.1 MULTIPLE NORMALIZED ROWS RECOVER DIMENSIONS LOST BY A SINGLE ROW

We use the interface Φ_ϵ and the assumptions of Section 4.4, with $D \geq 2$, $h \geq 1$ and an integer $s \geq 1$. In the upper bound the row encodings may be arbitrary differentiable maps, and additional affine row gains and biases are allowed; the construction does not need them.

Theorem 4.8 (Sharp local dimension, full statement). *At zero epsilon, $m_{\max}(D, h)$ in equation 6 is the largest dimension of a continuously differentiable variable that can be recovered locally with a differentiable decoder through this interface. Every dimension up to this bound is attained by affine encodings of the rows on a neighborhood of the origin. In particular, two rows suffice for full D-dimensional local recovery when $D \geq 4$, and three rows suffice when $D = 3$. No positive-dimensional smooth local recovery is possible for $D = 2$ at zero epsilon, irrespective of h. For the full-dimensional attaining constructions, the encoding can be chosen independently of epsilon, with noncollapsing row variances and bounded encoding derivatives, and for each fixed D, h, s there are $\rho > 0$ and $\epsilon_0 > 0$ such that on $\|q\| \leq \rho$ and $0 \leq \epsilon \leq \epsilon_0$,*

$$\tfrac{1}{2}\|q - q'\| \leq \|\Phi_\epsilon(q) - \Phi_\epsilon(q')\| \leq \tfrac{3}{2}\|q - q'\|. \tag{E.2}$$

Thus the local inverse is uniformly Lipschitz with constant two as epsilon tends to zero.

For two rows and $D \in 4\mathbb{N}$, Theorem 4.10 (Appendix E.3) makes the radius explicit and independent of D, for every $\epsilon \in [0, 1]$.

Upper bound. At any nonconstant row x, the Jacobian of N_0 has rank $D-2$: it annihilates the shift direction and the centered radial direction and is a nonzero scalar on their orthogonal complement. Repeated zero-epsilon normalization changes neither its image dimension nor the normalized row. The derivative of a row gain, bias and locally smooth absolute value cannot increase that rank. The derivative of the sum of h such row maps therefore has rank at most $h(D-2)$, and its output dimension gives the additional bound D. A differentiable local decoder recovering $q \in \mathbb{R}^m$ must compose with this derivative to I_m, hence $m \leq \min\{D, h(D-2)\}$. This proves the necessary condition for every differentiable encoding, not only the affine constructions below. For $D = 2$ each nonconstant normalized row is locally one of two points; the rank bound is zero. □

Signed tangent directions. Let $z \in \mathbf{1}^\perp$ have no zero coordinates and $\|z\|^2 = D$. At this unit-variance row,

$$DN_0(z) = \Pi_z = I - \mathbf{1}\mathbf{1}^\top/D - zz^\top/D.$$

Let $S_z = \operatorname{diag}(\operatorname{sign} z)$. The image of the derivative after absolute value is

$$\mathcal{T}_z = S_z \operatorname{im} \Pi_z, \qquad \mathcal{T}_z^\perp = \operatorname{span}\{\operatorname{sign} z, |z|\}.$$

Indeed, S_z is an orthogonal sign matrix, so it maps the two normal directions $\mathbf{1}, z$ to $\operatorname{sign} z, |z|$. This formula identifies when different rows supply complementary tangent directions.

One row. Its signed tangent image has dimension $D-2$. Choose a matrix B with columns forming a basis of the tangent space and encode $E(q) = z + Bq$. The derivative $S_z \Pi_z B$ has full column rank $D-2$. This realizes the bound for $h = 1$ locally; the inverse justification below also covers this rectangular case.

Three-dimensional features. Use the following zero-mean rows before rescaling each to norm $\sqrt{3}$:

$$r_1 = (1, 2, -3), \quad r_2 = (1, -3, 2), \quad r_3 = (-3, 1, 2).$$

Their tangent vectors may be chosen as $(-5, 4, 1)$, $(5, -1, -4)$ and $(1, -5, 4)$, respectively. They are orthogonal to both the corresponding row and $\mathbf{1}$. Applying each row's sign matrix gives the columns

$$M = \begin{pmatrix} -5 & 5 & -1 \\ 4 & 1 & -5 \\ -1 & -4 & 4 \end{pmatrix}, \qquad \det M = 40.$$

The first two columns are independent and all three span $\mathbb{R}^3$. Positive rescaling of a row changes neither its tangent space nor its signs. Thus one, two and three rows give derivative ranks one, two and three, respectively; extra constant rows do not reduce these ranks.

Four or more features. For $D = 4$ use raw zero-mean rows

$$r_1 = (1, 2, -3/2, -3/2), \qquad r_2 = (1, -2, 3, -2).$$

For $D > 4$ set $k = D - 4$ and use

$$r_1 = (1, 2, -5/2, -3/2, \underbrace{1/k, \ldots, 1/k}_{k}), \quad r_2 = (1, -3, 3, -2, \underbrace{1/k, \ldots, 1/k}_{k}).$$

Both rows have zero mean and no zero coordinate. Normalize each to squared norm D, calling the resulting rows z_1, z_2. To prove $\mathcal{T}_{z_1} + \mathcal{T}_{z_2} = \mathbb{R}^D$, it suffices to show that their two normal planes have zero intersection. The first four coordinates of their four raw normal vectors form the matrix

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & -1 & 2+u \\ -1 & 3/2+u & 1 & 3 \\ -1 & 3/2 & -1 & 2 \end{pmatrix}, \quad u = \begin{cases} 0 & D = 4, \\ 1 & D > 4. \end{cases}$$

Its determinant is $u - 4$, hence nonzero. The four full normal vectors are independent, so the intersection of the two normal planes is zero. Rescaling the magnitude columns by the positive row-normalization factors preserves independence. It follows that two rows suffice for derivative rank D, and further constant rows leave that rank unchanged.

Attainment and uniform stability. For any of the full-dimensional constructions let

$$T = [S_{z_1}\Pi_{z_1} \ \cdots \ S_{z_h}\Pi_{z_h}] \in \mathbb{R}^{D \times hD}.$$

It has full row rank. Choose the stacked encoding matrix

$$B = \begin{pmatrix} B_1 \\ \vdots \\ B_h \end{pmatrix} = T^\top (TT^\top)^{-1}, \qquad E_i(q) = z_i + B_i q.$$

Then $D\Phi_0(0) = TB = I_D$. Each B_i maps into the tangent space of z_i. For h larger than necessary, one may equivalently set the added $B_i = 0$ and keep their positive-variance rows fixed. All rows remain nonconstant and all absolute-value coordinates remain nonzero on a sufficiently small ball. For a fixed number s of normalizations, the map and its derivative are continuous jointly in q and $\epsilon \geq 0$ there, since every denominator is bounded away from zero. Consequently there exist $\rho, \epsilon_0 > 0$ such that

$$\|D\Phi_\epsilon(q) - I_D\|_2 \leq \tfrac{1}{2} \quad (\|q\| \leq \rho, \ 0 \leq \epsilon \leq \epsilon_0).$$

Integrating the derivative along the line segment between q and q' gives $\|\Phi_\epsilon(q) - \Phi_\epsilon(q') - (q - q')\| \leq \|q - q'\|/2$, which proves equation E.2 and injectivity on the ball. Hence $r_\epsilon(q) = \Phi_\epsilon(q) - \Phi_\epsilon(0) - q$ is $1/2$-Lipschitz and zero at the origin. For $\|y - \Phi_\epsilon(0)\| < \rho/2$, the map $q \mapsto y - \Phi_\epsilon(0) - r_\epsilon(q)$ sends the closed ball of radius ρ into itself and is a contraction, so its fixed point is the unique solution of $\Phi_\epsilon(q) = y$ there. Since $D\Phi_\epsilon(q)$ is invertible on the ball, the inverse function theorem makes this decoder differentiable with derivative $(D\Phi_\epsilon(q))^{-1}$, and equation E.2 bounds its Lipschitz constant by two.

For one row, or for two rows when $D = 3$, choose a full-column-rank tangent derivative of dimension $r = \min\{D, h(D-2)\}$. A fixed left inverse of that derivative projects the output to $\mathbb{R}^r$ with identity derivative at zero; apply the same argument to this projected output. Restricting any of the constructed encodings to fewer input coordinates proves attainment for every $m \le r$. The affine encodings have finite, epsilon-independent derivatives, and their row variances remain, for example, at least $1/2$ after shrinking the ball, so full-dimensional uniform stability does not use the collapsing-variance exception of Theorem 4.9. □

Why sign diversity matters. With no coordinatewise gains and all coordinates of every row made positive by uniform carriers, the IHW is the sum of the shifted normalized rows. Its coordinate sum is constant, so its local dimension is at most $D-1$; adding identical copies of the one-row two-anchor code of Appendix E.2 therefore does not attain the multirow bound, and the construction above instead retains stable mixed signs in different rows before absolute value. The result is local: an arbitrary state and its simultaneous sign reversal have the same IHW after this normalization, so global state recovery is not implied.

Relation to the architecture. The affine row encodings are realized through the generator projection and the normalization-only residual subpath as in Appendix E.3.4. As for the two-anchor code (Appendix E.4), a sufficiently wide GELU head approximates this smooth inverse uniformly on a smaller closed latent ball (Leshno et al., 1993); the fixed-width trained head is not asserted to implement it.

E.2 Normalization width versus stable recoverability

Theorem 4.9 (One row: injective but ill-conditioned, full statement). *Let $m \ge 1$ and consider one normalized IHW row with no side channel.*

(i) For $\epsilon > 0$ and $D = m+1$, the encoding $E(q) = (q, 0)$, followed by s normalizations and a carrier $b\mathbf{1}$ with $b > \sqrt{D}$, is exactly inverted by equation E.3, whose denominator is positive at every encoded point. Width $m+1$ is necessary for differentiable local recovery with positive epsilon, at points where the encoding, post-normalization operations and decoder are differentiable (in particular, the coordinates entering absolute value are nonzero). At zero epsilon, with nonzero row variance, the corresponding minimum is $m+2$.

(ii) At width $D = m+1$ and a point q_0, let E_ϵ and H_ϵ be differentiable epsilon-dependent encoding and post-normalization maps (the latter may include locally smooth affine and IHW operations) with $\|DE_\epsilon(q_0)\|_2 \le K$ and $\|DH_\epsilon(z_0)\|_2 \le G$ at $z_0 = N_\epsilon^{\circ s}(E_\epsilon(q_0))$, where $K, G > 0$, and let the encoded pre-normalization variance be at least $v_0 > 0$. Then every differentiable local left inverse $\mathcal{D}_\epsilon$ of $H_\epsilon \circ N_\epsilon^{\circ s} \circ E_\epsilon$ satisfies equation 7. Thus with fixed s and epsilon-independent v_0, K, G, every such decoder has sensitivity $\Omega(\epsilon^{-s})$ as epsilon tends to zero.

Proof of part (i). For $D = m+1$ let $x = (q, 0)$, $y = P_D x$, $v = \|y\|^2/D$, and $a^2 = A_s v + \epsilon^s$. Equation E.1 gives $z = y/a$ and

$$1 - A_s\|z\|^2/D = \epsilon^s/a^2 > 0.$$

Since $y_j - y_D = x_j - x_D = q_j$, solving for a gives

$$q_j = \frac{\epsilon^{s/2}(z_j - z_D)}{\sqrt{1 - A_s(\epsilon)\|z\|^2/D}}, \qquad z = \delta - b\mathbf{1}. \tag{E.3}$$

The squared norm of z is at most D because $A_s \ge 1$, so each coordinate is at least $-\sqrt{D}$ and $b > \sqrt{D}$ ensures the absolute value does not change $b\mathbf{1} + z$. This also works at $q = 0$ for positive epsilon. Every normalization is invariant to adding $t\mathbf{1}$, so its Jacobian has rank at most $D-1$. The chain rule for a differentiable left inverse requires rank at least m, giving $D \ge m+1$. At zero epsilon, the rank argument in the upper-bound proof of Theorem 4.8 with $h = 1$ also removes the radial direction, and no later differentiable operation can increase the resulting rank $D-2$; thus $D \ge m+2$. The code $E(q) = (q, 0, 1)$ with the same carrier attains this width for every $\epsilon \ge 0$. By

equation E.1 its normalized row is $P_D E(q)/a(q)$ with $a(q)^2 = A_s \|P_D E(q)\|^2/D + \epsilon^s > 0$ (also at zero epsilon, because the last two coordinates of $E(q)$ differ, so $P_D E(q) \neq 0$), the carrier keeps the absolute value inactive, and subtracting coordinate $m+1$ cancels the shared mean and carrier, so

$$q_j = \frac{\delta_j - \delta_{m+1}}{\delta_{m+2} - \delta_{m+1}}, \qquad \delta_{m+2} - \delta_{m+1} = \frac{1}{a(q)} > 0.$$

□

This two-anchor code avoids the weak radial coordinate. On $\|q\|_\infty \leq R$ and $\epsilon \in [0,1]$, $A_s \leq s$ and $\epsilon^s \leq 1$ give $a(q)^2 \leq s(mR^2+1)/D+1$, so the anchor gap is at least $a_* = [s(mR^2+1)/D+1]^{-1/2}$ uniformly in epsilon. A feature error $\eta \leq a_*/4$ changes each difference by at most 2η, hence the decoded q by at most $2(1+R)\eta/(a_* - 2\eta) \leq 4(1+R)\eta/a_*$ in the max norm.

Proof of part (ii). At the encoded point write $y = P_D E_\epsilon(q_0)$, $v = \|y\|^2/D \geq v_0$, and $a^2 = A_s v + \epsilon^s$. Differentiating equation E.1 yields

$$J = \frac{1}{a} P_D - \frac{A_s}{D a^3} y y^\top.$$

Let $P_r = yy^\top/\|y\|^2$ and $P_t = P_D - P_r$. The shift direction has eigenvalue zero, the $D-2$ centered directions orthogonal to y have eigenvalue $1/a$, and the radial direction has eigenvalue ϵ^s/a^3. Equivalently,

$$J = a^{-1} P_t + (\epsilon^s/a^3) P_r.$$

Since $D = m+1$, the map $P_t D E_\epsilon(q_0)$ has rank at most $m-1$. There is a unit vector $u \in \mathbb{R}^m$ in its kernel. The derivative of the encoded output in this direction therefore satisfies

$$\|D(H_\epsilon \circ N_\epsilon^{\circ s} \circ E_\epsilon) u\|_2 \leq GK\epsilon^s/a^3.$$

Differentiating the local inverse identity and applying it to u gives $1 \leq \|D\mathcal{D}_\epsilon\|_2 GK\epsilon^s/a^3$. Substitute $a^2 \geq A_s v_0 + \epsilon^s$ to obtain equation 7. □

The noncollapse condition is essential. Without the variance floor v_0, width $m+1$ can be uniformly conditioned. Let $U \in \mathbb{R}^{(m+1)\times m}$ have orthonormal columns in $\mathbf{1}^\perp$ and take $E_\epsilon(q) = \epsilon^{s/2} U q$. Then

$$N_\epsilon^{\circ s}(E_\epsilon(q)) = \frac{Uq}{\sqrt{1 + A_s(\epsilon)\|q\|^2/D}}.$$

On a bounded q domain this map and its inverse have epsilon-uniform conditioning for $\epsilon \in (0,1]$, while the pre-normalization variance tends to zero.

Why. The Jacobian of s normalizations at the encoded point has eigenvalue $1/a$ on the $D-2$ centered tangential directions but only ϵ^s/a^3 on the radial one, with $a^2 = A_s v + \epsilon^s$. At width $m+1$ the tangential part has rank at most $m-1$, so some input direction reaches the output only radially and every left inverse must amplify it by the reciprocal. At the common default $\epsilon = 10^{-5}$ with $v_0 = G = K = 1$, the bound exceeds 10^5 for $s = 1$ and 10^{10} for $s = 2$. The information is locally recoverable for each positive epsilon, but under the stated variance and derivative bounds the minimum-width decoder cannot remain uniformly well-conditioned as epsilon tends to zero. The construction below uses complementary rows instead of relying on this weak radial direction.

E.3 A DIMENSION-UNIFORM TWO-ROW CODE WITH A FINITE-HEAD DECODER

We use the construction of Theorem 4.10 in Section 4.4. For $D = 4k$, $k \geq 1$, the rows z_1, z_2 repeat $\sqrt{2/5}\,(2,1,-1,-2)$ and $\sqrt{2/5}\,(1,-2,2,-1)$; each is centered, has squared norm D, and has no zero coordinate. The matrices S_i, P_i, T and $B = T^\top(TT^\top)^{-1}$ are as defined there; the inverse exists by Appendix E.3.1. Partition B into $D \times D$ blocks B_1, B_2, so $E_i(q) = z_i + B_i q$ and Φ_ϵ is the IHW of these two rows. For $s \geq 1$ and $0 \leq \epsilon \leq 1$ put $A_s = \sum_{j=0}^{s-1} \epsilon^j$, so that $a_0 = \sqrt{A_s + \epsilon^s}$, and $\delta_0 = |z_1| + |z_2|$.

Theorem 4.10 (Uniform two-row access with a finite head, full statement). *Uniformly over $D \in 4\mathbb{N}$, $s \geq 1$ and $\epsilon \in [0,1]$, the construction above has $\|B\|_2 = \sqrt{5}$. On the ball $\|q\|_2 \leq 1/10$, both row*

variances are at least one and all row signs stay fixed. The affine readout $\Psi_\epsilon = a_0\Phi_\epsilon - \delta_0$ satisfies equation 8 and the Jacobian bound equation E.5 below, is bi-Lipschitz with constants $1/2$ and $3/2$, and has a local differentiable inverse with Lipschitz constant at most two. For the raw IHW Φ_ϵ, the corresponding constant is at most $2\sqrt{s+1}$. Any scalar target $v^\top q + c$ has error at most $\|v\|_2$ times the right side of equation 8 under an affine head on IHW, implementable by two GELU hidden units and an output bias.

E.3.1 Uniform conditioning of the tangent construction

Let $s_i = \operatorname{sign} z_i/\sqrt{D}$ and $v_i = |z_i|/\sqrt{D}$. The four-coordinate patterns give

$$\|s_i\| = \|v_i\| = 1, \qquad s_1^\top s_2 = 0, \qquad s_i^\top v_j = 0 \ (i,j \in \{1,2\}), \qquad v_1^\top v_2 = 4/5.$$

These identities can be checked in one block: the unscaled magnitude dot product is 8, the squared norms are 10, and each signed cross sum is zero. Repeating a block leaves their normalized inner products unchanged. Since S_i is orthogonal and P_i is an orthogonal projection,

$$TT^\top = \sum_{i=1}^{2} S_i P_i S_i = 2I_D - s_1 s_1^\top - s_2 s_2^\top - v_1 v_1^\top - v_2 v_2^\top.$$

On $\operatorname{span}\{s_1, s_2\}$ this has eigenvalue one. On the orthogonal plane spanned by v_1, v_2, their projection sum has eigenvalues $1 + 4/5$ and $1 - 4/5$, so $TT^\top$ has eigenvalues $1/5$ and $9/5$. It has eigenvalue two on the remaining $D - 4$ dimensions. Thus it is positive definite, with minimum eigenvalue $1/5$, and

$$TB = I_D, \qquad B^\top B = (TT^\top)^{-1}, \qquad \|B\|_2 = \sqrt{5}.$$

Each $B_i = P_i S_i (TT^\top)^{-1}$ has its image in $\{\mathbf{1}, z_i\}^\perp$. Hence for $u_i = B_i q$,

$$\mathbf{1}^\top u_i = z_i^\top u_i = 0, \quad \sum_i \|u_i\|^2 \le 5\|q\|^2, \quad \sum_i S_i u_i = q.$$

In particular $D^{-1}\|E_i(q)\|^2 = 1 + \|u_i\|^2/D \ge 1$. The smallest $|z_{i,j}|$ is $\sqrt{2/5}$, while $\|u_i\|_\infty \le \sqrt{5}/10 < \sqrt{2/5}$. The variance and fixed-sign assertions follow on the entire ball, not only at its center.

E.3.2 Repeated normalization and the affine error

For a centered row y of nonzero variance, equation E.1 gives $N_\epsilon^{\circ s}(y) = y/\sqrt{A_s\|y\|^2/D + \epsilon^s}$, also at $\epsilon = 0$. Set $t = \|q\|_2$, $\chi = A_s/(A_s + \epsilon^s) \in (0, 1]$ and $f_i = (1 + \chi\|u_i\|^2/D)^{-1/2}$. Fixed row signs yield the exact identity

$$\Psi_\epsilon(q) = \sum_i S_i[(z_i + u_i)f_i - z_i]. \tag{E.4}$$

Using $0 \le 1 - (1+a)^{-1/2} \le a/2$ for $a \ge 0$, and $\sum_i S_i u_i = q$, we obtain

$$\begin{aligned}\|\Psi_\epsilon(q) - q\| &\le \sum_i \frac{\|u_i\|^2}{2D}(\sqrt{D} + \|u_i\|) \\ &\le \frac{5t^2}{2\sqrt{D}} + \frac{5\sqrt{5}t^3}{2D}.\end{aligned}$$

The last step uses $\sum_i \|u_i\|^3 \le (\sum_i \|u_i\|^2)^{3/2} \le 5\sqrt{5}t^3$. This proves equation 8.

E.3.3 Derivative bound and exact local invertibility

Differentiating equation E.4 gives

$$D\Psi_\epsilon(q) = \sum_i S_i \left[f_i B_i - \frac{\chi f_i^3}{D}(z_i + u_i)u_i^\top B_i \right].$$

The first part differs from $I_D = \sum_i S_i B_i$ by at most $\sum_i \|u_i\|^2 \|B_i\|/(2D) \le 5\sqrt{5}t^2/(2D)$. For the second part, let $K = \sqrt{5}$ and $U^2 = \sum_i \|u_i\|^2 \le K^2 t^2$. The bound $\sum_i \|B_i\|^2 \le 2K^2$, Cauchy–Schwarz, and $\max_i \|u_i\| \le Kt$ imply

$$\frac{1}{D}\sum_i (\sqrt{D} + \|u_i\|)\|u_i\|\|B_i\| \le \frac{\sqrt{2}K^2 t}{\sqrt{D}} + \frac{\sqrt{2}K^3 t^2}{D}.$$

Adding the two bounds gives

$$\|D\Psi_\epsilon(q) - I_D\|_2 \le \frac{5\sqrt{2}\,t}{\sqrt{D}} + \frac{5\sqrt{5}(\sqrt{2}+1/2)t^2}{D} < \frac{1}{2}, \tag{E.5}$$

since its largest value on $D \ge 4$, $t \le 1/10$ is $\sqrt{2}/4 + \sqrt{5}(\sqrt{2}+1/2)/80 < 0.408$. Integration along the segment from q to q' within the ball proves $\|\Psi_\epsilon(q) - \Psi_\epsilon(q') - (q-q')\| \le \|q-q'\|/2$, and hence the bi-Lipschitz bounds with constants $1/2$ and $3/2$, and injectivity. Since equation E.4 has $\Psi_\epsilon(0) = 0$, the contraction and inverse-function argument in the proof of Theorem 4.8, applied with radius $1/10$, gives a differentiable inverse on $\|y\| < 1/20$ whose Lipschitz constant is at most two by the lower bi-Lipschitz bound. Since $a_0 \le \sqrt{s+1}$ on $\epsilon \in [0,1]$, passing from Ψ_ϵ to Φ_ϵ gives the stated bound for the raw feature.

E.3.4 Finite-head access, actual interface, and limits

The affine prediction $\widehat{f}(\delta) = v^\top(a_0\delta - \delta_0) + c$ has error at most $\|v\|$ times the right side of equation 8. With exact GELU $\sigma(t) = t\Phi_{\mathrm{normal}}(t)$, symmetry of the Gaussian cdf gives $\sigma(t) - \sigma(-t) = t$. Two hidden units with opposite affine preactivations therefore implement any scalar affine map exactly. The vector affine decoder needs at most $2D$ such units.

The affine generator projection allows these row encodings at a declared latent interface; constant scale factors in its parameterization can be absorbed into its weights and biases. Zero attention and generator-FFN contributions, with the existing residual path retained, yield the analyzed repeated rowwise LayerNorms. A terminal head can absorb the affine calibration without adding an operation to the architecture. Additional generator slots can be held at fixed nonconstant rows; their constant IHW contribution is subtracted in the head bias. Restricting q to a lower-dimensional coordinate subspace gives the same bounds there.

The bounds assume this unit-gain, zero-bias normalization subpath; arbitrary learned affine gains, intervening nonlinearities, or another LayerNorm on the summed IHW define different maps. In particular, post-IHW zero-epsilon LayerNorm has derivative rank at most $D - 2$ and destroys the full-D conclusion.

Construction. Let $S_i = \mathrm{diag}(\mathrm{sign}\, z_i)$ and $P_i = I - \mathbf{1}\mathbf{1}^\top/D - z_i z_i^\top/D$. Set $T = [S_1 P_1 \; S_2 P_2]$ and stack $B = T^\top(TT^\top)^{-1}$. The smallest eigenvalue of $TT^\top$ is exactly $1/5$, for every repeated width. This gives an identity Jacobian after affine decoding at $q = 0$. Tangential perturbations change each row's normalization scale only quadratically; explicit remainder and derivative bounds give equation 8 and the uniform ball. Width uniformity concerns a fixed Euclidean ball, not a growing coordinatewise box.

E.4 Realizing the two-anchor readout in the attention block

Applying absolute values loses signs, and LayerNorm removes a common shift and changes scale; recovering information after absolute values has precedents in affine phase retrieval (Gao et al., 2018). The two-anchor encoding $E(q) = (q, 0, 1)$, which attains the zero-epsilon width of Theorem 4.9 with an explicit ratio inverse (Appendix E.2), survives both operations and is available in the implemented block. At a terminal block whose signed generator action realizes $q(c, v) = \sum_l \tanh(w_l^\top c + b_l) M_l v \in \mathbb{R}^m$, use one head and $D = m + 2$ features, reserve two zero output coordinates in each dynamic matrix action, set every mixing row to $(1, \ldots, 1)$ and inherited/query generators to zero, and let the bias of the affine generator output projection supply the anchors $(0, 1)$. Every generator slot then equals $E(q)$, so h identical slots give h times the one-row IHW and the same ratios recover q. With zero generator FFN, unit LayerNorm gains, zero first bias and second bias $b\mathbf{1}$ with $b > \sqrt{D}$, the two normalizations are the case $s = 2$ with a positive carrier.

Additional blocks with zero attention-generator outputs and generator FFNs add normalizations, and unit-gain, uniform-bias normalizations, including an optional LayerNorm on IHW, preserve the coordinate-difference ratios; a nonzero uniform gain also cancels in the ratio while the carrier stays positive, but arbitrary coordinatewise gains need not, and extra rescalings require recomputing the anchor-gap bound. The construction is a parameter choice, not a claim that training discovers it or that an arbitrary encoder supplies the signed action. A two-layer GELU head with sufficiently large hidden width approximates the continuous ratio inverse uniformly on a bounded domain (Leshno et al., 1993); exact recovery belongs to the ratio decoder, not to the fixed-width head used in the experiments.

F Additional Theoretical Results

The main text does not use the results of this appendix. They record the geometry of linear readouts of a zonotope, a perturbation bound for matrix banks, and the transformation degrees of freedom of common statistical operators (Definition B.1).

F.1 Linear-readout geometry and fixed-map perturbations

Theorem F.1 (Linear-readout geometry and fixed-map perturbations)**.** *Let $Z = \{c + G\xi : \|\xi\|_\infty \le 1\}$ with generator columns G_g, $w \in \mathbb{R}^d$, and $\mathrm{IHW}(Z) = \sum_g \|G_g\|_1$. For the scalar linear readout $y(\xi) = w^\top(c + G\xi)$:*

(a) $\mathrm{diam}(w^\top Z) = 2\sum_g |w^\top G_g| \le 2\|w\|_2 \mathrm{IHW}(Z)$.

(b) If the coefficients are independent and uniform on $[-1, 1]$, then $\mathrm{Var}_\xi[y] = \frac{1}{3}\sum_g (w^\top G_g)^2 \le \frac{1}{3}\|w\|_2^2 \mathrm{IHW}(Z)^2$.

(c) IHW alone does not order output diameters: with $w = e_1$, changing the sole generator from e_1 to $2e_2$ increases IHW from 1 to 2 but reduces the diameter from 2 to 0.

(d) Fix attention scores and a linear map A, standard or MZ with frozen context, and perturb the values by $G\xi$ with $\|\xi\|_\infty \le 1$. The exact perturbation set is $\{AG\xi : \|\xi\|_\infty \le 1\}$. Its support in direction u is $\sum_g |u^\top AG_g|$, and it is contained in the Euclidean ball of radius $r = \|A\|_2 \sum_g \|G_g\|_2$. If $r > 0$, its support is strictly smaller than the ball's in every nonzero direction orthogonal to $\mathrm{span}(AG)$.

(e) If input perturbations also change the scores, fix q and A, let $v_j = W_V x_j$, $\alpha = \mathrm{softmax}_j(q^\top W_K x_j/\sqrt{d_k})$, and $o = A\sum_j \alpha_j v_j$. If $\|\delta x_j\|_2 \le \rho$, then

$$\tilde{o} - o = A\sum_j \alpha_j W_V \delta x_j + \mathcal{E}, \quad \|\mathcal{E}\|_2 \le \frac{2\|A\|_2\|q\|_2\|W_K\|_2\rho}{\sqrt{d_k}} \left(\max_j \|v_j\|_2 + \|W_V\|_2\rho\right).$$

Changing queries or context gates contributes additional terms.

Proof. **(a)–(c).** The readout is $y(\xi) = w^\top c + \sum_g \xi_g w^\top G_g$. Maximizing and minimizing each independent coefficient gives the exact interval $w^\top c \pm \sum_g |w^\top G_g|$, and Cauchy–Schwarz gives $\sum_g |w^\top G_g| \le \|w\|_2 \sum_g \|G_g\|_2 \le \|w\|_2 \mathrm{IHW}(Z)$. Independent uniform coefficients have mean zero, variance $1/3$ and zero cross-covariances, which gives the identity in (b); the bound follows from $(w^\top G_g)^2 \le \|w\|_2^2 \|G_g\|_1^2$ and $\sum_g \|G_g\|_1^2 \le \mathrm{IHW}(Z)^2$. The randomness is the stipulated zonotope coefficients, not an assumed distribution of future prediction errors. Part (c) is direct computation. Monotonicity does hold under inclusion, since $Z_1 \subseteq Z_2$ implies $w^\top Z_1 \subseteq w^\top Z_2$, and scaling a fixed generator matrix by a nonnegative factor scales both its IHW and its linear-readout diameter; neither statement orders arbitrarily oriented learned generators.

(d). A perturbation $G\xi$ maps to $AG\xi$, and every such point with $\|\xi\|_\infty \le 1$ is attained. Each independent coefficient can take the sign of $u^\top AG_g$, which gives the support. Moreover $\|AG\xi\|_2 \le \sum_g \|AG_g\|_2 \le r$. If $u \ne 0$ is orthogonal to every AG_g, the support in direction u is zero, while the ball's is $r\|u\|_2 > 0$.

(e). Add and subtract $A\sum_j \alpha_j \tilde{v}_j$. The remainder is $A\sum_j(\tilde{\alpha}_j - \alpha_j)\tilde{v}_j$. Each logit changes by at most $\|q\|_2\|W_K\|_2\rho/\sqrt{d_k}$. The softmax Jacobian maps a vector z to $\alpha_j(z_j - \sum_i \alpha_i z_i)$, with ℓ_1 norm at most $2\|z\|_\infty$; integrating this bound and bounding $\max_j \|\tilde{v}_j\|_2$ gives the claim. □

These statements describe a specified set, linear readout, coefficient distribution and fixed map. Learned generator features, without a verified input enclosure and sound propagation through all subsequent operations, do not certify the full network. The experimental head uses the IHW vector as an MLP input, a predictive feature interface rather than a calibrated distribution over future labels.

F.2 Coefficient perturbations of a matrix bank

Proposition F.2 (Coefficient perturbations of a matrix bank)**.** *For coefficient vectors $\alpha, \beta \in \mathbb{R}^L$ let $M(\alpha) = M_0 + \sum_l \alpha_l M_l$ and $B_M = \|M_0\|_2 + \sum_l \|M_l\|_2$. Then $\|M(\alpha)\|_2 \le B_M$ whenever $|\alpha_l| \le 1$, and*

$$\|M(\alpha)v - M(\beta)v'\|_2 \le \|M(\alpha)\|_2\|v - v'\|_2 + \sum_l \|M_l\|_2|\alpha_l - \beta_l|\,\|v'\|_2.$$

(a) With context gates $|\gamma_l| \le 1$, $\|M(\gamma(c))v - M(\gamma(c'))v'\|_2 \le B_M\|v - v'\|_2 + \sum_l \|M_l\|_2|\gamma_l(c) - \gamma_l(c')|\,\|v'\|_2$. For native gates, $|\gamma_l(c) - \gamma_l(c')| \le \|w_l\|_2\|c - c'\|_2$.

(b) Let $f(S) = M(\sigma(\mu))\mu$ and $\hat{f}(S) = M(\gamma(\mu))\mu$. If $\|\mu\|_2 \le R$ and $\sup_\mu |\sigma_l(\mu) - \gamma_l(\mu)| \le \eta_l$, then $\sup_S \|f(S) - \hat{f}(S)\|_2 \le R\sum_l \eta_l\|M_l\|_2$. Exact native-core representation holds when the target gates themselves have the form $\tanh(w_l^\top \mu + b_l)$; arbitrary Lipschitz gates need not have that form.

Proof. The operator bound follows by the triangle inequality. For the difference, write $M(\alpha)(v - v') + \sum_l(\alpha_l - \beta_l)M_l v'$ and apply the same inequality. Part (a) takes $\alpha = \gamma(c)$ and $\beta = \gamma(c')$ and uses that $\tanh$ is 1-Lipschitz. Part (b) takes $v = v' = \mu$, $\alpha = \sigma(\mu)$ and $\beta = \gamma(\mu)$, and then the supremum; when each target gate is native, all η_l vanish. □

F.3 Transformation degrees of freedom of statistical operators

Proposition F.3 (Affine dimensions of statistical matrix families)**.** *Let point clouds have at least $d + 1$ points and allow all full-rank covariance configurations. For regularization $\lambda > 0$, write $\hat{\Sigma} = \mathrm{Cov}(S) + \lambda I$. The operator families $\hat{\Sigma}^{-1}$ and $\hat{\Sigma}^{-1/2}$ both have TDOF $d(d+1)/2$. The family of top-k orthogonal PCA projectors, restricted to covariances with an eigengap at k, has TDOF $d(d+1)/2 - 1$ for $0 < k < d$ (and zero for $k = 0, d$). Diagonal inverse-standard-deviation maps have TDOF d. These are matrix-family dimensions, not complexities of scalar Mahalanobis distances or other downstream predictions; covariance-dependent maps do not generally factor through the mean alone.*

Proof. **(a) Inverse covariance.** Symmetric matrices form a vector space of dimension $d(d+1)/2$. Every positive-definite covariance is realizable by an appropriately transformed full-rank point cloud with at least $d + 1$ points: start with a centered full-rank cloud, whiten its covariance, and apply the positive-definite square root of the desired covariance. Thus $\hat{\Sigma} = \mathrm{Cov}(S) + \lambda I$ ranges over the open set $\{\Sigma : \Sigma \succ \lambda I\}$. Inversion is a smooth invertible map of this set onto $\{A : 0 \prec A \prec \lambda^{-1} I\}$, an open subset of the symmetric matrices. Differences of points in any nonempty open subset span the whole ambient vector space, proving TDOF $d(d+1)/2$.

(b) Whitening. The inverse positive-definite square root maps the same covariance domain onto $\{A : 0 \prec A \prec \lambda^{-1/2} I\}$; its inverse is $A \mapsto A^{-2}$. This again gives an open symmetric-matrix image, so its affine dimension is $d(d+1)/2$.

(c) Top-k orthogonal PCA projection. For $0 < k < d$, every rank-k orthogonal projector P occurs as the leading eigenspace of a covariance $\alpha P + \beta(I - P)$ with $\alpha > \beta > \lambda$. All such projectors are symmetric and have trace k. Their differences therefore lie in the traceless symmetric space, whose dimension is $d(d+1)/2 - 1$. To prove equality, for any $i \ne j$ choose $k-1$ other coordinate directions,

possible because $k \leq d-1$. The difference of the two projectors obtained by adding e_i or e_j is $e_i e_i^\top - e_j e_j^\top$, spanning all traceless diagonal directions. Instead add $(e_i + e_j)/\sqrt{2}$ or $(e_i - e_j)/\sqrt{2}$. The difference is $e_i e_j^\top + e_j e_i^\top$, spanning each symmetric off-diagonal direction. Together these span the full traceless symmetric space, which proves the claimed affine dimension. For $k = 0$ or $k = d$ the projector is constant, so its TDOF is zero. The Grassmann manifold dimension $k(d-k)$ and the orthonormal-frame dimension $dk - k(k+1)/2$ count different objects; neither is the affine image dimension of these projectors.

(d) Per-coordinate standardization. The diagonal entries of an inverse-standard-deviation operator can vary independently over nonempty open intervals by choosing diagonal covariances with independent positive marginal variances. The image is contained in, and has nonempty relative interior in, the d-dimensional diagonal-matrix space. Its differences span that space. This concerns the linear scaling operator after centering; the mean-subtraction offset is a separate part of an affine z-score map. □

Proposition F.4 (Canonical operator for a fixed nonzero query)**.** *Fix $a \neq 0$, $\beta > 0$ and a point cloud $B = (b_1, \ldots, b_m)$, allowing repeated points. Define $b_*(B) = \sum_j \mathrm{softmax}_j(-\beta\|a - b_j\|^2) b_j$ and the explicitly chosen map $P_a(B) = b_*(B)a^\top/\|a\|_2^2$. Then $P_a(B)a = b_*(B)$ and $T_a(B) = I - P_a(B)$ has TDOF d as B ranges over $(\mathbb{R}^d)^m$.*

Proof. For fixed $a \neq 0$, $P_a(B)a = b_*(B)$, so $T_a(B)a = a - b_*(B)$ is the soft matching residual. Specifying only this action would leave the map undetermined on $a^\perp$, and no such representation exists at $a = 0$ when $b_*(B) \neq 0$; the canonical formula removes that ambiguity. Every difference $P_a(B) - P_a(B')$ belongs to $\mathcal{U}_a = \{ba^\top/\|a\|_2^2 : b \in \mathbb{R}^d\}$, which has dimension d because $a \neq 0$. Conversely, set every point in B equal to an arbitrary b. All scores are identical, the softmax weights are $1/m$, and $b_*(B) = b$. Taking $b = 0, e_1, \ldots, e_d$ produces d independent matrix differences, so $\mathrm{TDOF}(T_a) = \mathrm{TDOF}(P_a) = d$. This is not a lower bound for scalar Chamfer prediction: the soft nearest neighbor already uses input-dependent scalar routing, which a fixed value projection does not preclude. □

G Extended Related Work

Zonotopes as external verification tools vs. as internal representations. Zonotopes have a long history in formal verification and abstract interpretation of neural networks (Bonaert et al., 2021; Mirman et al., 2018; Jordan et al., 2022; Chung & Kousik, 2025), where they are computed post-hoc around a trained model's activations to certify robustness, and in control for data-driven reachability (Alanwar et al., 2021; 2023; Kochdumper & Althoff, 2023). MZAttn places matrix-basis gates inside a trainable center/generator representation and uses generator-derived features at readout. The matrix-basis mixture itself has close precedents in conditional parameterization; distinctiveness must be assessed at the level of the complete feature-processing architecture and its empirical effects.

Set-input networks. Deep Sets (Zaheer et al., 2017), PointNet (Qi et al., 2017a), PointNet++ (Qi et al., 2017b), and Set Transformer (Lee et al., 2019) establish several ways to process sets through shared nonlinear maps and pooling or attention. Their fixed learned matrices coexist with input-dependent computations.

Context-adaptive linear maps. Hypernetworks (Ha et al., 2017; Schug et al., 2025) and dynamic filters (Jia et al., 2016) generate input-dependent weights with costs that depend on their generator architecture; a universal $O(d_k^4)$ parameter cost is not assumed. A FiLM layer (Perez et al., 2018) is diagonal in its modulation coordinates, but surrounding learned maps can create off-diagonal effective transformations. Mixture-of-experts weights can vary continuously within a fixed routing support, so the complete mixture need not be piecewise constant (Shazeer et al., 2017).

CondConv (Yang et al., 2019) uses input-dependent sigmoid weights on a fixed expert basis. For $\gamma_l(\mu) = \tanh(w_l^\top \mu)$, the identity

$$M_0 + \sum_l \gamma_l(\mu) M_l = \left(M_0 - \sum_l M_l\right) + \sum_l \mathrm{sigmoid}(2w_l^\top \mu)(2M_l)$$

exhibits the relationship between the matrix families. The full learned generator branch and IHW readout differ from directly applying this mixture to a vector. Our HyperNet baseline is a low-rank conditional-weight model, not HYLA. The enclosing matrix zonotope is convex, while the image of the actual gate network need not fill that zonotope or be convex itself.

Equivariant architectures. EGNN (Satorras et al., 2021), SchNet (Schütt et al., 2017), and similar E(n)- or SO(n)-equivariant architectures hard-code rotation and translation symmetry into the forward pass. An invariant-output network cannot represent a fixed-frame target that changes under rotations; equivariant models with appropriately transforming outputs are a different case. Appendix I.5.4 compares with EGNN.

Classical tools in the theory. Theorem 4.6 is an exact error optimum over a gated function class under a specified input law, not a learning-dynamics or random-feature guarantee from spherical-harmonic analysis (Vempala & Wilmes, 2019) or a spectral decomposition of a fixed bilinear interaction (Pearce et al., 2025, Sections 2–3.2). The pseudoinverse encoding of Theorem 4.10 has precedents in frame-based reconstruction for ReLU layers (Haider et al., 2023, Section 4.4); what is specific is a uniformly conditioned signed tangent map and control of the normalization remainder at the summed-IHW interface.

H Experimental Setup, Hyperparameters, and Compute

H.1 Implementation details

Hyperparameter summary. Default configuration shared across all attention-based architectures except where overridden per task or by the parameter-matched variants:

Setting	Value
Optimizer	AdamW (Loshchilov & Hutter, 2019)
Learning rate	10^{-4}
LR schedule	cosine (per-batch steps)
LR warmup	5 epochs (linear $0 \to 10^{-4}$)
Weight decay	10^{-5}
Gradient clip	1.0 (ℓ_2 norm)
Batch size	128 (synthetic) / 256 (geometric tasks)
Epochs (cap)	200
Early stopping	validation-loss patience 20
Dropout	0.1 (constant)
d_{model}	64 (default) / 80 (ST-Large)
Heads H	4
d_{ff}	128 (default) / 200 (ST-Large)
ISAB inducing points	16
MZAttn n_g	8 (default) / 2 (Slim) / 16 (Large)
MZAttn L	4 (default) / 1 (Slim) / 8 (Large)

MZ parameter initialization. Center matrices $M_c^{(h)}$ are initialized as I_{d_k}. Generator matrices $M_l^{(h)}$ are initialized from $\mathcal{N}(0, 0.02^2)$. Mixing weights W_{mix} are initialized from $\mathcal{N}(0, 0.1^2)$. All other linear layers use Xavier uniform initialization.

Generator FFN. The generator path uses a lighter FFN with hidden dimension $d_{\text{ff}}/2$ (compared to d_{ff} for centers).

Computational cost comparison. For the default configuration ($d = 64$, $H = 4$, $n_g = 8$, $L = 4$), the MZ-Set Transformer has approximately 372K parameters compared to 208K for the standard Set Transformer. The additional parameters come primarily from the generator projection layer (W_g: $d_{\text{in}} \times n_g d$), generator FFNs, and the MZ parameters ($H \times (1 + L) \times d_k^2 = 4 \times 5 \times 256 = 5{,}120$

matrix-zonotope parameters per attention layer, plus a small $O(LHd_k)$ for the gating network and the mixing matrix W_{mix}).

H.2 Per-task data and training settings

Task A: Set regression. Given a set $S = \{x_1, \ldots, x_n\}$ with $x_i \in \mathbb{R}^{16}$ and integer n sampled uniformly from 10 through 100, predict $y = \sum_{j=1}^{16} \max_{1 \le i \le n} x_{ij}$. Thus each feature is maximized across the set before summation; it is not the sum of within-token feature maxima. 10K train, 2K val, 2K test; 200 epochs; 5 seeds (3 seeds for parameter-matched models).

Task B: Set matching. Given two point clouds $A, B \subset \mathbb{R}^3$ (each 10–30 points), predict the Chamfer distance $\text{CD}(A, B)$. Set B is either a rotated-noisy version of A (50%) or an independent random cloud (50%). Both sets are packed into a single input with a binary type indicator feature, giving $d_{\text{in}} = 4$. 5K train, 1K val, 1K test; 200 epochs; 3 seeds.

Task C: Point cloud classification (ModelNet40). We classify 3D shapes from ModelNet40 (Wu et al., 2015), which comprises 40 categories with 9,843 training and 2,468 test instances. Each shape is represented by 1,024 randomly sampled surface points in $\mathbb{R}^3$. Training uses random y-axis rotation and jittering for augmentation; 200 epochs; 5 seeds.

Task D: Few-shot set retrieval. Given a support set S of $K{=}5$ examples from a target class and a query set Q of 20 elements from mixed classes, predict the fraction of Q belonging to the target class. Each element has 8 feature dimensions plus a binary type indicator, yielding $d_{\text{in}} = 9$. Class separability varies per sample (noise scale $\in [0.3, 1.5]$), creating natural relational ambiguity. 8K train, 1.5K val, 1.5K test; 200 epochs; 3 seeds.

Task E: ScanObjectNN classification (real-world). We evaluate on ScanObjectNN (Uy et al., 2019), a realistic 3D point cloud classification benchmark constructed from real indoor scans. We use the OBJ_BG variant (15 classes, $\sim$2,300 train, $\sim$600 test). Each shape is represented by 1,024 randomly sampled points in $\mathbb{R}^3$ with y-axis rotation and jitter augmentation. 200 epochs; 5 seeds.

Task F: Rotation-dependent set matching. Given two point clouds $A, B \subset \mathbb{R}^8$ where $B = R(\theta)A + \varepsilon$ for a sample-dependent rotation $R(\theta)$ (composed from Givens rotations in all $\binom{8}{2} = 28$ planes) and i.i.d. noise $\varepsilon \sim \mathcal{N}(0, 0.1^2 I)$, predict the post-alignment Chamfer distance. The task involves alignment-dependent geometry, motivating context-dependent transformations as an inductive bias. It does not require a learned predictor to explicitly reconstruct a rotation. Both sets are packed with a binary type indicator, giving $d_{\text{in}} = 9$. 5K train, 1K val, 2K test; 200 epochs; 3 seeds.

Task G: Minimum spanning tree weight in $\mathbb{R}^8$. Given a point set $S = \{x_1, \ldots, x_n\} \subset \mathbb{R}^8$ with $n \sim \text{Uniform}(10, 30)$, predict the total weight (sum of edge lengths) of the minimum spanning tree of S. This is a classical computational geometry quantity that depends on the global pairwise distance structure of the set, a natural relational task where the output is determined by the collective geometric arrangement rather than any single element. Points are generated with 1–3 random clusters (varying centers and spreads) to create diverse geometric configurations. No explicit context signal is broadcast; the model must infer the relational structure from pairwise distances in high-dimensional space. 5K train, 1K val, 2K test; 200 epochs; 3 seeds.

H.3 Compute resources

All experiments were run on a mix of consumer and datacenter GPUs. Synthetic benchmarks (Tasks A, B, F, G, hull, MST, rotation matching at $d \le 64$, sparse-geometric set predicates, generator-budget, HyperNet-rank and compact-configuration ablations, set-size generalization, ambiguity sweep, quadratic interaction targets, depth sweep) used a single NVIDIA RTX 5090 (32 GB). Higher-dimensional rotation matching ($d \in \{128, 256, 512\}$) and MST scaling ($d \in \{64, 128, 256\}$) used a single NVIDIA H200 (141 GB). Synthetic runs took 0.5–30 minutes per seed, depending on task and dimension. Rerunning the full sweep, including studies not reported here, was estimated at approximately 1,200 GPU-hours; this is not a measured total for the reported experiments.

I Additional Experiments

I.1 Complete main comparison

Table S1 expands the complete-model comparison in Table 2.

Table 2 compares complete models on seven tasks. MZAttn has the lowest mean MSE on regression, rotation matching and MST, 13.0%, 43.0% and 41.1% below ST-Large, at similar budgets across tasks (ST-Large 361–364K, HyperNet 359–362K, Perceiver 340–343K; MZAttn 365–372K). ST + FiLM is better on Chamfer with fewer parameters (48.0 versus 58.4 MSE, 224K) and on few-shot retrieval (3.06 versus 4.71, $\times 10^{-2}$), and ST-Large leads both classification tasks. Table S1 provides expanded metrics and representative parameter counts.

Table S1: Main comparison in three panels (Tasks A–G; mean ± population standard deviation; dashes: not reported; Deep Sets was evaluated only on ScanObjectNN). Bold marks the best mean for each metric. Parameter counts are representative and vary with task; the MZ count is 371,929 on Task A and 365,017 on Task B.

	A: Set regression		**B: Set matching**	
Model	MSE ↓	R^2 ↑	MSE ↓	R^2 ↑
Set Transformer	3.43±0.09	0.811±0.005	138.0±10.3	0.901±0.007
ST + FiLM	3.77±0.26	0.792±0.014	**48.0**±7.9	**0.966**±0.006
ST-Large	3.36±0.10	0.815±0.005	106.5±5.3	0.924±0.004
Perceiver	3.47±0.05	0.809±0.003	136.2±17.5	0.902±0.013
HyperNet	3.51±0.01	0.806±0.001	76.1±5.7	0.945±0.004
MZAttn (ours)	**2.92**±0.09	**0.839**±0.005	58.4±11.9	0.958±0.009

	F: Rotation matching		**G: MST weight**	
Model	MSE ↓	R^2 ↑	MSE ↓	R^2 ↑
Set Transformer	487.5±31.1	0.904±0.006	209.8±1.8	0.712±0.002
ST + FiLM	490.5±17.8	0.904±0.003	232.7±12.5	0.681±0.017
ST-Large	423.2±51.8	0.917±0.010	168.7±23.0	0.769±0.032
Perceiver	872.1±163.8	0.829±0.032	249.0±4.7	0.659±0.006
HyperNet	1373.0±141.6	0.730±0.028	291.2±9.2	0.601±0.013
MZAttn (ours)	**241.2**±3.0	**0.953**±0.001	**99.4**±1.9	**0.864**±0.003

	C: ModelNet40	**E: ScanObjectNN**	**D: Few-shot retrieval**		
Model	Acc. ↑	Acc. ↑	MSE ($\times 10^{-2}$) ↓	R^2 ↑	Params
Deep Sets	–	40.6±1.0%	–	–	36K
Set Transformer	77.5±0.6%	71.6±2.2%	4.81±1.85	0.363±0.246	208K
ST + FiLM	78.4±0.8%	68.9±3.2%	**3.06**±0.65	**0.594**±0.086	225K
ST-Large	**79.9**±1.0%	**72.9**±2.4%	4.28±0.92	0.432±0.122	362K
Perceiver	75.0±0.3%	67.3±1.2%	12.19±5.81	−0.616±0.771	341K
HyperNet	75.5±1.7%	66.1±0.9%	11.96±6.96	−0.585±0.923	360K
MZAttn (ours)	78.1±0.7%	69.4±3.7%	4.71±2.77	0.375±0.367	365K

I.2 Architecture and ablation studies

Interaction and bank diagnostics. The diagnostic target is $f(S) = w^\top[M_0 + \sum_{l=1}^{K} \mu_l(S) M_l]\mu(S)$ for a fixed random draw of w and M_l in dimension 32, where $\mu(S)$ is the set mean; K counts construction terms, not the scalar complexity of Section 4.2. In Figure 2 (left), one MZAttn layer keeps R^2 between 0.995 and 0.908 as K grows from 4 to 32, while Set Trans-

formers with one, two or four layers fall to 0.396–0.407. For $K \geq 8$ its lead over the best of the three depths grows from 0.17 to 0.50, and every MZAttn seed exceeds every Set Transformer seed. Additional depth does not close the gap in these evaluated settings. Both trained models include FFNs and LayerNorm, and MZAttn reads IHW; this is a complete-model comparison, not a direct empirical test of the restricted comparators in Section 4.1.

The middle panel of Figure 2 varies the bank size. A single matrix direction already reaches 0.978 at $K = 16$, and $L = 2$ or 4 gives the best means at $K = 32$ (0.960 and 0.961); $L = 32$ gives the lowest mean for both targets, although parameters grow from 247K to 383K. The $L = K$ bank of the left panel is therefore not a favorable choice for MZAttn. The number of matrix directions needed is thus far smaller than the number of target terms. A compact model with $h = 2$ and $L = 1$ stays within 0.022 of larger MZ configurations on three further tasks (Table S2). A variant that applies LayerNorm to the generator readout features learns the $d = 16$, $K = 8$ target with either an IHW or a signed readout, reaching 0.966 ± 0.008 and 0.970 ± 0.006 with 372K parameters in both.

I.2.1 Compact and full configurations

Table S2: Compact and full MZ configurations (R^2, reported means over three seeds). Max uses the Task A data with a 100-epoch cap instead of 200, so its values differ slightly from Table S1. Standard has fewer parameters, so this comparison does not isolate architecture from parameter budget.

Model	Max $R^2 \uparrow$	RotMatch $R^2 \uparrow$	MST $R^2 \uparrow$	Params
Standard	0.812	0.706	0.719	208K
MZ-Slim (h=2, L=1)	**0.871**	0.919	0.835	333K
MZ-Med (h=4, L=2)	0.863	0.933	**0.857**	344K
MZ-Full (h=8, L=4)	0.850	**0.936**	0.852	368K

Larger MZ budgets give no consistent gain in Table S2: across MZ variants the largest within-task spreads are 0.021, 0.017 and 0.022, and without seed uncertainties the table does not establish equivalence.

Matrix-bank and stored-slot budgets. Matrix-bank size L and stored-slot count $h = n_g$ control different parts of Equation 3. Table S2 changes both, the middle panel of Figure 2 varies L, and a single-seed set-matching sweep (seed 42) varies h: for $n_g = 2$, 4, 8 and 16 the test MSE is 45.80, 26.74, 32.46 and 55.05 (MAE 3.50, 3.33, 3.39 and 4.47; R^2 0.967, 0.981, 0.977 and 0.961). Additional seeds and train/validation diagnostics would be needed to attribute the degradation beyond $n_g = 4$ to overfitting. The available results favor considering compact configurations and show no monotonic performance benefit from increasing either budget. They do not supply a task-independent allocation rule or identify target TDOF from the best tested configuration.

I.2.2 HyperNet rank and network depth

HyperNet rank. Sweeping the HyperNet update rank $r \in \{16, 32, 64\}$ on raw-target hull ($d = 3, 6, 9$) and MST ($d = 64, 128, 256$) over three seeds leaves every HyperNet mean below the MZ-Full means of Tables S5 and S7, although the swept HyperNets have 0.49–0.51M, 0.76–0.78M and 1.31–1.32M parameters at $r = 16$, 32 and 64, more than the 360K HyperNet of those tables. On hull at $d = 3$ the HyperNet means range from 0.190 to 0.204, against 0.804 for MZ-Full, and the rank variation on MST at $d = 128$ is 0.29. At $d = 6$ and 9 the hull means lie in $[-0.214, -0.203]$ and $[-0.100, -0.095]$, and the MST means lie in $[-2.06, -1.92]$, $[-3.09, -2.80]$ and $[-3.99, -3.78]$ at $d = 64$, 128 and 256. Like the other raw-target results, this sweep does not establish a capacity limit (Figure 2, right, and Figure 5, left).

Depth. Table S3 compares one MZAttn layer with Set Transformers of one to six layers on set matching with noise $\sigma = 0.5$.

Table S3: Empirical depth sweep (mean±std, 3 seeds). One-layer MZAttn has lower MSE than the evaluated standard models at depths 1–6. The six-layer baseline has approximately twice its parameter count; the sweep does not establish a representation lower bound.

Model	Layers	MSE ↓	R^2 ↑	Params
MZAttn	**1**	**42.3**±1.6	**0.969**±0.001	244K
Set Transformer	1	53.3±0.7	0.961±0.001	140K
Set Transformer	2	57.3±4.6	0.958±0.003	208K
Set Transformer	3	68.6±16.4	0.950±0.012	276K
Set Transformer	4	59.6±3.7	0.956±0.003	344K
Set Transformer	5	63.4±8.3	0.953±0.006	412K
Set Transformer	6	57.1±5.9	0.958±0.004	480K

I.2.3 Parameterization diagnostics

The following diagnostics illustrate the parameterization of an MZ attention layer. In these probes the loss reads only the center stream, so the context gates and generator matrices receive no gradient and stay at their random initialization, and the probe target redraws its coefficients for every sample. The plots (Figures 3 and 4) therefore show how the gated matrix family varies with context by construction, not structure learned from a task.

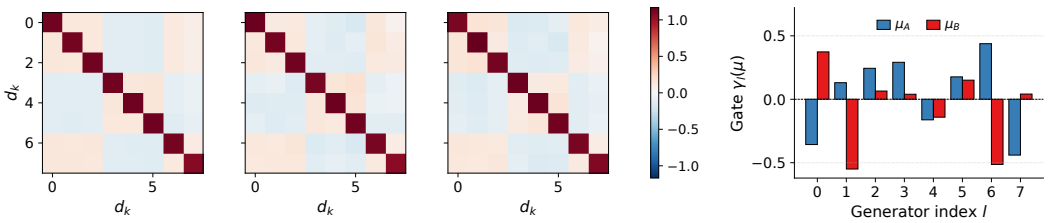

Figure 3: **The gated matrix family of an MZ attention layer (head 0).** Left: the center matrix M_c. Second and third from left: $M(\mu) = M_c + \sum_l \gamma_l(\mu) M_l$ for two synthetic contexts μ_A and μ_B. Right: their gates $\gamma_l(\mu)$. The two operators differ by 8.0% of $\|M_c\|_F$, and four of the eight gates change sign. The gates and generators are untrained in this probe, so the panels illustrate the parameterization rather than a learned adaptation.

Limits of bank-level diagnostics. The raw matrix family is not a parameterization-invariant description of the full predictor. For any positive constants s_l, replacing

$$M_l \leftarrow s_l M_l, \qquad W_{\mathrm{mix},gl} \leftarrow W_{\mathrm{mix},gl}/s_l$$

leaves every $B_g(\mu)$ and the output of Equation 3 unchanged, with gates and other parameters fixed. If mixing columns are shared across heads, the same rescaling is applied to the corresponding generator matrices in every such head. In the default data flow this leaves deterministic predictions unchanged, although optimization and regularization need not be equivalent. Raw generator norms and variation of $M_c + \sum_l \gamma_l M_l$ can change under this rescaling. The following plots therefore describe the reported parameterization; they do not identify a necessary predictive mechanism. This observation does not alter the definition of TDOF for a specified operator.

I.3 Robustness and generalization

Table S4 varies the input noise σ and the test set size on set matching.

I.4 Computational geometry and scaling

With training-split target normalization (Figure 2, right, and Figure 5, left), MZ-Full has the highest mean MST R^2 at every dimension, but its lead over the best baseline shrinks from 0.078 at $d = 16$

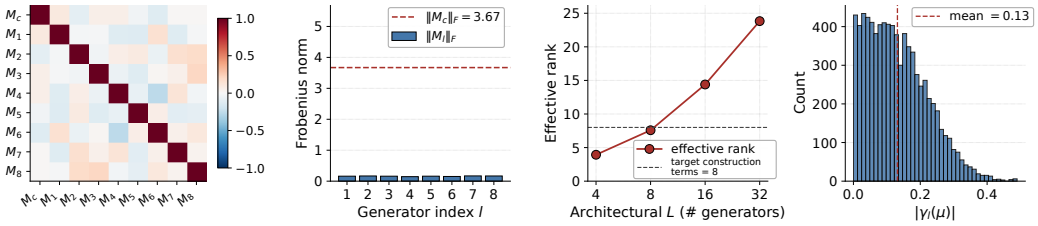

Figure 4: **Matrix-bank diagnostics of an MZ attention layer ($L = 8$, head 0, 1,000 probe contexts).** From left: cosine similarities between the flattened matrices $M_c, M_1, \ldots, M_8$; generator Frobenius norms, with the center norm dashed; effective rank against the architectural L in a separate sweep $L \in \{4, 8, 16, 32\}$ with 200 extracted operators, with a reference line at the 8 terms used to construct the probe target; and the distribution of gate magnitudes $|\gamma_l(\mu)|$. As in Figure 3, the generators and gates remain at their random initialization and only the center path is trained ($\|M_c\|_F = 3.67$). Generator norms lie in $[0.146, 0.169]$ (coefficient of variation 5.1%), about $1/23$ of the center norm. With random generators the effective rank grows with the architectural L, not with a rank that the target requires. Of the probed gate magnitudes, 78.6% exceed 0.05 (mean 0.13, 95th percentile 0.29).

Table S4: Noise and set-size sensitivity on set matching (one training run per setting). Top: MSE as a function of noise σ. Bottom: the $\sigma = 0.3$ models, trained on sets of size 10–30, tested on larger sets; their in-distribution result is the $\sigma = 0.3$ row.

σ	ST MSE	ST R^2	MZ MSE	MZ R^2	ΔMSE
0.05	62.06	0.957	**41.12**	**0.971**	-33.7%
0.1	62.58	0.956	**44.69**	**0.969**	-28.6%
0.3	59.87	0.957	**38.92**	**0.972**	-35.0%
0.5	57.25	0.958	**39.21**	**0.971**	-31.5%
1.0	58.28	0.955	**44.80**	**0.965**	-23.1%
Test size	ST MSE	ST R^2	MZ MSE	MZ R^2	ΔMSE
$[30, 50]$	24.87	0.978	**19.85**	**0.982**	-20.2%
$[50, 80]$	33.86	0.968	**24.07**	**0.977**	-28.9%
$[80, 120]$	34.56	0.969	**30.40**	**0.972**	-12.0%

to 0.043 at $d = 128$ and 0.007 at $d = 256$. Convex-hull volume is mixed: MZ-Full ties ST-Large at $d = 3$ (-0.003), leads the best baseline by 0.143 at $d = 6$, where the baselines vary strongly across seeds, and trails LA-ST at $d = 9$ (-0.049), where every model stays below 0.2. Without normalization, all baselines fall to negative MST R^2 at $d \geq 64$ while MZ-Full stays positive (Table S7); target scale therefore changes the size of the gaps. Sparse geometric predicates show

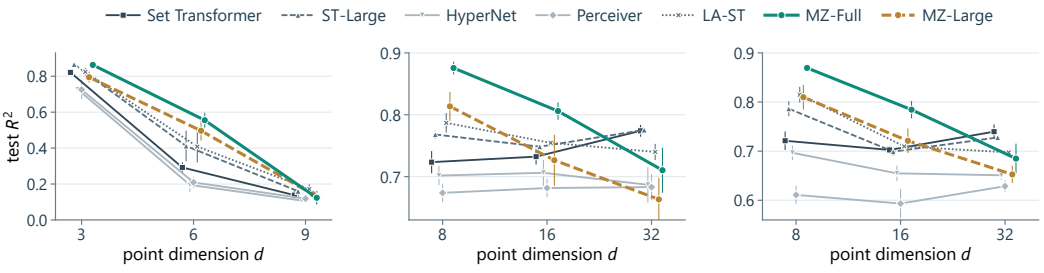

Figure 5: **Additional geometric scaling. Left:** convex-hull volume with training-split target normalization. **Middle and right:** raw-target diameter and k-nearest-neighbor radius. Lines show means and bars show sample standard deviations over five seeds; points are offset horizontally.

the same trend without normalization: MZ-Full's lead on diameter and neighbor radius ($+0.051$ to $+0.089$ at $d \leq 16$) reverses at $d = 32$ (-0.065 and -0.055; Figure 5, middle and right).

I.4.1 CONVEX HULL VOLUME

The convex-hull task predicts the volume of the hull of a point set in $\mathbb{R}^d$, computed numerically with Qhull (Barber et al., 1996). A hull has at most n vertices; combinatorial face complexity should not be confused with a larger number of input support points. Away from degeneracies and changes of combinatorial structure, volume varies smoothly with vertex positions. The experiment evaluates geometric regression, not exact-arithmetic certification.

Table S5: Raw-target convex hull volume (R^2, mean±std over 5 seeds; $n \in [15, 35]$). MZ-Large leads the listed means, with a 0.322 gap over ST-Large at $d = 3$. At $d = 9$, MZ-Full is near zero (0.003 ± 0.006). Normalization substantially changes the comparison (Figure 5, left); these raw-target results do not establish an expressiveness separation.

Model	Params	d=3 R^2	d=6 R^2	d=9 R^2
Standard	208K	0.273±0.081	−0.209±0.016	−0.099±0.002
ST-Large	362K	0.504±0.058	−0.178±0.012	−0.088±0.004
HyperNet	360K	0.136±0.082	−0.211±0.013	−0.097±0.003
Perceiver	341K	0.164±0.058	−0.221±0.007	−0.096±0.003
LA-ST (Linear Attn.)	359K	0.473±0.036	−0.126±0.022	−0.077±0.005
MZAttn-Full	368K	0.804±0.014	0.103±0.020	0.003±0.006
MZAttn-Large	457K	**0.826**±0.016	**0.191**±0.053	**0.056**±0.003

I.4.2 ROTATION MATCHING AT HIGH DIMENSION

We extend Task F to $d \in \{16, 32, 64, 128, 256, 512\}$ (Table S6). A rotation group has $d(d-1)/2$ local degrees of freedom, but that manifold dimension is not the affine matrix-span dimension in Definition B.1, and this sweep does not identify the cause of the poor high-dimensional fits. The Standard model's seed standard deviations of 0.47 at $d = 64$ and 3.5 at $d = 128$ show training variability that does not distinguish representational limits from optimization failure.

I.4.3 MST WEIGHT AT HIGH DIMENSION

We extend Task G to $d \in \{16, 32, 64, 128, 256\}$ (Table S7); a tree selector and its scalar weight are not assigned $O(d^2)$ operator TDOF by this setup. On raw targets MZ-Large leads MZ-Full at $d = 128, 256$, but with targets standardized by the training-split mean and standard deviation Full has the higher mean at every dimension, so the crossover does not show that a larger generator budget is required by increasing target complexity.

I.4.4 SPARSE GEOMETRIC SET PREDICATES (SGSP)

Motivated by the computational-geometry study, we use the task family *Sparse Geometric Set Predicates* (SGSP): targets computed from a continuous d-dimensional point cloud, without symbolic or one-hot encoding, that are locally smooth functions of a combinatorially selected subset of points or distances. Examples include extreme pairs, nearest neighbors, boundary points and small distance-subset cuts. At changes of the selected subset, the target can be nonsmooth; the number of selected elements can also grow with dimension or set size. The associated per-sample selector motivates context adaptation as an architectural hypothesis. It is neither a unique operator identified by the scalar target nor an impossibility result for fixed-value attention.

The middle and right panels of Figure 5 report diameter and neighbor radius. These targets use pairwise distances or neighbor sorting and are invariant to common translations and orthogonal rotations, so symmetry-aware models can exploit relevant structure; this sweep does not include that control. On a third predicate, the sum of the k smallest pairwise distances, every model stays below $R^2 = 0.16$; MZ-Full reaches 0.152, 0.010 and 0.014 at $d = 8$, 16 and 32. Distance concentration at $d = 32$ is a possible untested explanation for the reversal of MZ-Full's diameter and neighbor-radius lead at this dimension, not an established prediction ceiling.

Table S6: Rotation matching scaling (R^2, mean over 3 seeds). Several evaluated baselines develop large negative R^2 as d grows; MZAttn stays near the mean predictor at $d = 128, 256$, which is not accurate task prediction. At $d = 512$ all methods have strongly negative R^2; MZAttn has lower MSE than the listed alternatives under this protocol.

Model	d=16 R^2	d=32 R^2	d=64 R^2	d=128 R^2	d=256 R^2	d=512 R^2
Standard	0.084	−0.002	−1.20	−21.54	−61.69	−165.4
ST-Large	0.285	−0.001	−0.023	−9.47	−50.15	−145.1
Perceiver	−0.001	−0.006	−1.17	−19.18	−61.52	−165.2
HyperNet	0.020	−0.005	−1.10	−16.51	−61.48	−163.0
MZAttn	**0.826**	**0.519**	**0.071**	**−0.001**	**−0.005**	**−61.6**

Table S7: Raw-target MST scaling (R^2, mean±standard deviation over five seeds). The listed non-MZ baselines have negative means at $d \geq 64$, whereas MZ remains positive. The normalized comparisons in Figure 2 (right) give much smaller differences; these results do not establish a representation lower bound.

Model	d=16	d=32	d=64	d=128	d=256
Standard	0.295±0.072	−0.202±0.367	−1.326±0.245	−2.789±0.511	−3.221±0.305
ST-Large	0.526±0.044	0.274±0.140	−0.252±0.046	−1.230±0.430	−2.425±0.247
HyperNet	0.250±0.068	−0.211±0.244	−1.352±0.116	−1.985±0.251	−3.423±0.222
Perceiver	0.400±0.097	−0.082±0.180	−1.049±0.242	−2.361±0.248	−3.320±0.299
MZAttn-Full	**0.724**±0.010	**0.625**±0.027	**0.581**±0.008	0.484±0.019	0.384±0.042
MZAttn-Large	0.612±0.009	0.531±0.018	0.538±0.038	**0.497**±0.021	**0.435**±0.012

I.4.5 Large-scale set tasks

We evaluate three synthetic tasks (max, sum, range) from $d = 16$ with $n \in [10, 50]$, smaller sets than the $n \in [10, 100]$ of Task A, up to $d \in \{64, 128\}$ with n up to 500–1000. Each configuration is trained for 150 epochs with 3 seeds.

Table S8: Large-scale results (R^2, mean over three seeds). Max differences diminish at larger scales; all listed models have low R^2 on range.

Task	Configuration	Standard	ST-Large	Perceiver	MZAttn
Max	d=16, n=10–50	0.748	0.822	0.728	**0.843**
Max	d=64, n=200–500	0.744	0.744	**0.745**	**0.745**
Max	d=128, n=500–1000	**0.632**	0.631	0.628	0.627
Sum	d=64, n=200–500	0.939	0.946	0.910	**0.998**
Range	d=64, n=200–500	**0.083**	**0.083**	0.077	0.082
Range	d=128, n=500–1000	0.038	0.038	0.036	**0.039**

In Table S8, summation is linear (for fixed cardinality, $\sum_i x_i = n\mu$), so MZ's lead on sum is not evidence that an input-dependent matrix family is necessary. The table alone does not identify why the max scores converge at larger scales, and the range scores below 0.1 at $n \geq 200$ do not imply an irreducible prediction ceiling.

I.5 Transfer and symmetry-aware comparisons

Scope across task families. On support-conditioned operator regression, ST-Large reaches MSE 0.17 with 255K parameters against 0.47 for MZ-Full with 334K (Table S9). Default MZAttn leads no QM9 property (Ramakrishnan et al., 2014) (Table S10), LA-ST leads all three ResNet-18 (He et al., 2016) feature-geometry targets (Table S15), MZ-Full is last on rotation matching with real ModelNet40 shapes ($R^2 = -0.066$ against 0.690 for Perceiver; Appendix I.5.6), and image completion and held-out tabular meta-regression show no advantage (Appendix I.5.6). Relative to a Set

Transformer of the same width, MZAttn costs 1.7–1.9× the parameters, 2.4–2.7× the inference time and 2.2–5.5× the training-step time (Table S14).

I.5.1 Few-shot meta-regression

Unlike the fixed quadratic targets of Section 5.3, this support-conditioned linear-operator prediction task requires the operator to be inferred from a support set at inference time. The seven models have different budgets, so the comparison is not parameter-matched.

Setup. We fix a shared basis $\{M_1, \ldots, M_L\}$ of random matrices $M_l \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$. For each task τ, we draw coefficients $c_\tau \sim \mathcal{N}(0, I_L)$ and define the task-specific operator $A_\tau = \sum_{l=1}^{L} c_\tau[l]\, M_l$. The model observes K=30 support pairs (x_i, y_i) with $y_i = A_\tau x_i$ and must predict $y_q = A_\tau x_q$ for a new query x_q. A_τ is never given; it must be inferred from the support context. We set $d_{\text{in}} = d_{\text{out}} = 8$ and vary $L \in \{4, 8, 16, 32\}$ to control the intrinsic dimension of the operator family. All seven architectures are instantiated with d_{model}=64 and 2 or 4 encoder layers; 3 seeds; 50 epochs with patience 10.

Table S9: Few-shot meta-regression at $L = 32$ (MSE, means over 3 seeds; standard deviations are not displayed). Parameter counts differ and are given below the table. At depth 2, Standard and ST-Large both fit substantially better than the three baselines near MSE 30. At depth 4, ST-Large leads and MZ-Large is second. Perceiver remains near MSE 30 at both depths; the scores alone do not establish convergence.

Depth	Standard	ST-Large	HyperNet	Perceiver	LinearAttn	MZAttn-Full	MZAttn-Large
n=2	**0.79**	<u>0.83</u>	30.75	30.76	30.77	1.01	1.39
n=4	0.45	**0.17**	0.58	30.76	1.24	0.47	<u>0.36</u>

Bold: best in each row. Underline: 2nd best. Param counts shown for the L=32 (n=4) configuration: Standard 146K, ST-Large 255K, HyperNet 297K, Perceiver 314K, LinearAttn 380K, MZAttn-Full 334K, MZAttn-Large 485K; the L=32 (n=2) row uses the same architectures with about 50–55% of these counts.

The seven-model comparison does not show a general MZ advantage: at depth 2, Standard (0.79) and ST-Large (0.83) also fit better than MZ-Full (1.01).

I.5.2 QM9 molecular property prediction

QM9 (Ramakrishnan et al., 2014) contains ∼134K small organic molecules with DFT-computed properties. We parse the original QM9 molecular structures with RDKit (RDKit Contributors, 2024) to obtain true 3D atomic coordinates; each molecule is represented as a set of atoms with 9-dimensional features (5-dim atom type one-hot for H/C/N/O/F, 3D Cartesian coordinates in Å, Gasteiger partial charge), with up to 29 atoms per molecule. We use 50K molecules split 80/10/10 and report 12 properties, including dipole moment, polarizability, orbital energies and the HOMO–LUMO gap. All models use the same architecture as the main experiments (d=64, H=4, d_{ff}=128); results are averaged over 3 seeds with 100 training epochs.

In Table S10, MZ-Large uses $n_g = 16$ and $L = 8$, and ST-XLarge is a standard model of width 92; its HOMO and LUMO scores are 0.883 ± 0.005 and $0.937 \pm <0.001$ over three seeds. Specialized molecular models such as SchNet (Schütt et al., 2017) and DimeNet (Gasteiger et al., 2020) use geometric inductive biases tailored to molecular prediction. Our comparison evaluates a generic set-input architecture and does not claim to outperform those specialized models.

I.5.3 Slot Attention baseline

We compare with Slot Attention (Locatello et al., 2020) on four synthetic aggregation tasks, max, sum, median and range (Table S11).

Table S10: QM9 results on 12 targets (R^2, mean over 3 seeds). The table includes different budgets; ST-XLarge is evaluated only on HOMO/LUMO, and Deep Sets and Slot Attention only on μ, α and Gap (dashes: not evaluated). MZ-Large uses batch size 64 on all targets and exceeds ST-Large on HOMO/LUMO, while the larger ST-XLarge exceeds MZ-Large on both. Default MZ is competitive on several properties but not the highest mean in any column. On Gap, ST-Large 0.879 and MZ-Large 0.874 are close in mean; no statistical equivalence test is supplied. Larger MZ budgets improve HOMO/LUMO/Gap and worsen its other nine means. These outcomes do not identify a target TDOF.

Model	Params	μ	α	HOMO	LUMO	Gap	R^2
Deep Sets	35K	0.437	0.865	–	–	0.653	–
Slot Attention	59K	0.607	0.917	–	–	0.753	–
Set Transf.	208K	0.752	0.905	0.803	0.838	0.858	0.959
ST-Large	361K	**0.795**	0.914	0.842	0.852	0.879	**0.967**
ST-XLarge	443K	–	–	**0.883**	**0.937**	–	–
Perceiver	341K	0.766	**0.920**	0.774	0.819	0.824	**0.967**
HyperNet	360K	0.641	0.902	0.722	0.773	0.786	0.955
MZAttn	368K	0.761	0.911	0.802	0.830	0.855	0.947
MZAttn-Large	417K	0.500	0.698	0.875	0.926	0.874	0.823

Model	ZPVE	U_0	U_{298}	H_{298}	G_{298}	C_v
Deep Sets	–	–	–	–	–	–
Slot Attention	–	–	–	–	–	–
Set Transf.	0.954	0.856	0.856	0.856	0.859	0.940
ST-Large	**0.958**	0.864	0.862	0.863	0.858	**0.950**
ST-XLarge	–	–	–	–	–	–
Perceiver	0.955	**0.891**	**0.886**	**0.883**	**0.888**	0.948
HyperNet	0.934	0.871	0.863	0.869	0.876	0.922
MZAttn	0.952	0.877	0.875	0.871	0.867	0.940
MZAttn-Large	0.702	0.759	0.760	0.722	0.757	0.676

Table S11: Slot Attention baseline comparison on synthetic set tasks (R^2, mean $\pm$ std over 3 seeds). Models are reported at canonical sizes (Slot Attention 59K, Deep Sets 36K, Set Transformer 208K, MZAttn 372K) rather than parameter-matched; the parameter-matched comparison against ST-Large (362K) and HyperNet (360K) is in Table S1. Max uses the Task A data with a 100-epoch cap instead of 200. Slot Attention trails Set Transformer and MZAttn on all four tasks but exceeds Deep Sets on median; the different parameter counts limit architectural attribution.

Model	Max R^2 ↑	Sum R^2 ↑	Median R^2 ↑	Range R^2 ↑
Slot Attention	0.749±0.004	0.892±0.004	0.652±0.002	0.324±0.002
Deep Sets	0.823±0.002	0.995±<0.001	0.646±0.003	**0.348**±0.003
Set Transformer	0.817±0.001	0.960±0.002	**0.659**±0.008	0.344±<0.001
MZAttn (ours)	**0.847**±0.008	**0.999**±<0.001	0.654±0.001	0.343±0.001

I.5.4 Comparison with Euclidean-equivariant GNNs

Symmetry priors are strong competitors (Table S12). The E(n)-equivariant EGNN (Satorras et al., 2021) leads every rotation-invariant covariance statistic, most clearly the PCA error at $d = 16$ (0.934 versus 0.578) and the Mahalanobis distance, which neither attention model fits. Its invariant output cannot fit the fixed-frame quadratic targets, where MZ-Full reaches 0.900–0.958 and leads ST-Large by 0.12. On rotation matching, EGNN reaches 0.98–0.99 at $d = 8$–64, where MZAttn falls from 0.95 to 0.07 (Table S13).

We compare MZAttn to E(n)-equivariant graph neural networks (EGNN) (Satorras et al., 2021), a symmetry-enforcing baseline that uses pairwise distances $\|x_i - x_j\|$ as invariant features. All runs

Table S12: **Symmetry priors.** Mean test R^2 over three seeds under a common training protocol (population standard deviations at most 0.028). The invariant targets are computed from the sample covariance of a set in $\mathbb{R}^d$: its Frobenius norm, the rank-k PCA reconstruction error, and the Mahalanobis distance of a query point. The fixed-frame targets are the quadratic interaction targets of Section 5.3 with K terms in dimension 16, fitted by the default MZ-Full. EGNN has 303K parameters, ST-Large 361–362K and MZ-Full 367–372K. Bold marks the best mean in each column; dashes: not evaluated.

	Rotation-invariant targets					Fixed-frame quadratic		
Model	Cov. d=8	Cov. d=16	PCA d=8	PCA d=16	Mah. d=8	K=4	K=8	K=16
ST	–	–	–	–	–	0.818	0.774	0.767
ST-Large	0.896	0.724	0.736	0.105	−0.004	0.834	0.787	0.779
MZ-Full	0.969	0.964	0.835	0.578	−0.022	**0.958**	**0.904**	**0.900**
EGNN	**0.992**	**0.994**	**0.862**	**0.934**	**0.397**	−0.002	0.000	0.006

Table S13: Equivariant comparison on rotation matching and raw-target convex-hull volume (mean test R^2). Table S12 gives the covariance-statistic and fixed-frame quadratic comparisons.

	Rot. matching			**Hull volume**	
	d=8	d=32	d=64	d=3	d=9
ST-Large	0.92	0.00	−0.02	0.50	−0.09
MZAttn	0.95	0.52	0.07	0.80	0.003
EGNN	**0.99**	**0.99**	**0.98**	**0.95**	**0.85**

use 3 seeds; EGNN (h_{dim}=140, 4 message-passing layers) has ∼303K parameters versus 368K for MZAttn. The comparison keeps the raw-target protocol of Tables S5 and S6 and is not controlled jointly with the normalized sweeps in Figures 2 and 5.

I.5.5 Computational cost

Table S14 reports the overhead of MZAttn relative to the standard Set Transformer at three problem sizes.

Table S14: Computational overhead (ratio MZ/Standard). RTX 5090; batch size 64 for Small and Medium, 32 for Large.

Config	Params	Inference	Training	Memory
Small (d=64, n=30)	1.9×	2.4×	2.2×	2.5×
Medium (d=128, n=50)	1.9×	2.5×	3.8×	2.0×
Large (d=256, n=100)	1.7×	2.7×	5.5×	2.5×

I.5.6 Transfer across representations and tasks

Neural-Process image completion (MNIST, CIFAR-100). Following the Attentive Neural Process setup (Kim et al., 2019), each image defines a 2D function $f : (x, y) \mapsto$ pixel value; the model receives K context pixels with (xy, value) tokens and predicts values at T target coordinates. We compared Standard, ST-Large, HyperNet, Perceiver, LinearAttn, MZAttn-Full, and MZAttn-Large at matched depth (4 layers) using a Gaussian likelihood head. On MNIST (LeCun et al., 1998) (K=50, T=200, 5 seeds), ST-Large achieved the best test NLL of -1.657; MZAttn-Large reached -1.599 at $3.4\times$ the parameters of Standard (-1.617). On CIFAR-100 (Krizhevsky, 2009) (K=100, T=300, 3–5 seeds), Standard attained -1.147 NLL at 144K parameters while MZAttn-Full achieved only -1.140 NLL at 332K parameters; MZ lost by 0.007 NLL while using $2.3\times$ the parameters. These experiments show no advantage for the larger MZ models on image completion. Spatial locality is a possible explanation, but operator variability was not identified or isolated by this comparison.

Few-shot meta-regression on real tabular data (OpenML-CTR23 subset). We built a meta-learning benchmark from six real OpenML regression datasets (California Housing, Diabetes, Energy Efficiency, Abalone, Airfoil, Boston Housing), each restricted or padded to 8 features. Three were used for meta-training and three were held out for test. Each episode sampled $K{=}50$ context rows and $T{=}50$ query rows from a single dataset; models were trained to predict the standardized target under a Gaussian likelihood. Under the held-out split, both Standard and MZAttn-Full exhibit severe distribution shift (test NLL jumps from -0.43 at validation to $+18$–20 at test; 10 seeds with patience 12), showing poor transfer to these held-out datasets without identifying its cause. The two architectures have comparable test NLL (MZAttn-Full $+20.02$ vs. Standard $+18.66$; gap 1.36, pooled std 5.76) and cross-seed standard deviations (± 5.45 vs. ± 6.06), and MZAttn-Full has slightly lower test MSE (1.515 vs. 1.597). Under the easier same-distribution split (test held-out rows from training datasets), MZAttn-Full wins 5.5% on MSE but loses 1.4% on NLL, a mixed signal. Neither split supports a dominance claim.

Rotation matching on real shapes. Task F uses synthetic clouds. With real ModelNet40 shapes under random rotations in SO(3) (three seeds), only Perceiver fits the target well ($R^2 = 0.690$). The Set Transformer reaches 0.231 and MZ-Large 0.179, each with a seed standard deviation above 0.3, while ST-Large (0.003), HyperNet (-0.008) and MZ-Full (-0.066) do not fit it. The advantage on synthetic rotation matching does not transfer to real shapes in this setting.

Systematic real-world transfer study. A cross-representation grid of hull and MST targets, built from 2D pixel coordinates (MNIST, Fashion-MNIST (Xiao et al., 2017) and CIFAR-100, including affine and mixture variants), 8D polynomial lifts and CIFAR-100 patch features, and CIFAR-100 ResNet-18 layer-3 features ($d = 256$), does not establish a general transfer advantage: in its 11 reported cells MZ ranks between second (hull on CIFAR-100 $4{\times}4$ patch features) and fifth by mean score and is never first. MZ ranks 5, 4, 3 and 3 on hull with MNIST, Fashion-MNIST, affine Fashion-MNIST and CIFAR-100 pixels, 4 on both targets for the MNIST 3-digit mixture, 3 and 2 on hull with the MNIST polynomial lift and CIFAR-100 $4{\times}4$ patches, and 4, 4 and 5 on MST with the Fashion-MNIST mixture, CIFAR-100 variable-size patches and ResNet-18 features. The tiers describe ambient dimension and representation, not an isolated intervention on matrix-operator complexity.

CIFAR-100 ResNet-18 feature geometry. The ResNet-18 experiment uses ImageNet-pretrained (Deng et al., 2009) layer-3 features of CIFAR-100, pooled to a 4×4 grid: 16 points in $\mathbb{R}^{256}$ per image. Centered clouds have rank at most 15. Table S15 reports diameter, neighbor radius and MST for all models; two further targets with invalid labels are not reported.

Table S15: CIFAR-100 ResNet-18 feature geometry ($n = 16, d = 256$; mean R^2, seeds 42, 123, 456). Bold marks the largest mean, not statistical significance.

Task	Standard	ST-Large	HyperNet	Perceiver	LinearAttn	MZAttn-Full
Diameter	0.856	0.872	0.840	0.815	**0.882**	0.876
k-NN radius	0.892	0.899	0.898	0.848	**0.904**	0.898
MST weight	0.912	0.919	0.916	0.908	**0.922**	0.914