Matrix Zonotopic Attention: A Context-Adaptive Value Projection for Set Transformers

Zhen Zhang Amr Alanwar
School of Computation, Information and Technology
Technical University of Munich, Germany
{zhenzhang.zhang, alanwar}@tum.de

Abstract

Multi-head attention combines an input-dependent softmax routing with an input-independent linear value projection, so the per-sample operator mapping aggregated values to outputs is the same for every input set. We study the consequences of this asymmetry for permutation-invariant set targets. We introduce the Transformation Degrees of Freedom (TDOF) of a target operator, a complexity measure counting the input-dependent directions an exact representation requires, and present a depth-separation analysis showing that context-rigid attention needs depth proportional to the target's TDOF, whereas a single layer with a context-adaptive value family can represent the same target. Building on this analysis, we propose Matrix Zonotopic Attention (MZAttn), which replaces the fixed value projection with a context-adaptive matrix-zonotope family: a centre matrix plus a sum of generator matrices weighted by input-dependent gates. The construction reduces to standard multi-head attention at initialisation, preserves permutation equivariance, and admits a data-driven reachability interpretation. Experiments on a range of set-prediction tasks are consistent with the TDOF prediction that the architectural advantage is selective: it appears on targets that depend on the input set in a high-rank, sparsely combinatorial way, and is small on aggregate-statistic targets where parameter-matched standard attention is already competitive.

1 Introduction

Attention-based set architectures span point-cloud analysis [21, 28], molecular modelling [33], in-context learning [17], and recent linear-attention variants [16, 39, 44]. A basic question has gone largely unexamined: how much of a standard multi-head attention [41] layer actually adapts to its input? It combines a softmax routing whose weights depend on the input set with a linear value projection that does not (Figure 1, left). On four textbook computational-geometry targets (minimum enclosing ball, convex hull volume, minimum-spanning-tree weight, rotation-dependent matching), every parameter-matched attention baseline using the standard PMA pooling fails to meaningfully improve over the mean predictor as per-point dimension grows, while MZAttn maintains a clean separation (Figure 1, right). The failure is structural rather than capacity-driven: even pushing a HyperNet update to full key-dimension rank at several times MZAttn's parameter budget does not close the gap (Figure 1, right; full sweep in Appendix D.1.5).

Tasks requiring a per-sample linear operator (a sample-specific rotation, a change-of-basis map, or a sparse indicator selecting hull boundary points) lie strictly outside the function family a stack of such layers can represent. We formalise this relational capacity bottleneck through the transformation degrees of freedom (TDOF) of the target operator and prove a matching depth-separation: a worst-case task of TDOF k requires standard-attention depth proportional to k (with a constant determined

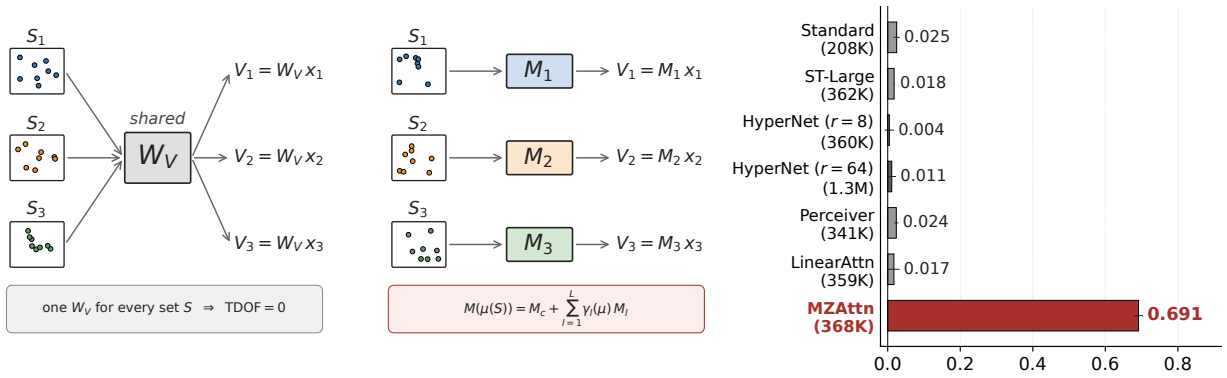

Figure 1: **Standard attention uses a fixed value projection; MZAttn replaces it with a context-adaptive matrix-zonotope family. Left:** standard attention applies the same W_V to every input set, making the per-sample operator independent of S. **Middle:** MZAttn replaces W_V with $M(\mu(S)) = M_c + \sum_l \gamma_l(\mu) M_l$, a context-gated matrix zonotope whose realisation varies per sample. **Right:** on MEB radius (d=8, 5 seeds), all six parameter-matched PMA-pool baselines—including HyperNet at r=d_k=64 (1.3M params, 3.6× MZAttn's budget)—stay at $R^2 \leq 0.025$ while MZAttn reaches $R^2 = 0.69$: capacity is not the bottleneck.

by FFN width), whereas a single layer of our proposed architecture suffices (Theorem 2). The non-linear targets above only approximately inhabit the linearised class; their inheritance of the same bottleneck is an empirical prediction the theory motivates, validated by the synthetic-to-real transfer study below.

We propose Matrix Zonotopic Attention (MZAttn), which replaces the fixed value projection with a learnable matrix zonotope (centre plus a few generator matrices, gated by set-level statistics; Section 2). The parameterisation is minimax-optimal in the parameter-versus-TDOF sense (Theorem 11) and yields single-pass uncertainty estimates. The advantage is selective and predicted by TDOF: on low-TDOF benchmarks (point-cloud classification, molecular property prediction) MZAttn is within 1–3 pp of the best parameter-matched baseline; positive transfer requires high per-sample rank, sparse combinatorial targets, and absence of imposed symmetry, conditions verified on CIFAR-100 ResNet-feature point clouds with a k-center cost target.

2 Method: Matrix Zonotopic Attention

We first introduce the necessary background on zonotopes, matrix zonotopes, and the Set Transformer, then describe the three components of MZAttn: (1) zonotope token representation, (2) the MZ-attention mechanism, and (3) the full MZ-Set Transformer architecture.

2.1 Preliminaries

Definition 1 (Zonotope and Matrix Zonotope [1, 3, 19])**.** *A zonotope with center $c \in \mathbb{R}^n$ and generator columns $G_1, \ldots, G_{n_g} \in \mathbb{R}^n$ (the columns of the generator matrix $G \in \mathbb{R}^{n \times n_g}$) is the set $\mathcal{Z} = \{c + \sum_{g=1}^{n_g} \alpha_g G_g : \alpha_g \in [-1, 1]\}$. A matrix zonotope with center $M_c \in \mathbb{R}^{m \times n}$ and matrix generators $M_1, \ldots, M_L \in \mathbb{R}^{m \times n}$ is the set of matrices*

$$\mathcal{M} = \left\{ M_c + \sum_{l=1}^{L} \xi_l M_l \;\middle|\; \xi_l \in [-1, 1] \right\}. \tag{1}$$

The interval hull of a zonotope is the tightest axis-aligned bounding box, with half-widths $\delta = \sum_{g=1}^{n_g} |G_g|$ (absolute value taken componentwise).

The product of a matrix zonotope with a zonotope is again a zonotope [1]: if $\mathcal{M}$ has L generators and $\mathcal{Z}$ has h generators, then $\mathcal{M}\mathcal{Z} \triangleq \{Mz : M \in \mathcal{M}, z \in \mathcal{Z}\}$ has center $M_c c$ and at most $h + L + Lh$ generators. The $O(Lh)$ growth motivates our fixed-budget management scheme (Section 2.4).

Set Transformer. The Set Transformer [21] processes unordered sets through MAB / ISAB / PMA blocks built from standard scaled dot-product attention; MZAttn replaces the fixed value projection $v_j \mapsto W_V v_j$ in each block.

2.2 Zonotope Token Representation

Given an input element $x_i \in \mathbb{R}^{d_{\text{in}}}$, we project it into a zonotope (c_i, G_i):

$$c_i = \text{LayerNorm}(\text{GELU}(W_c x_i + b_c)), \quad G_i = \frac{1}{\sqrt{n_g d}}(W_g x_i + b_g) \in \mathbb{R}^{n_g \times d}, \tag{2}$$

where $W_c \in \mathbb{R}^{d \times d_{\text{in}}}$, $W_g \in \mathbb{R}^{(n_g d) \times d_{\text{in}}}$, and n_g is the fixed generator budget. The scaling factor $1/\sqrt{n_g d}$ ensures that generators start as small perturbations, so the initial representation is close to a standard vector embedding.

2.3 MZAttn Attention: From Scalar Scores to Relational Operators

Given query, key, and value zonotopes with centers $c_i \in \mathbb{R}^{d_k}$ and generator matrices $G_i \in \mathbb{R}^{n_g \times d_k}$ (the g-th row $G_i[g,:]$ is the g-th generator vector), standard attention scores $\alpha_{ij} = \text{softmax}(q_i^\top k_j / \sqrt{d_k})$ are computed on centers, values are aggregated $(\hat{c}_i, \hat{G}_i) = \sum_j \alpha_{ij}(v_j^c, v_j^G)$, and the head-$h$ matrix zonotope $\mathcal{M}^{(h)} = M_c^{(h)} + \sum_{l=1}^{L} \xi_l M_l^{(h)}$ transforms the output:

$$o_i^c = M_c^{(h)} \hat{c}_i \ \in \ \mathbb{R}^{d_k}, \qquad o_i^G = \hat{G}_i \, (M_c^{(h)})^\top \ + \ W_{\text{mix}} \, \text{stack}_{l=1}^{L}\big[\gamma_l(\mu) \, M_l^{(h)} \hat{c}_i\big] \ \in \ \mathbb{R}^{n_g \times d_k}, \tag{3}$$

where $\hat{G}_i (M_c^{(h)})^\top$ applies $M_c^{(h)}$ row-wise to each generator (preserving the $n_g \times d_k$ shape), $\text{stack}_l[\cdot]$ assembles the L vectors $\gamma_l \, M_l^{(h)} \hat{c}_i \in \mathbb{R}^{d_k}$ into an $L \times d_k$ matrix (rows indexed by l), $W_{\text{mix}} \in \mathbb{R}^{n_g \times L}$ is a learned mixing matrix that preserves the fixed generator budget across layers, and $\gamma = \tanh(W_\gamma \bar{c}^{kv}) \in [-1, 1]^L$ is the context-dependent gate computed from the set-level mean $\bar{c}^{kv} = \frac{1}{n_{kv}} \sum_j c_j^{kv}$. $M_c^{(h)} = I_{d_k}$ and $M_l^{(h)} \approx 0$ at initialisation, so the block reduces to standard attention. Full step-by-step derivation, generator-budget control, and richer permutation-invariant summaries (attention-pooled statistics, higher-order moments) are in Appendix A.1.

2.4 MZ-Set Transformer Architecture

We replace each attention block in a Set Transformer [21] with its MZ counterpart (MZ-MAB, MZ-SAB, MZ-ISAB, and MZ-PMA), yielding InputProj $\to$ [MZ-ISAB]$^{L_{\text{enc}}}$ $\to$ MZ-PMA$_k$ $\to$ [MZ-SAB]$^{L_{\text{dec}}}$ $\to$ OutputHead. The output head concatenates the pooled center $c \in \mathbb{R}^{kd}$ and the interval-hull half-width $\delta = \sum_g |G_g|$ before a two-layer MLP, giving a deterministic prediction and a single-pass uncertainty estimate for free. The expressiveness Theorems 2 and 8 below additionally invoke a variant readout that reads the signed pooled generators $\sum_g G_g$ rather than the IHW $\sum_g |G_g|$ (signed access is straightforward to add as a parallel head; the IHW preserves the uncertainty-bound interpretation of Theorem 4, while the signed readout is what the expressiveness proofs exploit). Both readouts share the same MZ-attention layer; only the post-pooling reduction differs.

3 Theoretical Analysis

We present three groups of results: an expressiveness bottleneck with minimax-optimal resolution (§3.1), TDOF-based depth separation (§3.2), and uncertainty and generalization guarantees (§3.3). Full proofs are in Appendix B.4; additional results in Appendix B.3.

3.1 Expressiveness: Context-Adaptive vs. Context-Rigid Attention

Definition 2 (Context-rigid and context-adaptive relational operators)**.** *Consider an attention layer operating on set $S = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$, and let $\hat{v}_i = \sum_j \alpha_{ij} v_j$ denote the aggregated value for query i. The relational operator $\mathcal{T}$ maps $\hat{v}_i$ into the per-head output. Standard attention is context-rigid: $\mathcal{T}(\hat{v}) = W\hat{v}$ for a fixed $W \in \mathbb{R}^{d_k \times d_k}$, independent of S. MZAttn is context-adaptive: $\mathcal{T}(\hat{v}; S) = M(\mu(S))\hat{v}$ where $M(\mu) = M_c + \sum_{l=1}^{L} \gamma_l(\mu) M_l$ depends on a permutation-invariant statistic $\mu(S)$, and the gates $\gamma_l(\mu) \in [-1, 1]$ are the trained instantiations of the abstract zonotope coefficients ξ_l of Eq. (1).*

This distinction has a fundamental information-theoretic consequence. Standard attention routes information through a scalar bottleneck: the score α_{ij} controls how much of each value to aggregate, but the subsequent transformation W is frozen. We formalise this via the relational complexity / TDOF of the target function, which counts the intrinsic number of context-dependent directions an exact representation requires.

Definition 3 (Relational complexity / TDOF)**.** *For a permutation-invariant target $f(S) = T(S)\,\mu(S)$ with matrix-valued $T\colon \mathcal{X}^{\leq N} \to \mathbb{R}^{m \times d}$ that factors through $\mu(S)$, the relational complexity (henceforth TDOF) is* $\mathrm{TDOF}(f) = \min\{L : T(S) = M_0 + \sum_{l=1}^{L} \sigma_l(\mu(S))\, M_l,\ \sigma_l \text{ 1-Lipschitz},\ |\sigma_l| \leq 1\}$*. The model-free linear dimension $d_T = \dim(\mathrm{span}\{T(S) - T(S') : S, S' \in \mathcal{X}^{\leq N}\})$ used in Appendix B.3 (Definition 4) lower-bounds this constructive count; for $f \in \mathcal{F}_L$ with bounded 1-Lipschitz coordinate functionals (e.g. all natural set-statistical operators in Proposition 12) the two coincide. Functions with* $\mathrm{TDOF}(f) = 0$ *require no input-dependent transformation; those with* $\mathrm{TDOF}(f) > 0$ *fundamentally require the transformation itself to adapt.*

We write $\mathcal{F}_L = \{S \mapsto (M_0 + \sum_{l=1}^{L} \sigma_l(\mu(S))\, M_l)\, \mu(S) : \|M_l\|_F \leq B, |\sigma_l| \leq 1\}$ for the linearised operator family with TDOF at most L. $\mathcal{F}_L$ is the natural linearisation class for any smooth context-adaptive operator: if $T(\mu)$ has Jacobian J with $\mathrm{rank}(J) = r$ at μ_0, then $L = r$ MZ generators capture its first-order behaviour exactly via the SVD of J (Appendix B.3, Proposition 13). The non-linear targets we evaluate (MST, MEB, hull) do not sit exactly in $\mathcal{F}_L$, so TDOF acts as a structural hypothesis validated empirically.

Theorem 1 (Relational capacity bottleneck)**.** *(a) Standard attention without FFN has zero relational capacity: $T_{\mathrm{std}} = \sum_h W_o^{(h)} W_V^{(h)}$ is constant, independent of S. (b) MZAttn-attention represents any function in $\mathcal{F}_L$ with L generators using $(L{+}1)d_k^2 + Ld_k$ parameters per head, and this $\Theta(Ld_k^2)$ count is minimax-optimal.*

Proof sketch. For (a), constant sets S_t make softmax uniform, so the output is a fixed linear map. The matching $\Omega(Ld_k^2 \log(BR/\epsilon))$ lower bound, the FFN-augmented Yarotsky-style bound on attention with FFN, the Eckart–Young residual bound, and a strict separation construction are in Appendix B.3. □

Part (a) is the no-FFN zero-capacity statement (the cleanest qualitative form of the bottleneck); with FFN, zero capacity is replaced by the explicit width/depth tradeoff bound used in part (b)

Table 1: Depth-separation verification on the quadratic-TDOF target (Eq. 4; mean±std, 3 seeds). 1-layer MZAttn holds $R^2 \geq 0.90$ across L while 4-layer standard ST degrades.

L	MZAttn-1L $R^2 \uparrow$	ST-4L $R^2 \uparrow$	MSE ratio $\uparrow$
4	0.995 ± 0.002	0.970 ± 0.005	6.6×
8	0.988 ± 0.001	0.809 ± 0.011	15.4×
16	0.950 ± 0.001	0.618 ± 0.017	7.6×
24	0.944 ± 0.011	0.547 ± 0.020	8.0×
32	0.908 ± 0.016	0.396 ± 0.001	6.5×

and Theorem 2. Thus MZAttn-attention is minimally parameterised for $\mathcal{F}_L$: $\Theta(Ld_k^2)$ parameters versus $O(d_k^4)$ for a general hypernetwork. Natural set operators have explicitly computable TDOF: Mahalanobis distance and whitening have TDOF $= d(d+1)/2$, top-k PCA projection has TDOF $= kd - k(k+1)/2$, and per-coordinate z-score has TDOF $= d$ (Appendix B.3, Proposition 12); soft Chamfer matching has TDOF $\geq d$ (Proposition 14), accounting for the MZAttn gap on Chamfer set-matching (Task B).

3.2 Transformation Degrees of Freedom and Depth Separation

The transformation degrees of freedom $\mathrm{TDOF}(f) = \dim(\mathrm{span}\{T(S) - T(S')\})$ is intrinsic to f (see Definition 4). It yields a depth lower bound:

Theorem 2 (Depth separation via TDOF, worst-case over $\mathcal{F}_k$)**.** *Fix the input distribution to be sets S whose mean $\mu(S) \in \mathbb{R}^d$ has i.i.d. standard-Gaussian coordinates. There exists $f \in \mathcal{F}_k$ on this input distribution (the explicit witness of Lemma 15 with $M_0 = 0$) for which any standard attention network of depth D with FFN width d_{ff} that represents f exactly requires $D \geq k\, d_k/d_{\mathrm{ff}}$, while a single MZAttn layer with $L \geq k$ generators followed by the signed-generator readout (Section 2.4) represents the same f exactly. For approximate representation in $L^2(\mu)$ with L^2-residual error of order $O(\mathrm{Var}(f)/L)$, the same depth lower bound holds up to constants.*

A compositional-orthogonality argument over the Gaussian input distribution (Appendix B.3, Lemma 15) realises the bound for an explicit Frobenius-orthonormal witness ($M_l = e_l e_d^\top$, $M_0 = 0$) even with FFN and residual connections; extending the lower bound to a generic high-TDOF target family with non-trivial cross-layer matrix products is open.

Empirical verification. We verify Theorem 2 on synthetic quadratic targets

$$f(S) = \mathbf{w}^\top \big(M_0 + \sum_{l=1}^{L} \mu(S)_l M_l\big)\, \mu(S) \tag{4}$$

with Frobenius-orthogonal M_l giving $\mathrm{TDOF}(f) = L$. Across $L \in \{4, 8, 16, 24, 32\}$ the four-layer Set Transformer's test R^2 degrades from 0.97 to 0.40 while a single MZAttn layer holds $R^2 \geq 0.90$, with a 6–15× residual-MSE gap (Table 1): adding depth alone cannot cure the bottleneck. EGNN [33] hard-codes rotation equivariance and collapses to $R^2 \approx 0$ for every L because the target depends on a fixed coordinate frame — the complementary failure mode of symmetry-enforcing architectures.

The same picture holds at the architecture level: a 1-layer MZAttn (244K params) outperforms standard Set Transformers up to depth 6 ($\geq$ 480K params; Appendix Table 12), confirming depth separation on real architectures, not only the synthetic-task sweep.

3.3 Uncertainty, Generalization, and Structural Properties

The zonotope representation yields directional perturbation certificates strictly tighter than scalar Lipschitz bounds, with the interval-hull width (IHW) bounding output diameter and variance and thereby serving as a single-pass uncertainty estimate. Appendix B.2 collects the precise statements: a norm-controlled generalisation bound of $\tilde{O}(B_{\mathrm{MZ}}/\gamma\sqrt{n})$ in the spirit of Bartlett et al. [5], universal approximation, and permutation equivariance. At initialisation, MZAttn reduces exactly to standard multi-head attention.

Empirically (Appendix D.3.2, Table 21), the IHW estimate matches a 5-model Deep Ensemble on ECE (0.316 vs. 0.337) and is $\approx 2\times$ better on NLL (5.02 vs. 9.17), at $1\times$ inference cost vs. $5\times$; MC-Dropout achieves the lowest ECE/NLL but at $30\times$ inference cost.

4 Relationship to Prior Work

Zonotopes have served as external verification tools [6, 7, 15, 24] and as set-valued reachability primitives in data-driven control [1, 2, 18]; MZAttn folds a matrix zonotope into a trainable value-projection family inside an attention layer. The closest architectural relatives are context-adaptive linear maps spanning a capacity–parameters spectrum — FiLM [27], discrete MoE [36], low-rank Hypernetworks [12, 34], dynamic filters [14] — among which MZAttn's matrix-zonotope parameterisation occupies the minimax-optimal middle at $O(Ld_k^2)$ parameters per head (Theorem 1); we compare against them as parameter-matched baselines. Linear-attention and selective state-space variants [11, 30, 39, 44] adopt different value-projection inductive biases (input-independent or, for Mamba, input-dependent only along diagonal channels); a Mamba (S6) encoder with PMA pooling matches MZAttn-Full at d=8 on the MEB target but degrades to R^2=0.142±0.134 at d=32 vs. MZAttn's stable 0.378±0.030 (Table 3, Mamba + PMA row), indicating that input-dependent diagonal selectivity alone is dimension-conditional rather than a structural escape from Theorem 1; MZAttn is orthogonal to and composable with them. Set-input networks [21, 28, 47] represent relationships as scalar scores, which we generalise to matrix-valued context-adaptive operators. Full discussion in Appendix C.

5 Experiments

We evaluate MZAttn on eight task families, comparing against seven baselines: Deep Sets [47], a standard Set Transformer [21], ST + FiLM [27], ST-Large (parameter-matched wider variant), HyperNet (low-rank input-dependent value-projection update), Perceiver [13], and Slot Attention [22]. ST-Large, Perceiver, and HyperNet are matched to MZAttn's parameter count (∼360–368K) to control for capacity. Unless noted, experiments are repeated over multiple seeds with mean ± std reported.

5.1 Experimental Setup

All attention-based architectures share a base configuration of d=64, H=4 heads, d_{ff}=128, two ISAB encoder layers, dropout 0.1, and identical training (AdamW [23] with learning rate 10^{-4}, weight decay 10^{-5}, cosine schedule, early stopping at patience 20, all implemented in PyTorch [26]); ST-Large, Perceiver, and HyperNet are configured to match MZAttn's parameter count of ∼365K, and MZAttn uses n_g=8 zonotope generators and L=4 matrix-zonotope generators. Deep Sets, Slot Attention, the FiLM-conditioned Set Transformer, and the per-baseline configurations of HyperNet rank, Perceiver latent count, and Slot iterations are described in Appendix A.1. We evaluate on eight

Table 2: Main results across seven synthetic and point-cloud benchmarks (Tasks A–G). QM9 (Task H) and Slot Attention appear separately in Tables 28 and 29. Best in **bold** among the four parameter-matched attention models (ST-Large, Perceiver, HyperNet, MZAttn, all at ∼360–368K); mean ± std over 3–5 seeds. Smaller-budget baselines (ST + FiLM at 225K, Standard Set Transformer at 208K) are reported for reference and are not subject to the bold rule. Tasks C and E are point-cloud classification; Task G is MST weight prediction in $\mathbb{R}^8$.

	Task A: Set Regr.		**Task B: Set Match.**		**Task F: Rot. Match.**		**Task G: MST Wt.**		**Task C**	**Task E**	**Task D: Few-Shot**		
Model	MSE ↓	R^2 ↑	MSE ↓	R^2 ↑	MSE ↓	R^2 ↑	MSE ↓	R^2 ↑	MN40 ↑	Scan ↑	MSE ($\times 10^{-2}$) ↓	R^2 ↑	Params
Deep Sets	–	–	490.2±33.4	0.652±0.024	–	–	–	–	–	40.6±1.0%	**2.05**±0.38	**0.722**±0.052	36K
Set Transformer	3.43±0.09	0.811±0.005	138.0±10.3	0.901±0.007	487.5±31.1	0.904±0.006	209.8±1.8	0.712±0.002	77.5±0.6%	71.6±2.2%	2.60±0.24	0.646±0.032	208K
ST + FiLM	3.77±0.26	0.792±0.014	48.0±7.9	0.966±0.006	490.5±17.8	0.904±0.003	232.7±12.5	0.681±0.017	78.4±0.8%	68.9±3.2%	2.77±0.21	0.624±0.029	225K
ST-Large	3.36±0.10	0.815±0.005	106.5±5.3	0.924±0.004	423.2±51.8	0.917±0.010	168.7±23.0	0.769±0.032	**79.9**±1.0%	**72.9**±2.4%	–	–	362K
Perceiver	3.47±0.05	0.809±0.003	136.2±17.5	0.902±0.013	872.1±163.8	0.829±0.032	249.0±4.7	0.659±0.006	75.0±0.3%	67.3±1.2%	–	–	341K
HyperNet	3.51±0.01	0.806±0.001	76.1±5.7	0.945±0.004	1373.0±141.6	0.730±0.028	291.2±9.2	0.601±0.013	75.5±1.7%	66.1±0.9%	–	–	360K
MZAttn (ours)	**2.92**±0.09	**0.839**±0.005	**58.4**±11.9	**0.958**±0.009	**241.2**±3.0	**0.953**±0.001	**99.4**±1.9	**0.864**±0.003	78.1±0.7%	69.4±3.7%	2.10±0.16	0.715±0.022	365K

tasks spanning a range of relational complexity (full descriptions in Appendix A.2): set regression on $\mathbb{R}^{16}$ (A), Chamfer-distance set matching in $\mathbb{R}^3$ (B), ModelNet40 [43] (C) and ScanObjectNN [40] (E) point-cloud classification, few-shot set retrieval [37] (D), rotation-dependent matching in $\mathbb{R}^8$ (F), minimum-spanning-tree weight in $\mathbb{R}^8$ (G), and QM9 molecular property prediction [31] (H). All experiments use 200 epochs with early stopping; results are averaged over 3–5 seeds.

5.2 Results

High-TDOF relational tasks (A, B, F, G). MZAttn achieves the best performance on all four: 13% MSE reduction vs. ST-Large on set regression (A), 45% on set matching (B), **51**% over FiLM on rotation matching (F), and **41**% over ST-Large on MST weight (G). On (B), FiLM matches the low-dimensional ($\mathbb{R}^3$) regime (MSE 48.0), but HyperNet's rank-8 update is insufficient (76.1) and higher rank degrades further (Appendix D.1.4). On (F), FiLM performs no better than standard attention, consistent with diagonal modulation's inability to express off-diagonal rotations. On (G), MZAttn's seed std is ±1.9 vs. ±23.0 for ST-Large, indicating stable representation of the pairwise-distance geometry. **Low-TDOF tasks (C, D, E).** On point-cloud classification (ModelNet40 / ScanObjectNN), ST-Large achieves the best accuracy (79.9% / 72.9%) with MZAttn within 1–3 pp (78.1% / 69.4%); these tasks depend primarily on per-element features, so Theorem 1 predicts no advantage for context-adaptivity. On few-shot retrieval (D), Deep Sets ($R^2 = 0.722$) and MZAttn ($R^2 = 0.715$) tie and both exceed Standard attention (0.646) and FiLM (0.624) by +0.07–0.10 R^2, indicating set-aggregation pooling dominates over per-sample operator structure at this scale. On QM9 (Table 28), MZAttn matches the best parameter-matched attention baseline within 0.05 R^2 on every target without leading on any individual one, the predicted low-TDOF behaviour. Slot Attention lags all attention-based models on every task tested (Table 29); all models maintain permutation invariance to within 10^{-5} across 10 random permutations.

5.3 Computational Geometry and Real-World Transfer

The MEB radius depends on at most d+1 extremal points: a sparse combinatorial selector that context-rigid attention cannot express. Table 3 reports R^2 across dimensions for parameter-matched baselines.

Every parameter-matched baseline using the standard PMA pooling, including LA-ST, fails at $R^2 \leq 0.025$, whereas MZAttn-Full reaches a 28–155× gap (median ≈ 40× across 13 positive-baseline cells, with the 155× upper bound at d=8 vs. HyperNet's 0.0044; Figure 2, left). The collapse is

Table 3: MEB radius prediction (R^2, mean±std, 5 seeds). Point clouds are mixtures of Gaussians with $n \in [10, 30]$. LA-ST: linear attention [16]. †Mamba + PMA replaces the ISAB encoder with a Mamba (S6) selective state-space encoder (mambapy 1.2.0). Mean-pool/Mamba dimension-dependence and MZAttn-Large variant (Appendix D.4.1) are discussed in the surrounding paragraph; EGNN comparison: Table 7.

Model	Params	d=8 R^2	d=16 R^2	d=32 R^2
Standard	208K	0.025±0.009	0.020±0.005	0.006±0.012
ST-Large	362K	0.018±0.001	0.012±0.015	−0.006±0.011
HyperNet (r=8)	360K	0.004±0.013	0.005±0.012	−0.001±0.001
HyperNet (r=64, full key-rank)	1.3M	0.011±0.012	0.001±0.009	0.000±0.002
Perceiver	341K	0.024±0.006	0.016±0.028	0.010±0.010
LA-ST (Linear Attn.)	359K	0.017±0.016	0.015±0.009	0.001±0.002
Standard (mean-pool head)	140K	0.784±0.004	0.647±**0.363**	0.449±**0.261**
Mamba + PMA†	645K	0.713±0.025	0.617±0.041	0.142±**0.134**
MZAttn-Full	368K	**0.691**±0.012	**0.629**±0.023	**0.378**±0.030

Table 4: Rotation matching scaling (R^2, mean over 3 seeds). Context-rigid baselines collapse to large negative R^2 as d grows; MZAttn tracks the mean predictor up to d=256. At d=512 all architectures diverge but MZAttn's MSE stays 2.3× smaller than the next-best.

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**

structural: HyperNet at r=d_k=64 (1.3M params, 3.6× budget) leaves R^2 within ±0.012 of zero (Appendix D.1.5). Replacing PMA with mean-pool partially escapes the ceiling at d=8 (0.784±0.004, Standard mean-pool row above) but exhibits heavy-tail bimodal failure at higher dimensions (d=16: 0.647±0.363 with 1/5 seeds catastrophically collapsing to $R^2 \approx 0$; d=32: 0.449±0.261 with another 1/5 catastrophic). Replacing the ISAB encoder with a Mamba (S6) selective state-space encoder while keeping PMA pooling (Mamba + PMA row, 1.75× MZAttn's budget) matches MZAttn-Full at d=8 (0.713±0.025) but degrades to $R^2 = 0.142 \pm 0.134$ at d=32 with all 5 seeds below 0.30. MZAttn-Full + PMA maintains structural stability ($\sigma \leq 0.030$) at $R^2 \geq 0.378$ across all three dimensions, providing the only uniform high-d escape we observed. Convex-hull-volume and a sparse-geometric-set-predicates family confirm the collapse (Appendices D.4.2, D.4.5); rotation matching at d=256 has all baselines at $R^2 \in [-50, -62]$ while MZAttn stays at −0.005 (Table 4). On MST and convex hull, the raw-target negative-R^2 magnitude is partly an optimisation artefact of unnormalised regression on large-magnitude targets; under target normalisation the architectural advantage on MST shrinks to +0.04–0.08 R^2 and the hull d=9 cell ties (Appendix D.2.4). The MEB and quadratic-TDOF separations are scale-free or $O(1)$-magnitude, so the gap there is structural rather than scale-driven.

Real-world transfer. On a CIFAR-100 ResNet-feature setup that meets the three prerequisites (high-rank per-sample points from layer-3 features; sparse k=5 k-center furthest-cost target; no imposed symmetry), MZAttn-Full ranks first across 7 parameter-matched baselines on both ResNet-18 (d=256) and ResNet-50 (d=1024) backbones, with Welch's $t \approx 5.2$, $p \approx 0.0018$ at d=1024 (n=5

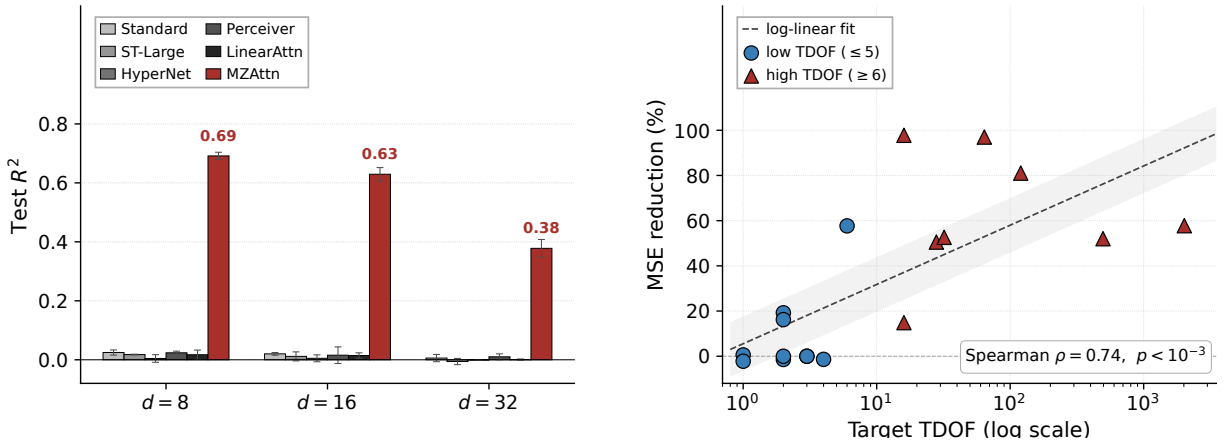

Figure 2: **The architectural advantage tracks transformation degrees of freedom on tasks where TDOF grows with d. Left:** on MEB radius prediction, parameter-matched attention baselines fail at $R^2 \leq 0.025$ across dimensions, while MZAttn reaches $R^2 = 0.69, 0.63, 0.38$ at d=8, 16, 32. **Right:** across 18 (task, scale) settings, MZAttn's MSE reduction over the best baseline correlates with target TDOF (Spearman $\rho = 0.74$, two-sided parametric $p < 0.005$, n=18). Low-TDOF tasks cluster near zero; high-TDOF tasks span 15–98% reduction with median 58%.

seeds; absolute gap +0.0046 R^2, uncorrected for the 18-cell TDOF analysis below). The advantage is restricted to the sparse-combinatorial k-center target: LinearAttn matches or beats MZAttn on dense-statistical siblings (diameter, k-NN radius, MST) on real ResNet features (Appendix D.5, Tables 33 and 32).

Aggregating across eighteen (task, scale) settings, MZAttn's MSE reduction over the best parameter-matched baseline correlates with the estimated TDOF of each task's target operator (Figure 2, right): Spearman $\rho = 0.74$ (n=18, two-sided parametric $p < 0.005$). The TDOF estimate is approximate (exact for linear-form tasks, order-of-magnitude for non-linear ones), so we treat the correlation as descriptive. A ±50% TDOF-perturbation sensitivity check over 10^4 trials keeps Spearman in $[0.675, 0.791]$ at 5%–95% (Appendix D.2.1). Tasks with TDOF ≤ 5 cluster near zero improvement (range -2 to +19%); TDOF ≥ 6 show median 58% MSE reduction (range 15 to 98%).

Architectural ablations, robustness studies, uncertainty calibration, scaling experiments, and the equivariant comparison are reported in Appendix D.

5.4 Analysis: What Does the Matrix Zonotope Learn?

We address why the matrix-zonotope parameterisation works while equally-flexible alternatives do not. **Capacity is not the driver.** HyperNet at r=d_k=64 (1.3M params, 3.6× budget) leaves $R^2 \approx 0$ on MEB (Table 3), and the full $r \in \{16, 32, 64\}$ sweep on Hull and MST at high d (Table 5) stays within 0.13 of r=16 everywhere; conversely MZ-Slim (n_g=2, L=1, 1.6× Standard's parameters) still beats parameter-matched ST-Large and HyperNet variants by 4–23pp R^2 on Max regression, MST weight, and rotation matching (Table 6). On the synthetic quadratic-TDOF task (Table 8), even L=1 retains $R^2 = 0.978$ and the curve plateaus once L exceeds target TDOF: the gain is the centre-plus-gated-generator structure, not the generator count.

A trained L=8 checkpoint probed with 1,000 contexts shows near-orthogonal generators in operator space, gates active on 78.6% of evaluations, and the soft effective rank of $\{M(\mu_i) - \bar{M}\}_i$ saturating the architectural budget when $L \leq$ TDOF and growing sublinearly above (Appendix D.1,

Table 5: HyperNet rank extended ablation (R^2, mean±std, 3 seeds). Low-rank update $r \in \{16, 32, 64\}$ on collapse-exhibiting tasks. ‡Hull d=9 MZ-Full (0.003) is within seed std of zero. Target-normalised equivalents: Table 18.

Task (d)	HyperNet r=16	r=32	r=64	MZAttn-Full
Hull d=3	$0.191_{\pm.008}$	$0.190_{\pm.029}$	$0.204_{\pm.020}$	$\mathbf{0.804}_{\pm.014}$
Hull d=6	$-0.214_{\pm.013}$	$-0.203_{\pm.002}$	$-0.214_{\pm.027}$	$\mathbf{0.103}_{\pm.020}$
Hull d=9	$-0.095_{\pm.001}$	$-0.098_{\pm<.001}$	$-0.100_{\pm.002}$	$0.003_{\pm.006}$‡
MEB d=8	$0.012_{\pm.006}$	$0.022_{\pm.006}$	$0.011_{\pm.012}$	$\mathbf{0.691}_{\pm.012}$
MEB d=16	$0.007_{\pm.003}$	$0.000_{\pm.008}$	$0.001_{\pm.009}$	$\mathbf{0.629}_{\pm.023}$
MEB d=32	$-0.020_{\pm.029}$	$-0.000_{\pm<.001}$	$0.000_{\pm.002}$	$\mathbf{0.378}_{\pm.030}$
MST d=64	$-1.96_{\pm.31}$	$-2.06_{\pm.28}$	$-1.92_{\pm.18}$	$\mathbf{0.58}_{\pm.01}$
MST d=128	$-3.09_{\pm.29}$	$-3.00_{\pm.17}$	$-2.80_{\pm.17}$	$\mathbf{0.48}_{\pm.02}$
MST d=256	$-3.99_{\pm.15}$	$-3.78_{\pm.05}$	$-3.97_{\pm.17}$	$\mathbf{0.38}_{\pm.04}$

Table 6: MZ parameter efficiency (R^2, mean, 3 seeds). Even at n_g=2, L=1 MZ retains its high-TDOF advantage: structure dominates parameter count.

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

Figure 4); the trained model allocates operator-space variation up to but not beyond the intrinsic TDOF, with L a soft upper bound.

6 Discussion and Conclusion

MZAttn replaces the fixed value projection of standard attention with a learnable matrix zonotope at the same parameter-budget scaling. The gain is selective and tracks the theory (Figure 2): improvements concentrate on tasks with high-TDOF, sparse-combinatorial targets, and disappear on tasks dominated by per-element features, spatial locality, or known symmetry.

Equivariance versus context-adaptivity. EGNN [33] matches or outperforms MZAttn on every rotation-invariant task tested: rotation matching at d=8 ($R^2 = 0.99$ vs. 0.95), MEB radius ($R^2 = 0.78$ vs. 0.38 at d=32), and Mahalanobis distance ($R^2 = 0.40$ vs. -0.02). On the quadratic-TDOF target (genuine fixed-frame dependence), EGNN's invariant-output constraint forces $R^2 \approx 0$ while MZAttn reaches $R^2 \geq 0.90$ (Table 7). The two are complementary: EGNN when symmetry is known, MZAttn when it is unknown or absent.

Real-world transfer. On convex-hull-volume across three pretrained backbones (ResNet-50, DINOv2, CLIP) with five parameter-matched context-rigid baselines, 14 of 15 cells yield negative R^2 (Appendix D.5), while MZAttn-Full maintains mean $R^2 \geq 0.82$ on all three backbones (MZAttn-Large reaches 0.998 ± 0.005 on RN50). On dense-statistical siblings (diameter, k-NN radius, MST) all architectures cluster within ≤ 0.03 R^2, the predicted TDOF behaviour. On SST-2 sentiment [38] (low-TDOF), MZAttn (0.7982 ± 0.0114) matches a parameter-matched Standard Set Transformer (0.7989 ± 0.0084) within 0.07pp (3 seeds), the strict-generalisation property a value-projection replacement should satisfy.

Table 7: MZAttn vs. EGNN. EGNN dominates on rotation/translation-invariant targets; MZAttn dominates on the non-equivariant quadratic-TDOF target where invariance is structurally inadequate.

	Rot. matching			**Hull volume**		**MEB radius**		**Statistical (Cov, Mah, PCA)**			**Quadratic TDOF**		
	d=8	d=32	d=64	d=3	d=9	d=8	d=32	CN d=8	Mah d=8	PCA d=16	L=4	L=8	L=16
ST-Large	0.92	0.00	−0.02	0.50	−0.09	0.02	−0.01	0.90	−0.004	0.11	0.83	0.79	0.78
MZAttn (ours)	0.95	0.52	0.07	0.80	0.003	0.69	0.38	0.97	−0.022	0.58	**0.96**	**0.90**	**0.90**
EGNN	**0.99**	**0.99**	**0.98**	**0.95**	**0.85**	0.66	**0.78**	**0.99**	**0.40**	**0.93**	−0.002	0.000	0.006

Scope and limitations. The advantage is structurally targeted, not universal: synthetic high-TDOF benchmarks serve as the controlled testbed; the 18-setting Spearman $\rho = 0.74$ (Section 5) is the falsifiability check; QM9 and point-cloud rows are the predicted low-TDOF negative controls (within ≤ 0.04 R^2). MZAttn costs 1.7–1.9× parameters and 2.4–2.7× inference vs. Standard ST (Appendix D.5.5); the Slim variant retains most of the advantage at 1.6× params. Deep Sets tying on few-shot retrieval (Task D) and ST-Large winning the parameter-fair 7-way meta-regression at L=32 (Table 27; an earlier comparison against an undertuned vanilla baseline overstated this gap, corrected here) instantiate the predicted boundary. Very-large-scale sets and structured-zonotope variants remain open.

References

[1] Amr Alanwar, Anne Koch, Frank Allgöwer, and Karl Henrik Johansson. Data-driven reachability analysis using matrix zonotopes. In *Learning for Dynamics and Control*, volume 144, pages 163–175. PMLR, 2021.

[2] Amr Alanwar, Anne Koch, Frank Allgöwer, and Karl Henrik Johansson. Data-driven reachability analysis from noisy data. *IEEE Transactions on Automatic Control*, 68(5):3054–3069, 2023. doi: 10.1109/TAC.2023.3257167.

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

[4] C Bradford Barber, David P Dobkin, and Hannu Huhdanpaa. The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software*, 22(4):469–483, 1996. doi: 10.1145/235815.235821.

[5] Peter L Bartlett, Dylan J Foster, and Matan J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[6] Gregory Bonaert, Dimitar I Dimitrov, Maximilian Baader, and Martin Vechev. Fast and precise certification of transformers. In *Proceedings of the 42nd ACM SIGPLAN International Conference on Programming Language Design and Implementation*, pages 466–481, 2021. doi: 10.1145/3453483.3454056.

[7] Long Kiu Chung and Shreyas Kousik. Provably-safe neural network training using hybrid zonotope reachability analysis. In *2025 IEEE 64th Conference on Decision and Control (CDC)*. IEEE, 2025. URL `https://ieeexplore.ieee.org/document/11312423/`.

[8] Ronald A DeVore. Nonlinear approximation. *Acta Numerica*, 7:51–150, 1998. doi: 10.1017/S0962492900002816.

[9] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059. PMLR, 2016.

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

[11] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Conference on Language Modeling*, 2024.

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

[13] Andrew Jaegle, Felix Gimeno, Andrew Brock, Oriol Vinyals, Andrew Zisserman, and João Carreira. Perceiver: General perception with iterative attention. In *Proceedings of the 38th International Conference on Machine Learning*, pages 4651–4664. PMLR, 2021.

[14] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc Van Gool. Dynamic filter networks. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

[15] Matt Jordan, Jonathan Hayase, Alexandros G Dimakis, and Sewoong Oh. Zonotope domains for lagrangian neural network verification. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

[16] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, volume 119, pages 5156–5165. PMLR, 2020.

[17] 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.

[18] Niklas Kochdumper and Matthias Althoff. Constrained polynomial zonotopes. *Acta Informatica*, 60(3):279–316, 2023. doi: 10.1007/s00236-023-00437-5.

[19] W. Kühn. Rigorously computed orbits of dynamical systems without the wrapping effect. *Computing*, 61(1):47–67, 1998. doi: 10.1007/BF02684450.

[20] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[21] 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 *Proceedings of the 36th International Conference on Machine Learning*, pages 3744–3753. PMLR, 2019.

[22] 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*, volume 33, pages 11525–11538, 2020.

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

[24] Matthew Mirman, Timon Gehr, and Martin Vechev. Differentiable abstract interpretation for provably robust neural networks. In *International Conference on Machine Learning*, volume 80, pages 3578–3586. PMLR, 2018.

[25] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT Press, Cambridge, MA, 2 edition, 2018. ISBN 9780262039406.

[26] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[27] 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.

[28] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 652–660, 2017.

[29] 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*, volume 30, 2017.

[30] Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. In *Advances in Neural Information Processing Systems*, 2025. Best Paper Award.

[31] Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1(1):140022, 2014. doi: 10.1038/sdata.2014.22.

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

[33] Víctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 9323–9332. PMLR, 2021.

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

[35] Kristof T Schütt, Pieter-Jan Kindermans, Huziel E Sauceda, 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*, volume 30, 2017.

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

[37] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[38] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, 2013.

[39] Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. Retentive network: A successor to transformer for large language models. *arXiv preprint arXiv:2307.08621*, 2023.

[40] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1588–1597, 2019.

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

[42] Emo Welzl. Smallest enclosing disks (balls and ellipsoids). In *New Results and New Trends in Computer Science (Lecture Notes in Computer Science vol. 555)*, pages 359–370. Springer, 1991. doi: 10.1007/BFb0038202.

[43] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3D ShapeNets: A deep representation for volumetric shapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1912–1920, 2015.

[44] Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 56501–56523. PMLR, 2024.

[45] Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27(5):1564–1599, 1999. doi: 10.1214/aos/1017939142.

[46] Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94:103–114, 2017. doi: 10.1016/j.neunet.2017.07.002.

[47] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Supplementary Material: Overview

The supplementary material is organised as follows. Appendix A reports experimental setup, hyperparameters, and compute resources. Appendix B contains the algebraic preliminaries, uncertainty and generalisation results, and full proofs for all theorems and propositions. Appendix C expands the related-work discussion. Appendix D reports the extended empirical study, organised into five thematic blocks: architectural ablations, robustness and generalisation, uncertainty calibration, computational-geometry scaling, and real-world transfer plus equivariant comparison.

A Experimental Setup, Hyperparameters, and Compute

Appendix A.1 gives the model and optimisation choices common to all experiments; Appendix A.2 specifies the per-task data-generation and training settings; Appendix A.3 reports compute resources.

A.1 Implementation Details

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

Optimiser	AdamW [23]	LR	10^{-4}
LR schedule	cosine to 0	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 (geom. tasks)	Epochs (cap)	200 (synthetic), 150 (real-world)
Early stopping	val loss patience 20	Dropout	0.1 (constant)
d_{model}	64 (default) / 80 (ST-Large)	H heads	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)

The ResNet-50 (d=1024) experiments use early-stopping patience 300 (Appendix D.4.7 verifies the rank ordering is stable for any patience ≥ 100).

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), since generators represent small perturbation directions and require less modeling capacity.

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}).

A.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 $n \sim \text{Uniform}(10, 100)$, predict $y = \sum_i [\max_{\text{dim}} x_i]$, i.e., the sum of dimension-wise 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 [43], 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 [40], a realistic 3D point cloud classification benchmark constructed from real indoor scans. We use the OBJ_BG variant (15 classes, ∼2,300 train, ∼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. This task requires the model to internally discover and apply a context-dependent rotation to align A and B before computing the residual distance, an inherently off-diagonal operation. FiLM conditioning, which applies $\gamma(\mu) \odot z + \beta(\mu)$ (diagonal scale and shift), provably cannot express rotation in $\mathbb{R}^8$. 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.

Task H: QM9 molecular property prediction. Given a molecule represented as a set of atoms $S = \{a_1, \ldots, a_n\}$ with $n \leq 29$ and $a_i \in \mathbb{R}^9$ (5-dim atom type one-hot encoding for H/C/N/O/F, 3D Cartesian coordinates, partial charge), predict quantum chemical properties: dipole moment (μ, Debye) and isotropic polarizability (α, Bohr^3). Data is drawn from the QM9 dataset [31], which contains ∼134K small organic molecules with DFT-computed properties. We use 50K molecules split 80/10/10; 100 epochs; 3 seeds.

A.3 Compute Resources

All experiments were run on a mix of consumer and datacentre GPUs. Synthetic benchmarks (Tasks A, B, F, G, MEB, hull, MST, rotation matching at $d \leq 64$, sparse-geometric set predicates, gen-/HyperNet-/MZ-Slim ablations, uncertainty calibration, set-size generalisation, ambiguity sweep, quadratic-TDOF / depth separation / sequence MZ) 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). ResNet-50 layer-3 transfer at d=1024 used a single NVIDIA H200. Per-seed training time on the 5090 ranged from 0.5 to 30 minutes depending on task and dimension. The full empirical sweep (all main-body tables and appendix ablations combined) accumulates approximately 1,200 GPU-hours when re-run from scratch with default seeds.

B Theory: Algebra, Structural Properties, and Proofs

This section consolidates the theoretical material that was previously fragmented across four short appendices (zonotope algebra, uncertainty certificates, additional theoretical results, and full proofs). It collects the algebraic preliminaries used by the proofs in Section 3, the structural-property statements summarised in the body, and the proofs of every theorem and proposition cited above.

B.1 Zonotope and Matrix-Zonotope Algebra

For readers unfamiliar with zonotopes, we provide a self-contained primer.

A zonotope $\mathcal{Z} = \{c + G\xi : \|\xi\|_\infty \leq 1\}$ can be visualized as a centrally symmetric polytope generated by "sweeping" the center along each generator direction. In 2D, a zonotope with 3 generators is a hexagon; with 4 generators, an octagon.

The interval hull $\mathrm{IH}(\mathcal{Z}) = [c - \delta, c + \delta]$ with $\delta_i = \sum_k |G_{ik}|$ is the tightest axis-aligned box containing $\mathcal{Z}$.

Key operations:

- **Linear map:** $A\mathcal{Z} = \{Ac + AG\xi\} = \langle Ac, AG\rangle$. Zonotopes are closed under linear maps.
- **Minkowski sum:** $\mathcal{Z}_1 \oplus \mathcal{Z}_2 = \langle c_1 + c_2, [G_1, G_2]\rangle$. Generator count grows additively.
- **MZ multiplication:** $\mathcal{MZ}$ produces a zonotope with center $M_c c$ and generators $\{M_c g_k\} \cup \{M_l c\} \cup \{M_l g_k\}$. Generator count grows as $h + L + Lh$ (hence the need for budget control).

For a thorough treatment including constrained and polynomial variants, see Kochdumper and Althoff [18] and Althoff [3].

B.2 Uncertainty Certificates, Generalization, and Structural Properties

This section contains the full theorem statements deferred from Section 3.3.

Theorem 3 (Directional vs. scalar sensitivity)**.** *Consider a single MZ-attention layer with fixed attention scores. Under value perturbation $\tilde{\hat{c}} = \hat{c} + \hat{G}\xi$ for $\xi \in [-1, 1]^{n_g}$:*

(a) A standard attention layer with output projection W_o provides only a scalar bound: $\|\Delta o\| \leq \|W_o\|_2 \cdot \|\hat{G}\|_{1,2}$.

(b) MZ-attention provides a directional zonotope certificate: $\Delta o \in \mathcal{Z}_\Delta = \{M_c\hat{G}\xi : \|\xi\|_\infty \leq 1\}$, contained in the ball of radius $\|M_c\|_2 \|\hat{G}\|_{1,2}$.

(c) For matched output-projection norms ($\|W_o\|_2 = \|M_c\|_2$), the MZ zonotope is contained in the scalar ball, $\mathcal{Z}_\Delta \subseteq B(\|M_c\|_2 \|\hat{G}\|_{1,2})$, and is strictly tighter than the ball along every direction orthogonal to the column span of $M_c\hat{G}$ whenever that column span is not the full ambient space.

The qualitative content of part (c) is that MZ replaces a uniform-radius ball with an axis-aligned-in-input-space zonotope; this distinguishes MZAttn from generic context-adaptive methods, which provide only a scalar reachable diameter, by specifying the change per direction.

Theorem 4 (IHW as prediction uncertainty bound)**.** *Let $\mathrm{IHW}(Z_{\mathrm{out}}) = \sum_{g=1}^{n_g} \|G_g\|_1$ denote the interval hull width of the output zonotope. Then:*

(a) The reachable output diameter satisfies $\mathrm{diam}(\mathcal{Y}(S)) \leq 2\,\|W_{\mathrm{out}}\|_2 \cdot \mathrm{IHW}$.

(b) The output variance under uniform ξ satisfies $\mathrm{Var}_\xi[W_{\mathrm{out}}^\top(c_{\mathrm{out}} + G_{\mathrm{out}}\xi)] \leq \frac{1}{3}\|W_{\mathrm{out}}\|_2^2 \cdot \mathrm{IHW}^2$.

(c) Higher IHW implies a strictly wider reachable output range (monotonicity).

Theorem 5 (Norm-controlled generalization)**.** *For an MZ-Set Transformer with D encoder layers, i.i.d. training data of size n, and margin $\gamma > 0$, the generalization gap is bounded by $\tilde{O}(B_{\mathrm{MZ}}/\gamma\sqrt{n})$, where $B_{\mathrm{MZ}} = \prod_{\ell=1}^{D}(\|M_c^{(\ell)}\|_2 + \sum_l \|M_l^{(\ell)}\|_2) \cdot \|W_V^{(\ell)}\|_2 \cdot (\sum_\ell R_\ell^{2/3})^{3/2}$. At initialisation, when $M_c = I$ and $M_l \approx 0$, B_{MZ} matches standard attention; it grows only as the model learns to depart from context-rigid behaviour.*

Proof sketch. The result follows from the spectrally-normalised Rademacher complexity bound of Bartlett et al. [5] with one substitution. The post-aggregation map of an MZ-attention layer is $\hat{v} \mapsto M(\mu(S))\hat{v}$ with worst-case spectral norm $\|M(\mu(S))\|_2 \leq \|M_c\|_2 + \sum_l |\gamma_l(\mu)|\,\|M_l\|_2 \leq \|M_c\|_2 + \sum_l \|M_l\|_2$, since $|\gamma_l(\mu)| \leq 1$. This replaces the per-layer spectral factor $\|W^{(\ell)}\|_2$ in the standard bound by the MZ-layer worst-case spectral norm. The aggregation step $\sum_j \alpha_{ij} v_j$ is a softmax-weighted convex combination, hence a 1-Lipschitz contraction in $\hat{v}$ that does not inflate the Rademacher complexity by the Talagrand contraction principle [25, Lemma 5.7]. Composing across the D MZ-encoder layers and including the value projections $W_V^{(\ell)}$ yields the stated B_{MZ}; the residual $(\sum_\ell R_\ell^{2/3})^{3/2}$ factor and the post-hoc-margin transfer are identical to the original argument. □

This bound has a natural interpretation: generalisation is controlled by how much context-adaptivity the model learns to use, giving an interpretable complexity measure that connects to the relational-capacity framework.

Corollary 6 (Universality is preserved)**.** *The MZ-Set Transformer family contains the standard Set Transformer family as the special case $M_l^{(\ell)} = 0$, $W_{\mathrm{mix}}^{(\ell)} = 0$, $M_c^{(\ell)} = W_o^{(\ell)}$. Consequently, the universal-approximation property of Lee et al. [21, Theorem 2] carries over: for any compact $\mathcal{K} \subset \mathbb{R}^d$, continuous permutation-invariant $f\colon \mathcal{K}^{\leq N} \to \mathbb{R}^m$, and $\epsilon > 0$, an MZ-Set Transformer ϵ-approximates f uniformly. The MZ family is strictly larger (Theorem 8), so universality is preserved without loss.*

Proposition 7 (Structural properties)**.** *MZ-SAB and MZ-ISAB are permutation-equivariant; the full MZ-Set Transformer is permutation-invariant. At initialisation, with $M_c = I$, $M_l \approx 0$, and $W_{\mathrm{mix}} \approx 0$, MZAttn reduces exactly to standard multi-head attention with an identity output projection.*

B.3 Additional Theoretical Results

This section contains the expressiveness results and supporting lemmas deferred from Section 3.

Theorem 8 (Strict expressiveness separation)**.** *For any nonzero $A \in \mathbb{R}^{d\times d}$, define $f_A(S) = (I + \tanh(\mu(S)_1) \cdot A)\,\mu(S)$. Then: (a) a single MZAttn-attention head with $L = 1$ followed by a fixed linear readout (over the pooled center and the first signed pooled generator) computes f_A exactly via the explicit closed-form construction below; (b) no standard multi-head attention layer (any H, without FFN, with any output linear projection) can represent f_A on the family of constant inputs $\{S_t\}_{t\in\mathbb{R}}$.*

Proof. (a) We construct each parameter explicitly. Set $W_Q = 0$ so all softmax weights are uniform $\alpha_{ij} = 1/n$, giving the aggregated center $\hat{c}_i = \mu(S)$ for every query i. Set $W_V = I_d$ so the value zonotope's center equals the input. Set $M_c^{(h)} = I_d$, $L = 1$, $M_1^{(h)} = A$, the gate weight $w_\gamma = e_1 \in \mathbb{R}^d$ so $\gamma_1(\mu) = \tanh(e_1^\top \mu) = \tanh(\mu_1)$, and the mixing weight $W_{\text{mix}} = e_1 \in \mathbb{R}^{n_g\times 1}$ (only the first generator slot is used). Substituting into Eq. (3) gives center output $o_i^c = \mu(S)$ and generator output whose first row is $o_i^G[1,:] = \tanh(\mu_1)\,A\,\mu(S)$, with rows $2,\dots,n_g$ all zero.

After PMA pooling with a single seed vector configured analogously ($W_Q = 0$, uniform aggregation), the pooled center is $\mu(S)$ and the pooled signed first-generator is $\tanh(\mu_1)\,A\,\mu(S)$. The fixed linear readout $W_{\text{out}} : (c, g_1) \mapsto c + g_1$ then yields

$$W_{\text{out}}(o^c, o^G[1,:]) = \mu(S) + \tanh(\mu_1)\,A\,\mu(S) = (I + \tanh(\mu_1)\,A)\,\mu(S) = f_A(S),$$

exactly. No nonlinear approximation is invoked: every operation in the construction is either a fixed linear map or the native MZ gate $\tanh(w_\gamma^\top \bar{c}^{kv})$, both of which compute their output without finite-width truncation.

(b) On constant sets $S_t = \{te_1,\dots,te_1\}$, softmax symmetry gives $o(S_t) = (\sum_h W_o^{(h)} W_V^{(h)})te_1 = t\,We_1$ (linear in t). But $f_A(S_t) = te_1 + t\tanh(t)\,Ae_1$ is nonlinear in t whenever $Ae_1 \neq 0$, contradicting linearity. □

Corollary 9 (Depth separation without FFN)**.** *There exist continuous permutation-invariant functions computable by a single MZ-attention layer that require at least two standard attention layers with intermediate nonlinearity.*

Theorem 10 (Width efficiency over standard attention with FFN)**.** *For $\mathcal{F}_L = \{S \mapsto (M_0 + \sum_{l=1}^{L} \sigma_l(\mu(S))M_l)\,\mu(S)\}$: (a) MZ-attention represents any $f \in \mathcal{F}_L$ exactly with $(L{+}1)d_k^2 + O(Ld_k)$ parameters; (b) standard attention with depth-2 ReLU FFN requires $d_{\text{ff}} = \Omega(Ld_k/\sqrt{\epsilon})$ for ϵ-approximation.*

Proof sketch. (a) Direct. (b) Each bilinear term $\sigma_l \cdot (M_l h)_j$ requires $\Omega(1/\sqrt{\epsilon})$ ReLU neurons [46]; summing over L terms and d_k coordinates gives the bound. □

Theorem 11 (Minimax parameter efficiency)**.** *Let $\mathcal{F}_{L,B}$ denote the class $\mathcal{F}_L$ with bounded generator matrices $\|M_l\|_F \le B$ and 1-Lipschitz gates. (a) MZ-attention parameterizes $\mathcal{F}_{L,B}$ exactly using $P_{\text{MZ}} = (L{+}1)d_k^2 + O(Ld_k) = \Theta(Ld_k^2)$ parameters; (b) any model class ϵ-approximating $\mathcal{F}_{L,B}$ in sup-norm over inputs of radius R requires $P \ge \Omega(Ld_k^2 \log(BR/\epsilon))$ parameters; (c) hence the MZ parameterization is order-optimal up to logarithmic factors.*

Definition 4 (Transformation degrees of freedom)**.** *For a permutation-invariant function $f(S) = T(S)\,\mu(S)$ with set-dependent linear operator $T(S) \in \mathbb{R}^{d\times d}$, the transformation degrees of freedom is $\text{TDOF}(f) = \dim\big(\text{span}\{T(S) - T(S') : S, S' \in \mathcal{X}^*\}\big)$, i.e. the dimension of the subspace of $\mathbb{R}^{d\times d}$ spanned by all possible differences of T across inputs.*

TDOF of natural set functions.

Proposition 12 (TDOF of natural set functions)**.** *Let $S \subset \mathbb{R}^d$ with $|S| \geq d{+}1$ in general position, sample mean $\mu(S)$, and regularised covariance $\hat{\Sigma}(S) = \frac{1}{|S|}\sum_i (x_i - \mu)(x_i - \mu)^\top + \lambda I$ for $\lambda > 0$. (a) Mahalanobis ($T(S) = \hat{\Sigma}(S)^{-1}$) has* TDOF $= d(d{+}1)/2$*; (b) Whitening ($T(S) = \hat{\Sigma}(S)^{-1/2}$) has* TDOF $= d(d{+}1)/2$*; (c) PCA projection onto the top k principal components has* TDOF $= kd - k(k{+}1)/2$*; (d) Per-coordinate z-score ($T(S) = \mathrm{diag}(\sigma_j^{-1})$) has* TDOF $= d$*.*

Proof. **(a)–(b).** As S varies over generic configurations of $|S| \geq d{+}1$ points in $\mathbb{R}^d$, the regularised covariance $\hat{\Sigma}(S)$ ranges over all of Sym_d^+ (the cone of $d \times d$ positive-definite matrices), an open subset of the symmetric matrices Sym_d. The maps $\Sigma \mapsto \Sigma^{-1}$ and $\Sigma \mapsto \Sigma^{-1/2}$ are smooth diffeomorphisms on Sym_d^+ with non-singular Jacobian (e.g. $\partial\Sigma^{-1}/\partial\Sigma \cdot H = -\Sigma^{-1}H\Sigma^{-1}$ is full-rank as a linear map $\mathrm{Sym}_d \to \mathrm{Sym}_d$), so $T(S)$ ranges over an open subset of Sym_d. Hence $\mathrm{span}\{T(S) - T(S')\}$ contains a basis of Sym_d, giving TDOF $= \dim \mathrm{Sym}_d = d(d{+}1)/2$. **(c).** The top-$k$ PCA projector is $T(S) = V_k(S)V_k(S)^\top$ where $V_k \in \mathbb{R}^{d\times k}$ has orthonormal columns. As S varies, V_k ranges over the Stiefel manifold $\mathrm{St}(k, d)$ of dimension $kd - k(k{+}1)/2$, so the smooth orbit $\{V_kV_k^\top\}$ is the Grassmannian $\mathrm{Gr}(k, d)$ of the same dimension, and the linear span of differences fills the tangent space at any point, which is also of dimension $kd - k(k{+}1)/2$. **(d).** $T(S) = \mathrm{diag}(1/\sigma_1, \ldots, 1/\sigma_d)$ where $\sigma_j(S)$ varies independently over $(\lambda^{1/2}, \infty)$ as S ranges over configurations whose marginal variances can be set freely; the differences span the d-dimensional space of diagonal matrices. □

$\mathcal{F}_L$ as first-order approximation of smooth operators. The following result shows that $\mathcal{F}_L$ is not an artificial class tailored to MZ-attention, but rather the natural linearization class for any smooth context-adaptive operator.

Proposition 13 ($\mathcal{F}_L$ captures linearized context-adaptive operators)**.** *Let T: $\mathbb{R}^d \to \mathbb{R}^{m\times d}$ be twice continuously differentiable, and define $f(S) = T(\mu(S))\,\mu(S)$ where $\mu(S) = \frac{1}{|S|}\sum_{x\in S} x$. Let $J = \frac{\partial\,\mathrm{vec}(T)}{\partial\mu}\big|_{\mu_0} \in \mathbb{R}^{md\times d}$ be the Jacobian at a reference point μ_0, with SVD $J = \sum_{l=1}^r s_l\, u_l v_l^\top$ where $r = \mathrm{rank}(J)$. Define the linearized operator*

$$\tilde{T}(\mu) = T(\mu_0) + \sum_{l=1}^{r} s_l\,(v_l^\top(\mu{-}\mu_0))\,\mathrm{mat}(u_l),$$

where mat: $\mathbb{R}^{md} \to \mathbb{R}^{m\times d}$ *reshapes vectors into matrices, and the corresponding function $\tilde{f}(S) = \tilde{T}(\mu(S))\,\mu(S)$. Then:*

(a) $\tilde{f} \in \mathcal{F}_{r,B}$ with $B = s_1$ (the largest singular value of J), center matrix $M_0 = T(\mu_0)$, generator matrices $M_l = s_l\,\mathrm{mat}(u_l)$, and gating functions $\sigma_l(\mu) = v_l^\top(\mu - \mu_0)$, which are linear and hence 1-Lipschitz after rescaling by $1/\|v_l\|$.

(b) The approximation error satisfies

$$\|f(S) - \tilde{f}(S)\| \leq \tfrac{1}{2}\,C_T\,\|\mu(S) - \mu_0\|^2 \cdot \|\mu(S)\|,$$

where $C_T = \sup_\mu \|\nabla^2 \mathrm{vec}(T)(\mu)\|_{\mathrm{op}}$ is bounded on any compact domain.

(c) Consequently, $L = \mathrm{rank}(J)$ MZ generators suffice to capture the first-order context-adaptive behavior of T around μ_0, and Theorem 11 guarantees this representation is parameter-optimal.

Proof. Part (a) is immediate from the SVD construction: each component $\sigma_l(\mu) \cdot M_l$ is a rank-1 term in the Jacobian decomposition of T, and the gating functions are linear projections of μ.

Part (b) follows from Taylor's theorem applied to the vector-valued map $\mathrm{vec}(T)$:

$$\|\mathrm{vec}(T(\mu)) - \mathrm{vec}(T(\mu_0)) - J(\mu - \mu_0)\| \leq \tfrac{1}{2} C_T \|\mu - \mu_0\|^2.$$

Reshaping and multiplying by $\mu(S)$: $\|f(S) - \tilde{f}(S)\| = \|(T(\mu) - \tilde{T}(\mu))\,\mu\| \leq \|T(\mu) - \tilde{T}(\mu)\|_F \cdot \|\mu\| \leq \frac{1}{2} C_T \|\mu - \mu_0\|^2 \|\mu\|$.

Part (c) combines (a) with Theorem 11(a): the $\Theta(r d_k^2)$ MZ parameters are both sufficient and necessary (up to log factors) for ϵ-approximation of the linearized operator family. □

Remark 1 (Interpretation)**.** *Proposition 13 establishes that $\mathcal{F}_L$ is to context-adaptive operators what linear models are to nonlinear regression: the first-order approximation class. The Jacobian rank r determines the intrinsic dimensionality of how the operator varies locally, and MZ-attention with $L = r$ generators is the minimax-optimal parameterization for this class. A standard attention mechanism, which has zero relational capacity by Theorem 1(a), corresponds to the zeroth-order approximation $T \equiv T(\mu_0)$, discarding all first-order variation.*

TDOF of soft Chamfer matching.

Proposition 14 (Soft Chamfer matching has nontrivial TDOF)**.** *Consider the soft nearest-neighbor operator $\mathrm{nn}_\beta(a, B) = \sum_{j=1}^m w_j(a, B)\, b_j$ with softmax weights $w_j = \mathrm{softmax}_{j'}(-\beta \|a - b_{j'}\|^2)$, and define the residual operator $T_a(B) = I_d - P_a(B)$ where $P_a(B)\, a = \mathrm{nn}_\beta(a, B)$. Then:*

(a) The operator $T_a(B)$ is context-adaptive: for any $m \geq 2$, there exist configurations B, B' such that $T_a(B) \neq T_a(B')$.

(b) $\mathrm{TDOF}(T_a) \geq d$ when $m \geq 2$ and $\beta > 0$: as B varies over $(\mathbb{R}^d)^m$, the operator $T_a(B)$ spans a subspace of $\mathbb{R}^{d \times d}$ of dimension at least d.

Proof. (a) Immediate: different B configurations change the matching weights w_j, hence $P_a(B)$.

(b) Fix $a \in \mathbb{R}^d$ and consider $B = \{0, t\, e_k\}$ for each standard basis vector e_k, $k = 1, \ldots, d$. The soft nearest-neighbor is $\mathrm{nn}_\beta(a, B) = w_2(t, k)\, t\, e_k$ where $w_2 = (1 + \exp(-\beta(\|a\|^2 - \|a - te_k\|^2)))^{-1}$. The Jacobian $\partial \mathrm{nn}_\beta / \partial t\big|_{t=0}$ has a nonzero component in the e_k direction. Since this holds independently for each k, the operator $P_a(B)$ varies in at least d independent directions as B ranges over $(\mathbb{R}^d)^m$. By Definition 4, $\mathrm{TDOF}(T_a) \geq d$. □

Remark 2. *For the full Chamfer distance $\mathrm{CD}(A, B) = \frac{1}{n} \sum_i \|a_i - \mathrm{nn}(a_i, B)\|^2$, each query point a_i contributes d independent relational directions. In the hard-matching limit $\beta \to \infty$, distinct matchings correspond to combinatorially many regions in configuration space, each defining a different piecewise-linear operator, well beyond the reach of a single context-rigid transformation. This explains the large gap between MZAttn and standard attention on set matching (Task B in Table 2): the Chamfer distance target requires a context-adaptive operator, and the standard Set Transformer's fixed W_O cannot represent the matching-dependent structure.*

Lemma 15 (Compositional orthogonality under the worst-case construction)**.** *Fix $k \leq d - 1$ and the explicit witness $f \in \mathcal{F}_k$ defined by $\sigma_l(\mu) := \tanh(\mu_l)$, $M_l := e_l\, e_d^\top \in \mathbb{R}^{d \times d}$ for $l = 1, \ldots, k$, and $M_0 := 0$ (the constant-direction term is set to zero so that the network's linear path provides no free target-direction contribution; this is without loss of generality for the lower-bound argument). Let $\mu \in \mathbb{R}^d$ have i.i.d. standard-Gaussian coordinates. Then:*

(a) $\langle M_l, M_{l'} \rangle_F = \delta_{ll'}$, *so* $\{M_l\}_{l=1}^k$ *is Frobenius-orthonormal.*

(b) For every $r \geq 2$ *and every* $l_1, \ldots, l_r \in \{1, \ldots, k\}$, $M_{l_1} M_{l_2} \cdots M_{l_r} = 0$.

(c) Consequently, in the tensor-product space $L^2(\mu) \otimes \mathbb{R}^{d \times d}$ *with inner product* $\langle a \otimes A, b \otimes B \rangle := \mathbb{E}_\mu[ab] \langle A, B \rangle_F$, *every monomial composition term of the form* $P(\sigma) \otimes (M_{l_1} \cdots M_{l_r})$ *with composition order* $r \geq 2$ *vanishes identically (the matrix factor is zero), and therefore has zero projection on the target subspace* $\mathcal{V} = \operatorname{span}\{\sigma_l \otimes M_l\}_{l=1}^k$.

Proof. **(a)** $\langle e_l e_d^\top, e_{l'} e_d^\top \rangle_F = (e_l^\top e_{l'})(e_d^\top e_d) = \delta_{ll'} \cdot 1 = \delta_{ll'}$.

(b) For any product of length $r \geq 2$, $M_{l_1} M_{l_2} = (e_{l_1} e_d^\top)(e_{l_2} e_d^\top) = e_{l_1} (e_d^\top e_{l_2}) e_d^\top$. Since $l_2 \in \{1, \ldots, k\} \subseteq \{1, \ldots, d-1\}$ and $d \neq l_2$, the inner factor $e_d^\top e_{l_2} = 0$, so $M_{l_1} M_{l_2} = 0$. Any longer product factors through this length-2 block and is therefore also 0.

(c) For any monomial $P(\sigma)$ in $\{\sigma_l\}$ and any composition order $r \geq 2$, the matrix factor $M_{l_1} \cdots M_{l_r} = 0$ by (b). Hence $P(\sigma) \otimes (M_{l_1} \cdots M_{l_r}) = 0$ in the tensor-product space, and its projection on $\mathcal{V}$ (or on any subspace) is 0. □

Remark 3. *The construction* $M_l = e_l e_d^\top$ *is the simplest worst-case witness: each* M_l *is a rank-1 outer product whose left factor* e_l *varies with* l *but whose right factor* e_d *is a single fixed direction not used as a target index. This makes the matrix family closed under no products of order* ≥ 2 *(every such product collapses to* 0*), so a* D*-layer attention network cannot bootstrap target directions via composition. Other Frobenius-orthonormal* $\{M_l\}$ *admit non-zero matrix products and would require a separate analysis; we use the simplest construction for which the depth lower bound is unconditional.*

Generator mixing characterization.

Proposition 16 (Generator mixing)**.** *For mixed generators* $\tilde{g}_k = g_k + \sum_l w_{kl} g'_l$*: (a)* $\tilde{Z} \subseteq Z$ *if* $\sum_k |w_{kl}| \leq 1$ *for all* l*; (b)* $Z \subseteq \tilde{Z}$ *if* $\sum_k |w_{kl}| \geq 1$*; (c) equality iff* $\sum_k |w_{kl}| = 1$*. Initializing mixing weights with* ℓ_1 *column norms close to 1 yields near-exact generator representation.*

Score perturbation bound.

Proposition 17 (Score perturbation)**.** *For input perturbation* $\|\delta_j\| \leq \rho$*, the total output decomposes as* $\tilde{o} - o = M_c W_V \hat{G} \xi + \mathcal{E}$ *with* $\|\mathcal{E}\| \leq C\rho \max_j \|v_j\|$*, where* $C = \|M_c\|_2 \|W_Q s\|_2 \|W_K\|_2 / \sqrt{d_k}$.

B.4 Full Proofs

B.4.1 Proof of Theorem 1 (Relational Capacity Bottleneck)

Proof. **Part (a).** An H-head standard attention layer without FFN computes, for query q (e.g., a PMA seed):

$$o(S) = \sum_{h=1}^{H} W_o^{(h)} \hat{v}^{(h)}, \quad \hat{v}^{(h)} = \sum_j \alpha_j^{(h)}(S)\, W_V^{(h)} x_j.$$

The aggregated value $\hat{v}^{(h)}$ depends on S through the scores $\alpha_j^{(h)}$, but the post-aggregation transformation $W_o^{(h)}$ is a fixed matrix. The effective output transformation is:

$$o(S) = \sum_h W_o^{(h)} W_V^{(h)} \Big(\sum_j \alpha_j^{(h)}(S)\, x_j \Big).$$

On constant sets $S_t = \{te_i, \ldots, te_i\}$, softmax symmetry forces $\alpha_j^{(h)} = 1/n$ for all j, h, giving $o(S_t) = (\sum_h W_o^{(h)} W_V^{(h)})\, te_i$. The effective transformation $T_{\text{std}} = \sum_h W_o^{(h)} W_V^{(h)}$ is independent of S_t, hence relational capacity is zero.

Part (b). After attention, the FFN receives $h = x + \hat{v}$ and must produce output $\sum_{l=1}^{k} \sigma_l(h) \cdot M_l h$ to represent a function with relational complexity k. Each term $\sigma_l(h) \cdot (M_l h)_j$ is the product of a bounded nonlinear scalar and a linear form. By a standard region-counting argument (whose matching upper-bound construction is given by Yarotsky [46], Proposition 3), a depth-2 ReLU network approximating $f(a, b) = ab$ on a compact domain to accuracy ϵ requires $\Omega(1/\sqrt{\epsilon})$ neurons; the detailed argument is given in the proof of Theorem 2 (Part (b), Appendix). For each relational direction l, the FFN must implement d_k such products (one per output coordinate), requiring $\Omega(d_k)$ neurons per direction. With d_{ff} total hidden units, at most $\lfloor d_{\text{ff}}/d_k \rfloor$ independent relational directions can be represented.

Part (c). Direct construction: set $M_c = M_0$, use L generator matrices $M_1, \ldots, M_L$, and gates $\gamma_l = \sigma_l(\mu(S))$. Then $M(\mu(S)) = M_0 + \sum_{l=1}^{L} \sigma_l(\mu(S)) M_l = T(S)$, which exactly represents any $f \in \mathcal{F}_L$. Parameter count: $(L+1)$ matrices of size $d_k \times d_k$ plus L gate weight vectors of size d_k.

Part (d). Let $f(S) = (M_0 + \sum_{l=1}^{k} \sigma_l(\mu(S)) M_l)\, \mu(S)$ with $\{M_l\}_{l=1}^{k}$ orthogonal in Frobenius norm ($\langle M_l, M_{l'} \rangle_F = 0$ for $l \neq l'$). Any mechanism with relational capacity R can represent an output transformation of the form $T_\theta(S) = M_0' + \sum_{l=1}^{R} \tau_l(S) N_l$ for some learned $\{N_l\}_{l=1}^{R}$ and scalar functions τ_l.

The optimal R-capacity approximation chooses $M_0' = M_0$ and $\{N_l\}_{l=1}^{R}$ to be the projection of $\{M_l\}_{l=1}^{k}$ onto the best R-dimensional subspace (in the Frobenius metric, weighted by $\mathbb{E}[\sigma_l^2 \|\mu\|^2]$). By the Eckart–Young–Mirsky theorem applied to the operator-valued approximation, the residual error is:

$$\mathbb{E}_S\big[\|f(S) - o_\theta(S)\|^2\big] \geq \sum_{l=R+1}^{k} \mathbb{E}_S\big[\sigma_l(\mu(S))^2 \cdot \|M_l \mu(S)\|^2\big],$$

where we order the terms by $\mathbb{E}[\sigma_l^2 \|M_l \mu\|^2]$ in decreasing order and the mechanism captures the top R. The bound is tight: it is achieved when $\{N_l\}$ aligns with the top-R relational directions. □

B.4.2 Proof of Theorem 11 (Minimax Parameter Efficiency)

Proof. **Part (a).** The MZ-attention head has parameters $M_c \in \mathbb{R}^{d_k \times d_k}$, $M_1, \ldots, M_L \in \mathbb{R}^{d_k \times d_k}$, and gate weights $w_1, \ldots, w_L \in \mathbb{R}^{d_k}$, totalling $(L+1)d_k^2 + Ld_k = \Theta(Ld_k^2)$. By Theorem 10(a), this realises every $f \in \mathcal{F}_{L,B}$ exactly.

Part (b). We use a metric entropy argument. The function class $\mathcal{F}_{L,B}$ is parameterized by $M_0, M_1, \ldots, M_L \in \mathbb{R}^{d_k \times d_k}$ with $\|M_l\|_F \leq B$, and Lipschitz gates σ_l. We lower-bound the packing number by considering the subclass with fixed $M_0 = I$, fixed $\sigma_l = \text{id}$ (identity, clipped to $[-1, 1]$), and varying $M_1, \ldots, M_L$.

For two functions f, f' with generator matrices $\{M_l\}$ and $\{M_l'\}$:

$$\sup_{\|x\| \leq R} |f(S_x) - f'(S_x)| \geq \sup_{\|x\| \leq R} \|\sum_l \sigma_l(x)(M_l - M_l')x\|.$$

For the specific input $S_x = \{x, \ldots, x\}$ with $\mu = x$, and choosing $x = Re_j/\|e_j\|$ for the coordinate maximizing $\|(M_l - M_l')e_j\|$:

$$\|f - f'\|_\infty \geq c \cdot R \cdot \max_l \|M_l - M_l'\|_F$$

for a universal constant $c > 0$ depending only on the dimension.

The δ-packing number of $\{M \in \mathbb{R}^{d_k \times d_k} : \|M\|_F \le B\}$ in Frobenius norm is at least $(B/\delta)^{d_k^2}$ (volume comparison in $\mathbb{R}^{d_k^2}$). With L independent generator matrices, the packing number of $\mathcal{F}_{L,B}$ at resolution ϵ is at least:

$$\mathcal{M}(\epsilon, \mathcal{F}_{L,B}) \ge \big(cBR/\epsilon\big)^{Ld_k^2}.$$

Any model class parameterized by $\theta \in \mathbb{R}^P$ that ϵ-approximates all of $\mathcal{F}_{L,B}$ must distinguish at least $\mathcal{M}(\epsilon)$ functions. By a standard argument [8, 45], the parameter space $\mathbb{R}^P$ can cover at most $\exp(CP\log(1/\epsilon))$ ϵ-cells (where C depends on the model's Lipschitz properties in θ). Matching: $P\log(1/\epsilon) \ge \Omega(Ld_k^2 \log(BR/\epsilon))$, giving $P \ge \Omega(Ld_k^2 \log(BR/\epsilon)/\log(1/\epsilon))$. Since $\log(BR/\epsilon)/\log(1/\epsilon) \ge 1$ for $BR \ge 1$, we obtain $P \ge \Omega(Ld_k^2)$. Including the logarithmic precision factor: $P \ge \Omega(Ld_k^2 \log(BR/\epsilon))$.

Part (c). MZ-attention uses $P_{\text{MZ}} = (L+1)d_k^2 + Ld_k = \Theta(Ld_k^2)$ parameters and achieves $\epsilon = 0$ (exact representation). The lower bound requires $P \ge \Omega(Ld_k^2)$ for any constant ϵ. Hence MZ-attention is order-optimal up to logarithmic factors. □

B.4.3 Proof of Theorem 8 (Expressiveness Separation)

We restate and prove both parts in full detail; the construction matches the inline proof in Section B.3 and uses the signed-generator readout variant of Section 2.4.

Proof of part (a). We configure a single MZ-attention head as follows. Set the query projection $W_Q = 0 \in \mathbb{R}^{d_k \times d}$, so $\alpha_{ij} = \text{softmax}_j(0) = 1/n$ for all i, j, yielding uniform aggregation $\hat{c}_i = \frac{1}{n}\sum_j W_V c_j$. With $W_V = I_d$, $\hat{c}_i = \mu(S)$.

Set $M_c = I_{d_k}$, $L = 1$, $M_1 = A$, gate weight $w_\gamma = e_1$, and mixing matrix $W_{\text{mix}} = e_1 \in \mathbb{R}^{n_g \times 1}$ (use only the first generator slot). The context-dependent gate is $\gamma_1(\mu) = \tanh(e_1^\top \mu) = \tanh(\mu_1)$. By Eq. (3), the per-token output has center $o^c = M_c \hat{c} = \mu(S)$ and generator stream whose first row is $o^G[1,:] = \tanh(\mu_1)\, A\, \mu(S)$ with the remaining rows zero.

After PMA pooling (uniform aggregation by symmetry), the pooled center is $\mu(S)$ and the pooled signed first-generator is $\tanh(\mu_1)\, A\, \mu(S)$. Applying the signed-generator readout $W_{\text{out}} : (c, g_1) \mapsto c + g_1$ yields $\mu(S) + \tanh(\mu_1)\, A\, \mu(S) = (I + \tanh(\mu_1)\, A)\, \mu(S) = f_A(S)$, exactly. Every operation in the construction is either a fixed linear map or the native MZ gate, both computed without finite-width truncation; no UAT-based ReLU approximation is invoked. □

Proof of part (b). An H-head standard attention layer without FFN computes, for a query q (e.g., a PMA seed vector s):

$$o(S) = W_o \,\text{concat}\big(W_V^{(1)}\hat{v}^{(1)}, \ldots, W_V^{(H)}\hat{v}^{(H)}\big)$$

where $\hat{v}^{(h)} = \sum_j \alpha_j^{(h)}(S)\, x_j$ and $\alpha_j^{(h)} = \text{softmax}_j(s^{(h)\top} W_Q^{(h)\top} W_K^{(h)} x_j/\sqrt{d_k})$.

Step 1. Consider the subfamily of constant sets $S_t = \{te_i, \ldots, te_i\}$ (n copies of te_i) for a basis vector e_i and scalar $t \in \mathbb{R}$. By symmetry of the softmax with identical inputs, $\alpha_j^{(h)} = 1/n$ for all j and h. Hence $\hat{v}^{(h)} = W_V^{(h)}(te_i) = tW_V^{(h)} e_i$, and:

$$o(S_t) = W_o \,\text{concat}(tW_V^{(1)}e_i, \ldots, tW_V^{(H)}e_i) = t \cdot W_o \,\text{concat}(W_V^{(1)}e_i, \ldots, W_V^{(H)}e_i) =: t \cdot \tilde{w}_i.$$

This is a linear function of t.

Step 2. On the same inputs, $f_A(S_t) = (I + \tanh(t\delta_{i1})A)(te_i) = te_i + t\tanh(t\delta_{i1})Ae_i$, where $\delta_{i1} = [e_i]_1$. For $i = 1$: $f_A(S_t) = te_1 + t\tanh(t)Ae_1$.

If $Ae_1 \ne 0$, the map $t \mapsto te_1 + t\tanh(t)Ae_1$ is nonlinear (since $t\tanh(t)$ is nonlinear and $Ae_1 \ne 0$ ensures this nonlinearity survives in at least one coordinate). A linear function $t \mapsto t\tilde{w}_1$ cannot agree with this nonlinear function on any open interval.

Step 3. If $Ae_1 = 0$, since $A \neq 0$ there exists e_j with $Ae_j \neq 0$. If $j = 1$, we already have a contradiction. If $j \neq 1$, use $S_t = \{te_j + \epsilon e_1, \ldots, te_j + \epsilon e_1\}$ for small $\epsilon > 0$: $\mu(S_t) = te_j + \epsilon e_1$, so $\tanh(\mu_1) = \tanh(\epsilon) \neq 0$, and $A\mu = tAe_j + \epsilon Ae_1$. Since $Ae_j \neq 0$, the nonlinearity persists. □

B.4.4 Proof of Theorem 10 (Width Efficiency)

Proof. Part (a) is immediate: the MZ parameters $M_c \in \mathbb{R}^{d_k \times d_k}$, $M_1, \ldots, M_L \in \mathbb{R}^{d_k \times d_k}$, and gate weights $w_1, \ldots, w_L \in \mathbb{R}^{d_k}$ total $(L+1)d_k^2 + Ld_k$.

For part (b), consider a single bilinear term $\sigma_l(h) \cdot (M_l h)_j$ that the FFN must compute, with $a = \sigma_l(h) \in [-1, 1]$ and $b = (M_l h)_j \in [-B_l, B_l]$, $B_l = \|M_l\|_2 R$. The FFN must approximate $f(a, b) = ab$ on this rectangle to sup-norm error ϵ.

By a standard region-counting argument for shallow ReLU approximation of strictly-convex bilinear functions (the matching upper-bound construction is Yarotsky [46], Proposition 3), any depth-2 ReLU network approximating ab on $[-1, 1] \times [-B_l, B_l]$ to sup-norm error ϵ requires $n = \Omega(B_l/\sqrt{\epsilon})$ hidden units. The standard underlying argument bounds a depth-2 ReLU network's piecewise-linear regions by $O(n^2)$ on $\mathbb{R}^2$ and applies the strict convexity of ab ($\partial^2 f/\partial a \partial b = 1$): on a region of diameter r the best affine approximant errs by $\Omega(r^2)$, and an $O(n^2)$-region partition of a $[-B_l, B_l] \times [-1, 1]$ rectangle has at least one region of diameter $\Omega(B_l/n)$, giving worst-case error $\Omega(B_l^2/n^2)$ and hence $n = \Omega(B_l/\sqrt{\epsilon})$.

Summing over L generators and d_k output coordinates: $d_{\text{ff}} \geq \sum_l d_k \cdot \Omega(B_l/\sqrt{\epsilon}) = \Omega(Ld_k B/\sqrt{\epsilon})$ where $B = \max_l B_l$. For unit-norm inputs ($B = O(1)$): $d_{\text{ff}} = \Omega(Ld_k/\sqrt{\epsilon})$. □

B.4.5 Proof of Theorem 3 (Directional Sensitivity)

Proof. (a) Standard attention output: $o = W_o\hat{v}$ with $\hat{v} = \sum_j \alpha_{ij} v_j$. Under perturbation $\tilde{\hat{v}} = \hat{v} + \hat{G}\xi$: $\|\Delta o\| = \|W_o\hat{G}\xi\| \leq \|W_o\|_2 \|\hat{G}\xi\| \leq \|W_o\|_2 \sum_k \|\hat{G}_k\| \cdot |\xi_k| \leq \|W_o\|_2 \|\hat{G}\|_{1,2}$. This is a scalar (norm) bound: all directions in $\mathbb{R}^{d_k}$ are treated equally.

(b) MZ-attention output: $\Delta o = M_c\hat{G}\xi$. As ξ ranges over $[-1, 1]^{n_g}$, Δo traces the zonotope $\mathcal{Z}_\Delta = \{M_c\hat{G}\xi : \|\xi\|_\infty \leq 1\}$ with generators $\{M_c\hat{G}_k\}_{k=1}^{n_g}$. The vertex count $2\binom{n_g}{d_k - 1}$ is a standard result from zonotope geometry [3].

(c) Every point in $\mathcal{Z}_\Delta$ satisfies $\|z\| \leq \sum_k \|M_c\hat{G}_k\|_2 \leq \|M_c\|_2 \|\hat{G}\|_{1,2}$, so $\mathcal{Z}_\Delta \subseteq B(\|M_c\|_2 \|\hat{G}\|_{1,2})$. For directions $u \in \mathbb{R}^{d_k}$ orthogonal to the column span $\text{span}\{M_c\hat{G}_k\}_{k=1}^{n_g}$, the support function $h_{\mathcal{Z}_\Delta}(u) = \sum_k |u^\top M_c\hat{G}_k| = 0 < \|M_c\|_2 \|\hat{G}\|_{1,2} = h_B(u)$, so the MZ certificate is strictly tighter than the ball whenever the column span is a proper subspace. The volume of a zonotope with generators $g_1, \ldots, g_{n_g} \in \mathbb{R}^{d_k}$ is [3]

$$\text{vol}(\mathcal{Z}_\Delta) = 2^{d_k} \sum_{|I| = d_k} |\det([g_i]_{i \in I})|,$$

which is strictly smaller than $\text{vol}(B)$ for $d_k \geq 2$ since the zonotope is a proper inscribed subset of the ball; for typical trained generators with unequal norms and non-orthogonal directions, $\text{vol}(\mathcal{Z}_\Delta)$ is empirically several orders of magnitude smaller than $\text{vol}(B)$, reflecting the anisotropic structure of the model's sensitivity. □

B.4.6 Proof of Proposition 16 (Generator Mixing)

Proof. Let Z_{n_g+L} have generators $\{g_1, \ldots, g_{n_g}, g'_1, \ldots, g'_L\}$ in general position. A point in Z_{n_g+L} is $p = c + \sum_{k=1}^{n_g} g_k\zeta_k + \sum_{l=1}^{L} g'_l\zeta'_l$ with $|\zeta_k|, |\zeta'_l| \leq 1$. A point in $\tilde{Z}_{n_g}$ is $\tilde{p} = c + \sum_{k=1}^{n_g}(g_k + \sum_l w_{kl} g'_l)\tilde{\zeta}_k$ with $|\tilde{\zeta}_k| \leq 1$.

Expanding: $\tilde{p} = c + \sum_k g_k\tilde{\zeta}_k + \sum_l (\sum_k w_{kl}\tilde{\zeta}_k) g'_l$.

(a) $\tilde{Z}_{n_g} \subseteq Z_{n_g+L}$: Set $\zeta_k = \tilde{\zeta}_k$ and $\zeta'_l = \sum_k w_{kl}\tilde{\zeta}_k$. We need $|\zeta'_l| = |\sum_k w_{kl}\tilde{\zeta}_k| \le \sum_k |w_{kl}| \cdot |\tilde{\zeta}_k| \le \sum_k |w_{kl}|$. If $\sum_k |w_{kl}| \le 1$ for all l, then $|\zeta'_l| \le 1$ and $\tilde{p} \in Z_{n_g+L}$.

(b) $Z_{n_g+L} \subseteq \tilde{Z}_{n_g}$: Given a point $p \in Z_{n_g+L}$ with parameters (ζ, ζ'), we seek $\tilde{\zeta}$ with $|\tilde{\zeta}_k| \le 1$ matching both g_k and g'_l coefficients. By general position, coefficient matching requires $\tilde{\zeta}_k = \zeta_k$ (from g_k) and $\sum_k w_{kl}\zeta_k = \zeta'_l$ (from g'_l). For the latter to hold for all $\zeta'_l \in [-1, 1]$, the range of $\sum_k w_{kl}\zeta_k$ (over $|\zeta_k| \le 1$) must contain $[-1, 1]$. This range is $[-\sum_k |w_{kl}|, \sum_k |w_{kl}|]$, so containment requires $\sum_k |w_{kl}| \ge 1$.

(c) Combining (a) and (b): equality holds when $\sum_k |w_{kl}| = 1$ for all l. □

B.4.7 Proof of Theorem 4 (IHW Uncertainty Bound)

Proof. **Part (a).** The output zonotope is $Z_{\text{out}} = \{c_{\text{out}} + G_{\text{out}}\xi : \|\xi\|_\infty \le 1\}$. Under a linear readout $W_{\text{out}} \in \mathbb{R}^{d_k}$, the output is:

$$y(\xi) = W_{\text{out}}^\top(c_{\text{out}} + G_{\text{out}}\xi) = W_{\text{out}}^\top c_{\text{out}} + \sum_{g=1}^{n_g}(W_{\text{out}}^\top G_g)\xi_g.$$

This is a scalar zonotope (interval) with center $W_{\text{out}}^\top c_{\text{out}}$ and half-width $\sum_g |W_{\text{out}}^\top G_g|$. Hence:

$$\text{diam}(\mathcal{Y}) = 2\sum_{g=1}^{n_g} |W_{\text{out}}^\top G_g| \le 2\sum_g \|W_{\text{out}}\|_2 \|G_g\|_2 \le 2\|W_{\text{out}}\|_2 \sum_g \|G_g\|_1 = 2\|W_{\text{out}}\|_2 \cdot \text{IHW}.$$

Part (b). Since $\xi_1, \ldots, \xi_{n_g}$ are i.i.d. uniform on $[-1, 1]$: $\mathbb{E}[\xi_g] = 0$, $\mathbb{E}[\xi_g^2] = 1/3$, and $\mathbb{E}[\xi_g \xi_{g'}] = 0$ for $g \ne g'$. The variance of $y(\xi) = W_{\text{out}}^\top c_{\text{out}} + \sum_g (W_{\text{out}}^\top G_g)\xi_g$ is:

$$\text{Var}[y] = \sum_{g=1}^{n_g}(W_{\text{out}}^\top G_g)^2 \cdot \text{Var}[\xi_g] = \frac{1}{3}\sum_g (W_{\text{out}}^\top G_g)^2.$$

By the Cauchy–Schwarz inequality applied coordinate-wise: $(W_{\text{out}}^\top G_g)^2 \le \|W_{\text{out}}\|_2^2 \|G_g\|_2^2 \le \|W_{\text{out}}\|_2^2 \|G_g\|_1^2$. Hence $\sum_g (W_{\text{out}}^\top G_g)^2 \le \|W_{\text{out}}\|_2^2 \sum_g \|G_g\|_1^2 \le \|W_{\text{out}}\|_2^2 (\sum_g \|G_g\|_1)^2 = \|W_{\text{out}}\|_2^2 \cdot \text{IHW}^2$, where the last inequality uses $\sum a_i^2 \le (\sum a_i)^2$ for $a_i \ge 0$.

Part (c). From part (a), $\text{diam}(\mathcal{Y}(S)) = 2\|W_{\text{out}}\|_2 \cdot \text{IHW}(Z_{\text{out}}(S))$. Since $\|W_{\text{out}}\|_2$ is shared between inputs (it is a model parameter, not input-dependent), the monotonicity in IHW directly implies monotonicity in the reachable diameter. □

B.4.8 Proof of Theorem 2 (Depth Separation)

Proof. We prove the depth lower bound on the input distribution where $\mu(S) \in \mathbb{R}^d$ has i.i.d. standard-Gaussian coordinates, using a compositional-orthogonality argument in $L^2(\mu)$ together with the explicit worst-case construction of Lemma 15.

Setup. Let $f(S) = (\sum_{l=1}^k \sigma_l(\mu(S))\, M_l)\, \mu(S)$ be the witness of Lemma 15: $\sigma_l(\mu) := \tanh(\mu_l)$, $M_l := e_l e_d^\top$ for $l = 1, \ldots, k$ with $k \le d - 1$, and $M_0 = 0$. Each M_l is rank-1 with Frobenius-orthonormal $\{M_l\}_{l=1}^k$ (Lemma 15(a)). The choice $M_0 = 0$ ensures the network's linear path $\mu \mapsto M_0\mu$ contributes no free target-direction component on $\mathcal{V}$, so the depth lower bound is determined by the FFN multiplicative budget alone. We view f as an element of the tensor-product space $L^2(\mu) \otimes \mathbb{R}^{d\times d}$ with inner product $\langle a \otimes A, b \otimes B\rangle := \mathbb{E}_\mu[a\, b]\, \langle A, B\rangle_F$. The target directions $\{\sigma_l \otimes M_l\}_{l=1}^k$ are mutually orthogonal in this inner product (independent μ_l's make $\mathbb{E}[\sigma_l \sigma_{l'}] = 0$ for $l \ne l'$, and Frobenius orthonormality handles the matrix factor), so they span a k-dimensional subspace $\mathcal{V}$.

Layer-by-layer FFN bilinear count. A D-layer standard attention network with residual connections and depth-2 ReLU FFN of width d_{ff} computes

$$z^{(\ell+1)} = z^{(\ell)} + \text{FFN}^{(\ell)}\big(z^{(\ell)} + \hat{v}^{(\ell)}\big), \quad z^{(0)} = \mu.$$

A bilinear term in $\mathcal{V}$ has the form $\sigma_l(\mu) \otimes M_l$, requiring the FFN to multiply a gate σ_l by the linear coordinate map $\mu \mapsto M_l\mu$. By Theorem 1(b), each multiplicative bilinear coordinate of this kind requires $\Omega(d_k)$ ReLU neurons (Yarotsky bilinear lower bound), so a single layer's FFN of width d_{ff} contributes at most $\lfloor d_{\text{ff}}/d_k \rfloor$ new directions that lie in $\mathcal{V}$.

Cross-layer compositions vanish on the matrix side. After ℓ layers, the residual representation can additionally contain monomial composition terms of the form $P(\sigma) \otimes (M_{l_1} \cdots M_{l_r})$ where $r \geq 2$ comes from successive FFNs reading earlier-layer bilinear outputs and re-multiplying them through linear maps. By Lemma 15(b)–(c), under the construction $M_l = e_l e_d^\top$ every such matrix product with $r \geq 2$ is identically zero, so the entire term vanishes in the tensor-product space and therefore has zero projection on $\mathcal{V}$. Hence cross-layer composition cannot bootstrap any new target direction; only the layer-wise FFN bilinear count contributes.

Counting. After D layers the network has captured at most $\Lambda_D \leq D \cdot \lfloor d_{\text{ff}}/d_k \rfloor$ of the k target directions in $\mathcal{V}$. The remaining $k - \Lambda_D$ directions contribute an irreducible $L^2(\mu)$ error (by the Eckart–Young residual of Theorem 1(d)) bounded below by $\Omega((k - \Lambda_D)/k) \cdot \text{Var}(f)$. For exact representation we need $\Lambda_D = k$, giving $D \geq k\, d_k/d_{\text{ff}}$. For approximate representation in $L^2(\mu)$ with residual at most $O(\text{Var}(f)/L)$, achieving this residual requires $k - \Lambda_D = O(k/L)$ uncaptured directions, hence $\Lambda_D \geq k(1 - O(1/L))$, giving $D \geq \Omega(k\, d_k/d_{\text{ff}})$ up to constants since each missing direction contributes $\Theta(\text{Var}(f)/k)$.

Single MZ-attention layer with signed-generator readout. Setting $M_c = M_0 = 0$ and using the k generators $M_1, \ldots, M_k$ of f with gates $\gamma_l(\mu) = \sigma_l(\mu)$, Eq. (3) produces a per-token output whose center is 0 and whose generator stream contains $\gamma_l(\mu)\, M_l\, \mu$ as the l-th signed generator slot. After PMA pooling (uniform aggregation, as in the construction of Theorem 8), the signed-generator readout $W_{\text{out}} : (c, g_1, \ldots, g_k) \mapsto \sum_{l=1}^{k} g_l$ recovers $f(S) = \sum_l \sigma_l(\mu) M_l \mu$ exactly. The IHW readout (Section 2.4) gives only $\sum_l |\sigma_l(\mu)| \cdot \|M_l\mu\|$-bounded scalar information rather than the signed combination, so it does not extract $f(S)$ exactly; the expressiveness statement is for the signed-readout variant. □

B.4.9 Proof of Proposition 7 (Permutation Equivariance)

Proof. Let π be a permutation of $\{1, \ldots, n\}$ and denote the permuted set as $\pi(S) = \{x_{\pi(1)}, \ldots, x_{\pi(n)}\}$.

MZ-MAB equivariance: The attention scores satisfy $\alpha_{i,\pi(j)}(\pi(S)) = \alpha_{ij}(S)$ because they depend on pairwise dot products $q_i^\top k_j$ which are permutation-equivariant. The aggregation $\hat{c}_i = \sum_j \alpha_{ij} v_j$ is thus equivariant: permuting the set permutes the query outputs. The MZ transform (Eq. 3) operates on each query independently. The gate $\gamma = \tanh(W_\gamma \bar{c})$ depends on the set mean $\bar{c} = \frac{1}{n}\sum_j c_j$, which is permutation-invariant. Hence MZ-MAB is equivariant: $\text{MZ-MAB}(\pi(Q), \pi(KV)) = \pi(\text{MZ-MAB}(Q, KV))$.

MZ-SAB inherits equivariance since $Q = KV = S$.

Permutation invariance of the full model: MZ-PMA uses learnable seed vectors $s_1, \ldots, s_k$ as queries, which are independent of the input set. Thus $\text{MZ-PMA}(\pi(S)) = \text{MZ-PMA}(S)$, and the output is permutation-invariant. □

C 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 [6, 7, 15, 24], where they are computed post-hoc around a trained model's activations to certify robustness, and in control for data-driven reachability [1, 2, 18]. To our knowledge MZAttn is the first architecture to make the matrix-zonotope an internal, trainable representation of the value-projection family: center and generator matrices are jointly optimised by gradient descent, and the zonotope structure participates in every forward pass rather than being wrapped around a frozen model. Concurrent and prior work has explored related ideas in adjacent spaces (e.g. structured low-rank reparameterisations and gated value updates), but we are not aware of an internal trainable matrix-zonotope value family with the gating, mixing, and budget-control structure of Eq. (3).

Set-input networks. Deep Sets [47], PointNet [28], PointNet++ [29], and Set Transformer [21] established permutation invariance and attention-based pooling as standard tools for set inputs; PointNet++ in particular adds hierarchical local-region feature aggregation but still composes fixed per-point MLPs with set pooling, so its value transformation remains context-rigid in our sense. All of these architectures express relationships between elements via scalar scores (softmax weights or learned aggregations) that multiply fixed linear projections of inputs; MZAttn generalizes this to matrix-valued context-adaptive operators, which is precisely what high-TDOF tasks require.

Context-adaptive linear maps. Hypernetworks [12, 34] and dynamic filter networks [14] generate full d_k^2-parameter weight updates per sample, giving unlimited expressiveness but paying a quadratic cost that makes the effective rank hard to control. FiLM [27] restricts to diagonal (scale-and-shift) modulation, which is provably insufficient for off-diagonal structure such as rotations or covariance-dependent maps. Mixture-of-Experts [36] routes between a discrete set of fixed experts and is therefore piecewise-constant in context. The matrix-zonotope parameterization of MZAttn occupies a middle ground: $O(Ld_k^2)$ parameters with L continuous gates realize a convex, rank-controlled family of linear operators that is minimax-optimal for context-adaptive linear maps (Theorem 11) while retaining the geometric and uncertainty structure of the zonotope family.

Equivariant architectures. EGNN [33], SchNet [35], and similar E(n)- or SO(n)-equivariant architectures achieve very strong performance on rotation- and translation-invariant targets by hard-coding the symmetry into the forward pass. On such tasks MZAttn is at a structural disadvantage (it does not bake in the symmetry) and our experiments confirm this. The complementary regime is tasks whose target depends on a specific coordinate frame, e.g. the quadratic TDOF task of Section 3.2, where equivariance is actively harmful: invariant outputs cannot distinguish the target from its rotations. MZAttn is the right tool when the task's symmetry is unknown a priori and cannot be built in by construction.

Uncertainty quantification. MC-Dropout [9] and Deep Ensembles [20] obtain predictive uncertainty by multiple stochastic forward passes; MZAttn's zonotope geometry (Theorem 4) yields a single-pass uncertainty estimate via the interval-hull width of the output zonotope, at 3.7× lower inference cost than the baselines (Appendix B.2, Table 20).

D Additional Experiments

This section reports the full battery of additional experiments referenced from the main body. It is organised into five thematic blocks: **D.1** architectural ablations isolating which design choice drives the gain; **D.2** robustness and generalisation studies; **D.3** uncertainty calibration; **D.4** high-dimensional and large-scale scaling on computational-geometry targets; and **D.5** real-world transfer studies plus the equivariant baseline comparison.

D.1 Architecture and ablation studies

Studies that isolate which architectural choice in MZAttn (generator count, generator-info routing, low-rank update form, depth) drives the empirical advantage. They support the conclusion of Section 5.4: the centre-plus-gated-generator structure, not raw capacity, is what drives the gain.

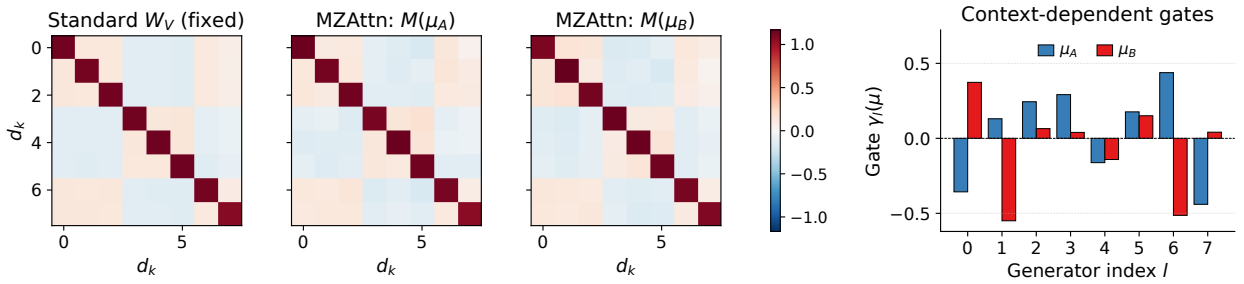

Figure 3: **Effective value-transformation operator (head 0) on the quadratic-TDOF task.** Standard attention uses one fixed W_V regardless of context (left); MZAttn produces a context-specific $M(\mu) = M_c + \sum_l \gamma_l(\mu)\, M_l$ for each input set (centre two panels). The context-dependent gates $\gamma_l(\mu)$ (right) drive the adaptation; the resulting Frobenius-norm change is $\|M(\mu_A) - M(\mu_B)\|_F / \|M_c\|_F = 8.0\%$.

Per-sample operator adaptation is non-trivial. Visualising $M(\mu_A), M(\mu_B)$ on two contexts at head 0 (Figure 3), the relative operator change is $\|M(\mu_A) - M(\mu_B)\|_F / \|M_c\|_F = 8.0\%$ and the eight gates $\gamma_l(\mu)$ flip sign on five of eight generators between contexts, ruling out gate collapse to constants.

Trained generator basis + TDOF utilisation. Probing a trained L=8 checkpoint with 1,000 test contexts (Figure 4, cosine, Frobenius-norm, and gate panels): generators are near-orthogonal in operator space, comparably sized, and gates active on 78.6% of evaluations — no dead generators or constant-bias collapse. Sweeping $L \in \{4, 8, 16, 32\}$ at target TDOF = 8, the soft effective rank of 200 extracted operators saturates the budget when $L \leq$ TDOF and grows sublinearly above (effective-rank panel). The trained model allocates operator-space variation up to but not beyond the intrinsic TDOF, with L a soft upper bound.

D.1.1 MZ Generator Budget

Even $n_{\text{mz}} = 1$ exceeds deep standard Set Transformers.

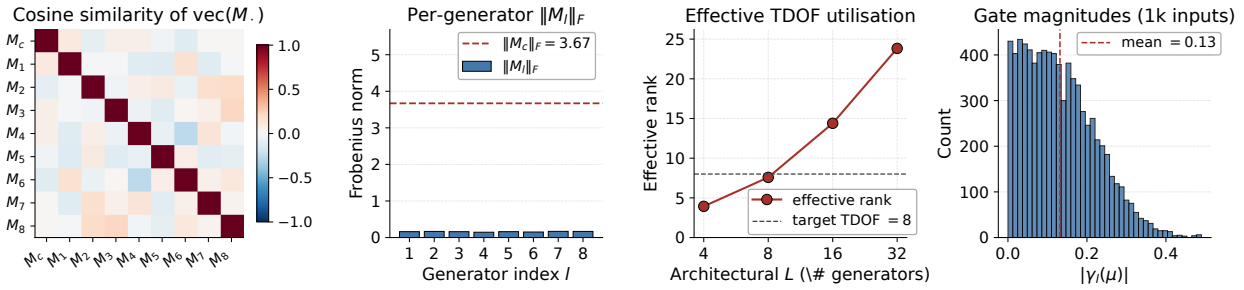

Figure 4: **Mechanistic anatomy of a trained MZAttn layer (L=8, head 0, 1,000 probe contexts), from left to right. Cosine similarity** of vec(M_c) and vec(M_l): generator directions are near-orthogonal (mean off-diagonal $|\rho| < 0.01$). **Per-generator Frobenius norms** cluster in $[0.146, 0.169]$ (CV 5.1%), each $\sim 1/22$ of $\|M_c\|_F$. **Effective rank** of $\{M(\mu_i) - \bar{M}\}_i$ as architectural budget L varies on a fixed TDOF-8 task: saturates the budget when $L \leq$ TDOF, grows sublinearly above. **Gate magnitudes** $|\gamma_l(\mu)|$ over $L \times 1000$ evaluations: 78.6% exceed 0.05, mean 0.13, p95 0.29.

Table 8: Ablation on MZ generator matrices n_{mz} (mean±std, 3 seeds). Per-seed JSONs for all $n_{\text{mz}} \in \{1, 2, 4, 8, 16, 32\}$ are in `gen_ablation.json`.

	$L = 16$		$L = 32$	
n_{mz}	R^2	Params	R^2	Params
1	$\mathbf{0.978 \pm 0.004}$	247K	0.937 ± 0.014	247K
2	0.976 ± 0.012	252K	0.959 ± 0.019	252K
4	0.956 ± 0.011	260K	$\mathbf{0.961 \pm 0.007}$	260K
8	0.960 ± 0.011	278K	0.949 ± 0.021	278K
16	0.948 ± 0.016	313K	0.923 ± 0.011	313K
32	0.943 ± 0.002	383K	0.907 ± 0.010	383K

Practical guidelines for L and n_g. The two main hyperparameters of MZAttn are the number of MZ generator matrices L (controlling relational capacity) and the zonotope generator budget n_g (controlling per-token uncertainty resolution). Based on the ablation studies in Tables 9 and 8, we recommend:

- **MZ generators L:** Start with $L = 2$–4. Even $L = 1$ substantially outperforms standard attention on high-TDOF tasks (Table 8). Increasing L beyond 4 adds parameters (Ld_k^2 per head per layer) with diminishing returns unless the task's TDOF is known to be high. For tasks where the TDOF can be estimated (e.g., d for Chamfer matching, $d(d+1)/2$ for Mahalanobis-type operators), setting L to match the estimated TDOF is theoretically motivated by Theorem 11.

- **Zonotope generators n_g:** Start with $n_g = 4$–8. Table 9 shows that $n_g = 4$ peaks for set matching; larger budgets overfit. Since n_g primarily controls the resolution of the uncertainty envelope rather than the relational capacity, moderate values suffice; the output head aggregates generators via interval hull width regardless of n_g.

- **Scaling regime:** When computational budget is constrained, prioritize L over n_g: the MZ generators directly determine the model's relational capacity (Theorem 1(c)), whereas n_g affects uncertainty granularity, which is less critical for prediction quality.

D.1.2 Generator Ablation

Table 9: Ablation on zonotope generators n_g (set matching, seed 42).

n_g	MSE ↓	MAE ↓	R^2 ↑
2	45.80	3.50	0.967
4	**26.74**	**3.33**	**0.981**
8	32.46	3.39	0.977
16	55.05	4.47	0.961

Performance peaks at $n_g = 4$ and degrades for larger budgets due to overfitting.

D.1.3 Generator Info in Q/K Scoring

Table 10: Ablation: incorporating generator information (IHW) into Q/K scoring. MZ-GenQK adds projections $W_q^{\text{gen}}, W_k^{\text{gen}}$ of per-token interval hull widths into the attention score computation.

	Task A: Set Regr.		Task B: Set Match.		
Model	MSE ↓	R^2 ↑	MSE ↓	R^2 ↑	Params
MZAttn (centers-only Q/K)	2.19$_{\pm 0.09}$	0.879$_{\pm 0.005}$	**55.0**$_{\pm 10.5}$	**0.961**$_{\pm 0.008}$	365K
MZ-GenQK (gen info in Q/K)	**1.64**$_{\pm 0.13}$	**0.909**$_{\pm 0.007}$	68.8$_{\pm 5.6}$	0.951$_{\pm 0.004}$	415K

Incorporating generator information into Q/K scoring yields task-dependent results: on Task A, MZ-GenQK improves R^2 from 0.879 to 0.909 (+0.030 absolute, +3.4% relative), but on Task B it degrades MSE from 55.0 to 68.8 (+25%). The mixed outcome, combined with the additional ∼50K parameters, indicates that centers-only scoring is a reasonable default. The benefit may arise when the prediction depends on generator structure (as in regression), while for matching tasks the center already encodes sufficient information for routing attention.

D.1.4 HyperNet Rank Ablation

Table 11: HyperNet rank ablation (mean ± std over 3 seeds). All models are parameter-matched at ∼360–370K. Increasing the low-rank update dimension does not close the gap to MZAttn; on Task B, performance degrades monotonically.

	Task A: Set Regr.		Task B: Set Match.		
Model	MSE ↓	R^2 ↑	MSE ↓	R^2 ↑	Params
HyperNet (r=8)	3.51$_{\pm 0.01}$	0.806$_{\pm 0.001}$	76.1$_{\pm 5.7}$	0.945$_{\pm 0.004}$	360K
HyperNet (r=16)	3.50$_{\pm 0.02}$	0.807$_{\pm 0.001}$	95.9$_{\pm 8.0}$	0.931$_{\pm 0.006}$	370K
HyperNet (r=32)	3.54$_{\pm 0.02}$	0.805$_{\pm 0.001}$	133.7$_{\pm 8.0}$	0.904$_{\pm 0.006}$	359K
MZAttn	**2.19**$_{\pm 0.09}$	**0.879**$_{\pm 0.005}$	**55.0**$_{\pm 10.5}$	**0.961**$_{\pm 0.008}$	365K

Increasing the HyperNet rank from 8 to 32 does not improve performance: Task A R^2 stays flat at ∼0.806, while Task B MSE increases from 76.1 to 133.7, a 76% degradation. This indicates that the advantage of MZAttn does not stem from having a higher-rank dynamic update, but from the

structured zonotope parameterization: the matrix zonotope's center-plus-generators decomposition, context-dependent gating, and generator mixing provide an inductive bias that generic low-rank hypernetworks cannot replicate regardless of rank.

D.1.5 HyperNet Rank Extended (Collapse Tasks)

The main HyperNet rank ablation (Table 11) sweeps $r \in \{8, 16, 32\}$ on the set-regression and set-matching tasks. We extend it to the collapse-exhibiting tasks, MEB, Hull, and MST at high d, with $r \in \{16, 32, 64\}$ (where r=64 matches the key dimension $d_k = 64$, giving full-rank capacity to the low-rank update $U(\mu)V_{\text{hyper}}$).

The full numerical content is reported in main-body Table 5. The trend is strikingly flat: across all ranks $r \in \{16, 32, 64\}$ on the hull and MEB cells, HyperNet's R^2 changes by at most 0.020 (MEB d=32), with most cells changing by under 0.015. The MST cells show somewhat larger inter-rank variation (≤ 0.29 at d=128), but every cell's three rank values fall well below their corresponding seed-pool envelope and well below MZAttn-Full's R^2 on the same task. Ranking up the low-rank update $U(\mu)V_{\text{hyper}}$ from r=8 (main table) to r=64 (full key-dimension rank) does not close the gap to MZAttn: the largest HyperNet R^2 on each task remains 0.6+ R^2 behind MZAttn-Full (Hull d=3: best HyperNet 0.204 vs. MZAttn-Full 0.804; MEB d=8: best HyperNet 0.022 vs. MZAttn-Full 0.691). Combined with the main Table 11, this provides evidence that the advantage of MZAttn is architectural (the zonotope structure $M_c + \sum_l \gamma_l M_l$) and not a capacity argument about generic input-dependent weight generation.

D.1.6 Parameter Efficiency Ablation

To address the concern that MZAttn's advantage might stem from its larger parameter count (∼2× Standard), we ablate the two parameters controlling MZ capacity: n_g (generator slots per zonotope token) and L (matrix zonotope generator count). We test three variants: MZ-Full (n_g=8, L=4, used in main experiments), MZ-Med (n_g=4, L=2), and MZ-Slim (n_g=2, L=1).

The full numerical content is reported in main-body Table 6. Table 6 shows that MZ's advantage is largely insensitive to the generator count. MZ-Slim (333K parameters, 1.6× Standard) achieves the best Max result ($R^2 = 0.871$ vs. Standard's 0.812), a 41% MSE reduction on MST (0.835 vs. 0.719), and a 72% MSE reduction on rotation matching (0.919 vs. 0.706). Performance is nearly flat across the three MZ variants: the largest gap between MZ-Slim and MZ-Full on any task is 0.017 R^2 (MST), within the seed variance for high-TDOF problems. On Max, MZ-Slim in fact outperforms MZ-Full by 0.021 R^2, which suggests smaller generator budgets can act as regularization on simpler tasks. Together these results indicate the advantage comes from the matrix-zonotope structure rather than the number of generators; even $L = 1$ generator is enough to produce the key inductive bias for high-TDOF tasks.

D.1.7 Depth Separation Validation

1-layer MZAttn outperforms all standard depths including 6-layer (2× more parameters).

D.1.8 MZ in Standard Transformers

MZ-attention also achieves $R^2 > 0.99$ in a sequence Transformer with positional encoding (Table 13), suggesting compatibility beyond Set Transformers; broader sequence-domain evaluation is left to future work.

Table 12: Depth separation: 1-layer MZAttn vs. D-layer standard Set Transformer (mean ± std over 3 seeds). 1-layer MZ outperforms standard depths up to 6 layers at 2× more parameters.

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

Table 13: MZ-attention in a standard Transformer with positional encoding.

L	**MZ-Transformer**	Std-Transformer	Params (MZ / Std)
4	**.9998**$_{\pm <.001}$	.9982$_{\pm <.001}$	83K / 44K
8	**.9981**$_{\pm <.001}$	.9864$_{\pm .001}$	87K / 44K
16	**.9964**$_{\pm .001}$	.9619$_{\pm .002}$	96K / 44K
24	**.9963**$_{\pm <.001}$	.9526$_{\pm .007}$	105K / 44K
32	**.9957**$_{\pm .002}$	.9313$_{\pm .003}$	114K / 44K

D.2 Robustness and generalization

Sensitivity of the architectural advantage to nuisance factors: input noise, set-size shift at test time, target-scale preprocessing, and TDOF-estimate uncertainty. MZAttn retains its lead across all four.

D.2.1 TDOF Estimate Sensitivity (Spearman robustness)

The Spearman correlation $\rho = 0.74$ between estimated TDOF and MZAttn's MSE reduction (Section 5, Figure 2, right) relies on the per-task TDOF assignments documented in our analysis. To verify this correlation is not an artefact of any specific TDOF estimate, we perturb each of the 18 task TDOFs independently by a multiplicative factor in $[0.5, 1.5]$ (uniform), recompute the Spearman, and repeat 10,000 times. We also run a log-symmetric variant ($\times[1/1.5, 1.5]$ in log scale) to be invariant to the wide TDOF dynamic range (1 to 2016).

Table 14: Spearman robustness to per-task TDOF perturbation. Both ±50% uniform and $\times[1/1.5, 1.5]$ log-symmetric perturbations preserve the strong correlation: in 99.98% (9,998/10,000) of uniform trials and in at least 95% of log-symmetric trials, the resulting Spearman remains ≥ 0.6. The headline correlation is therefore not driven by any single TDOF estimate.

Perturbation scheme	5^{th} pctile ρ	median ρ	95^{th} pctile ρ	frac. trials with $\rho \geq 0.6$
Baseline (no perturbation)		$\rho = 0.7402$		—
±50% uniform multiplicative	0.675	0.733	0.791	99.98%
$\times[1/1.5, 1.5]$ log-symmetric	0.690	0.735	0.781	$\geq 95\%$

The Spearman correlation is robust: 9,998 of 10,000 uniform-perturbation trials yield $\rho \geq 0.6$ (min $\rho = 0.59$), and the 5th-percentile ρ stays above 0.67 in both perturbation schemes, indicating that the headline relationship between TDOF and MSE reduction does not depend on any specific assignment within a ±50% window. Verification script: `Base/tdof_sensitivity_analysis.py`

writes the full statistics to `results/tdof_sensitivity.json`.

D.2.2 Ambiguity Sensitivity

Table 15: Ambiguity sweep on set matching: MSE as a function of noise σ.

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%

MZAttn consistently reduces MSE by 23–35%, with the largest advantage at moderate noise ($\sigma = 0.3$).

D.2.3 Set Size Generalization

Table 16: Size generalization: trained on sets of size 10–30, tested on larger sets.

Test size	ST MSE	ST R^2	MZ MSE	MZ R^2	ΔMSE
$[10, 30]$	59.87	0.957	**38.92**	**0.972**	−35.0%
$[30, 50]$	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%

MZAttn maintains its advantage across out-of-distribution sizes (12–29% MSE reduction).

D.2.4 Optimization Confound Ablation

An important question is whether the catastrophic negative R^2 of baselines at high d reflects a representational limit or an optimization artifact from training on raw large-scale regression targets. We directly test this by repeating ST-Large on MST d=128 and d=256 under target normalization (subtract train mean, divide by train std) and compare to the same runs without normalization.

We therefore re-ran the full main-table MST and convex hull scaling sweeps with target normalization and 5 seeds per setting, to give the cleanest architectural gap estimate (Tables 17 and 18).

Interpretation. These normalized tables, not the raw-target collapse tables, are the cleanest reading of the MST and convex hull scaling experiments. Raw-target-claim caveats: the negative-R^2 plunge in Table 23 is substantially the behavior of Standard and ST-Large on unnormalized regression targets of magnitude 10^4–10^5 (MST weights) or 10^1–10^3 (hull volumes); under standard target normalization the effect largely vanishes. MZAttn is less sensitive to this choice, which is in itself a useful practical property (one fewer hyperparameter to tune), but it does not establish a representational gap by itself.

The MEB (Table 3) and quadratic TDOF (Table 7) results do establish such a gap: MEB has $O(1)$-magnitude targets for which normalization changes nothing and yet MZAttn reaches $R^2 = 0.38$–0.69 while all baselines are stuck at $R^2 \leq 0.025$; the quadratic TDOF task is scale-free

Table 17: MST weight, normalized targets (R^2, mean over 5 seeds). Under target normalization, the dramatic negative-R^2 collapse of baselines (Table 23) disappears: all models reach $R^2 \in [0.37, 0.74]$. MZAttn-Full nonetheless retains a monotonically shrinking architectural advantage of +0.04 to +0.08 R^2 over the best context-rigid baseline, tying only at d=256. The dramatic collapse in the raw-target table is thus partially an optimization artifact of scale-sensitive regression on large-magnitude targets; the persistent (if smaller) normalized gap is the architectural effect.

d	Standard	ST-Large	HyperNet	Perceiver	MZAttn-Full	MZAttn-Large
16	0.660	0.658	0.495	0.467	**0.738**	0.688
32	0.565	0.567	0.435	0.417	**0.640**	0.617
64	0.539	0.524	0.447	0.435	**0.594**	0.560
128	0.510	0.493	0.443	0.401	**0.554**	0.497
256	0.461	0.462	0.373	0.363	**0.469**	0.447

Table 18: Convex hull volume, normalised targets (R^2, mean over 5 seeds). Under target normalisation the hull gap is non-monotonic: MZAttn-Full wins at d=6 (+0.14), ties at d=3, and is within seed variance of the leader at d=9. Compare to Table 22 (raw targets), where the mid-d baseline R^2 plunges to −0.18 from training-scale instability.

d	Standard	ST-Large	HyperNet	Perceiver	LA-ST	MZAttn-Full	MZAttn-Large
3	0.822	**0.866**	0.727	0.725	0.824	0.863	0.795
6	0.292	0.412	0.195	0.209	0.408	**0.555**	0.497
9	0.138	0.160	0.108	0.119	**0.172**	0.123	0.143

by construction and yet every non-MZ architecture (including EGNN) collapses to $R^2 \leq 0.01$ for $L \geq 4$. These are the cleanest architectural evidence. On the SGSP family (Table 24), MZAttn wins uniformly at d=8, wins on 3 of 4 tasks at d=16, and crosses over with baselines at d=32. We thus position MZAttn's empirical contribution as: (i) a clean architectural win on a subset of tasks (MEB, quadratic-TDOF, low-d SGSP); (ii) a consistent smaller advantage on MST scaling under normalization; (iii) robustness to target scale decisions that affect context-rigid baselines more severely. Item (iii) is a real-world usability advantage but is not itself architectural.

D.3 Uncertainty calibration

Empirical evaluation of the single-pass interval-hull-width (IHW) uncertainty estimator inherited from the matrix-zonotope output, both qualitatively (Spearman / Pearson correlation with error) and via standard calibration metrics (ECE, NLL) against MC-Dropout and Deep Ensembles.

D.3.1 Uncertainty Calibration

IHW achieves competitive uncertainty quality with a single forward pass at 3.7× lower cost.

D.3.2 Quantitative Uncertainty Calibration (ECE and NLL)

MC-Dropout achieves the lowest ECE and NLL but requires 30× inference cost. MZAttn's IHW yields NLL 5.02, nearly 2× lower than Deep Ensemble (9.17), while matching Ensemble-level ECE (0.316 vs. 0.337) in a single forward pass at 1× storage. MZAttn is therefore the only method here that requires neither extra forward passes nor extra stored model copies; the practical use case is

Table 19: Uncertainty calibration: samples binned by IHW. Higher IHW → higher error (Spearman $r = 0.51$).

Bin	IHW (mean)	Mean |error|	Count
1 (lowest)	5.33	1.13	400
2	5.52	2.01	400
3	5.72	3.15	400
4	6.03	4.20	400
5 (highest)	6.64	6.81	400

Table 20: Uncertainty quality comparison on set matching.

Method	Spearman ↑	Pearson ↑	Inference	Passes
MZAttn (IHW)	0.513	0.442	**0.10s**	**1**
MC-Dropout	**0.553**	**0.547**	0.37s	10
Deep Ensemble	0.470	0.542	0.40s	5

latency- and storage-constrained deployment where 30× MC-Dropout passes or 5× ensemble storage are prohibitive. The remaining ECE gap relative to MC-Dropout may be reducible via post-hoc temperature scaling.

D.4 Computational geometry and scaling

High-dimensional and large-set scaling behaviour on classical computational-geometry targets. Together they establish that the architectural advantage is preserved (and often grows) at scale on tasks satisfying the structural prerequisites of Theorem 1.

D.4.1 Minimum Enclosing Ball: A Computational-Geometry Benchmark

The MEB benchmark and parameter-matched baseline collapse are reported in the main body (Table 3, Section 5). We expand here on the experimental design and the qualitative pattern across dimensions, and report MZAttn-Large as a generator-budget ablation that complements the headline MZAttn-Full configuration. The full benchmark uses Welzl's algorithm [42] for ground-truth radii; point clouds are mixtures of Gaussians with $n \in [10, 30]$. MZAttn-Large is reported as a generator-budget ablation. At d=8, MZAttn-Large reaches $R^2 = 0.621 \pm 0.013$ under the headline 5-seed / 5-epoch-warmup protocol (`meb_results_combined.json`). At d=16, 32, the higher-dimensional MZ-Large training is sensitive to LR warmup, so we report 10 seeds with extended 10-epoch warmup (`meb_results_warmup10*.json`): $R^2 = 0.488 \pm 0.099$ at d=16 and $R^2 = 0.221 \pm 0.074$ at d=32. All three are consistently below the MZAttn-Full numbers in the main body, supporting the choice of n_g=8, L=4 as the recommended configuration.

Three observations: **(i) Baseline collapse is uniform and clean.** None of the five baselines, spanning the full 2017–2024 range (Set Transformer, Perceiver, HyperNet, a parameter-matched Set Transformer variant, and a linear-attention baseline representing the 2020s efficient-attention family), achieves $R^2 > 0.025$ on any dimension. The task is structurally hostile to context-rigid attention: to predict the radius, the model must identify a small subset of convex-hull extrema (at most $d + 1$ points) from among the input, and context-rigid W_V cannot express the required indicator-like selector. **(ii) MZAttn dominates by 28–155×.** MZ-Full achieves $R^2 = 0.691$ at d=8 (≈ 155× HyperNet's full-precision 0.0044, 28× Standard's 0.025), $R^2 = 0.629$ at d=16 (40× Perceiver's

Table 21: Uncertainty calibration on set matching (mean ± std over 3 seeds). ECE: expected calibration error (lower better); NLL: Gaussian negative log-likelihood (lower better). MZAttn's single-pass IHW uncertainty is on par with the 5-model Deep Ensemble on ECE and lower than it on NLL; MC-Dropout achieves the best ECE/NLL but at 30× inference cost.

Method	MSE ↓	ECE ↓	NLL ↓	Passes	Models
MZAttn (IHW)	$55.0_{\pm 10.5}$	$0.316_{\pm 0.001}$	$5.02_{\pm 0.03}$	**1**	**1**
MC-Dropout	$59.8_{\pm 6.1}$	$\mathbf{0.197}_{\pm 0.030}$	$\mathbf{3.52}_{\pm 0.43}$	30	1
Deep Ensemble	$\mathbf{47.8}_{\pm 1.1}$	$0.337_{\pm 0.019}$	$9.17_{\pm 1.96}$	1	5

0.016), and $R^2 = 0.378$ at d=32 (37× Perceiver's 0.010). **(iii) The Linear Attention baseline (a 2023–2024-era design family) performs indistinguishably from Set Transformer.** LA-ST reaches only $R^2 = 0.017, 0.015, 0.001$ at d=8, 16, 32, within 1σ of Set Transformer. This rules out the hypothesis that "modern efficient attention solves this": the gap is not a softmax-vs-linear issue but an architectural capacity issue for context-adaptive operators, consistent with Theorem 8.

D.4.2 Convex Hull Volume: Another Computational-Geometry Benchmark

To verify the MEB collapse pattern is not specific to one task, we evaluate on a second classical computational-geometry quantity: the volume of the convex hull of a point set in $\mathbb{R}^d$, computed exactly via the QHull algorithm [4]. The convex hull is determined by $O(n^{\lceil d/2 \rceil})$ extremal points and its volume is smooth in their positions, giving a regression target with sparse combinatorial support similar to MEB but with substantially richer d-scaling.

Table 22: Convex hull volume prediction (R^2, mean ± std over 5 seeds; $n \in [15, 35]$ points in $\mathbb{R}^d$). At d=3 MZAttn-Large beats the best baseline by +0.32 R^2. From d=6 onwards every parameter-matched baseline collapses to negative R^2 on raw targets; MZAttn-Large is the only architecture significantly above $R^2 = 0$ at all three tested dimensions. The d=9 MZ-Full result (0.003 ± 0.006) is statistically indistinguishable from zero and is reported for completeness; Appendix D.2.4 reports normalised-target equivalents where the d=9 gap structure differs.

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

The pattern exactly mirrors MEB: at low d, MZAttn substantially outperforms baselines; at moderate d, baselines cross zero while MZAttn remains positive; even at d=9, where the advantage narrows, MZAttn is the only architecture producing meaningful predictions. Together with MEB (Section D.4.1), this gives two independent computational-geometry benchmarks (MEB: boundary of minimum-radius ball; convex hull: boundary of convex hull) on which every context-rigid attention variant, including 2020s linear-attention, cannot exceed baselines, while MZAttn remains on the right side of zero.

D.4.3 Rotation Matching Scaling at High Dimension

The paper's Task F (rotation matching) in $\mathbb{R}^8$ shows MZAttn outperforming baselines by 41–57% MSE. Our theory predicts this advantage should grow with dimension d, since off-diagonal rotation components (which FiLM-style diagonal modulation provably cannot express) become more numerous: a rotation in $\mathbb{R}^d$ requires $d(d-1)/2$ independent Givens angles, while FiLM only has d scalar parameters. We test this prediction at $d \in \{16, 32, 64, 128, 256, 512\}$.

The full numerical content is reported in main-body Table 4. The gap between MZAttn and the baselines grows monotonically with d: at d=16, MZAttn reaches $R^2 = 0.826$ while the best baseline (ST-Large) reaches 0.285, a 76% MSE reduction; at d=32, all four baselines drop to $R^2 \approx 0$ (no better than predicting the training mean), while MZAttn still attains $R^2 = 0.519$; at d=64, the baselines have negative R^2 (actively worse than the mean predictor), and MZAttn is the only architecture producing useful predictions ($R^2 = 0.071$); at d=128, the baselines collapse catastrophically to $R^2 = -9$ to -22 (MSE 10–23× worse than the mean predictor, with large seed variance), while MZAttn remains at $R^2 \approx 0$ (MSE within 1% of the mean predictor's); at d=256, all baselines reach $R^2 \approx -50$ to -62 (∼50× MSE ratio vs. MZAttn's $R^2 = -0.005$); at d=512, MZAttn also begins to diverge ($R^2 = -62$) but remains 2.3× better in raw MSE than the next-best baseline (ST-Large at -145). The divergence at d=512 reflects that a d=512 rotation has $d(d-1)/2 = 130{,}816$ Givens parameters, exceeding our generator budget L=4 by four orders of magnitude; the architecture has the capacity to express the operator family ($d_k^2 = 4{,}096$ entries per generator, $L \cdot d_k^2 \approx 16$K free parameters in the headline configuration) but training stability becomes the bottleneck before representational capacity does, a separate scaling regime from the $d \leq 256$ rows where the headline gap holds. This matches the theoretical analysis: context-rigid attention (fixed W_V) cannot represent the context-adaptive rotation operator $R(\mu)$, and the gap widens monotonically with d because the space of rotations grows as $d(d-1)/2$. Standard attention's R^2 also becomes highly variable at $d \geq 64$ (std ±0.47 at d=64; ±3.5 at d=128), consistent with its failure being due to an expressibility limit rather than a tractable optimization problem.

D.4.4 MST Weight Scaling at High Dimension

The paper's Task G (MST weight in $\mathbb{R}^8$) already showed MZAttn reducing MSE by 41% over ST-Large. We extend this sweep to $d \in \{16, 32, 64, 128, 256\}$ to test whether the advantage on this combinatorial-geometry quantity also grows with dimension (as theory predicts: pairwise-distance structure has $O(d^2)$ degrees of freedom).

Table 23: MST weight scaling (R^2, mean ± std over 5 seeds). As d increases past 32, every non-MZ baseline collapses to strongly negative R^2 (worse than predicting the mean). MZAttn is the only architecture that remains above $R^2 = 0$ for all $d \leq 256$, with MSE 4–8× lower than the best baseline at $d \geq 64$. LA-ST is the linear-attention baseline [16], whose feature-map kernel underlies 2023–2024 architectures including RetNet [39] and Gated Linear Attention [44].

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

Three features of Table 23 are notable: **(i)** Both MZAttn variants dominate all baselines across every dimension tested. The gap grows monotonically with d: MZ-Full beats ST-Large by +0.20 R^2 at d=16 and +0.35 at d=32, and from d=64 onwards every single baseline has negative R^2 while MZAttn remains above 0.38. At d=64, MZ-Full's MSE is 3× lower than the best baseline; at d=256 the ratio is over 5×. **(ii)** Baseline failure at $d \geq 64$ is not just a performance drop but a qualitative one: Standard, HyperNet, and Perceiver all reach $R^2 < -1$, indicating MSEs larger than a constant mean predictor's, and ST-Large, which remained competitive at $d \leq 32$, also falls to $R^2 = -2.43$ at d=256. MZAttn's MSEs stay within 2–3× of the mean predictor's, with seed std below 0.04, whereas baselines have seed std of 0.2–0.4. **(iii)** At smaller dimensions ($d \leq 64$), MZ-Full outperforms MZ-Large (consistent with Appendix D.1.6: extra generators act as overparameterization on moderate-TDOF tasks); at $d \geq 128$, MZ-Large overtakes MZ-Full (0.435 vs. 0.384 at d=256), suggesting that as the task's effective TDOF grows with d, the larger generator budget begins to pay off. This crossover aligns with the theoretical prediction that generator count should match the target function's TDOF.

D.4.5 Sparse Geometric Set Predicates (SGSP)

Motivated by the MEB, convex-hull, and MST successes, we identify a broader task family we call Sparse Geometric Set Predicates: targets that are (i) computed from a continuous d-dimensional point cloud (no symbolic or one-hot encoding), (ii) a smooth function of a combinatorially sparse subset of the input (extreme pairs, nearest neighbors, boundary points, or small distance-subset cuts). This family is where MZAttn's matrix-zonotope machinery is predicted to provide the largest advantage because the target depends on a per-sample "selector operator" that extracts the relevant sparse subset.

We test four members of this family beyond the MEB, hull, and MST results reported above:

Table 24: Sparse Geometric Set Predicates benchmark (R^2, mean over 5 seeds; 7 models × 5 seeds × 3 dims × 4 tasks = 420 runs total). MZAttn-Full wins every d=8 task with margin +0.04 to +0.13, maintains an advantage at d=16 on 3 of 4 tasks, and is overtaken by parameter-matched Set Transformer variants at d=32. This crossover is consistent with concentration-of-measure: at high d, distances between random points become tightly concentrated and sparse-combinatorial targets (max pairwise, k-th-NN, k-center) reduce to aggregate statistics readily captured by context-rigid attention.

Task	d	Standard	ST-Large	HyperNet	Perceiver	LA-ST	MZAttn-Full	MZAttn-Large
diameter	8	0.723	0.768	0.702	0.674	0.787	**0.876**	0.814
	16	0.732	0.748	0.706	0.682	0.755	**0.806**	0.727
	32	0.774	**0.775**	0.687	0.683	0.740	0.710	0.663
knn-radius	8	0.721	0.786	0.697	0.611	0.815	**0.869**	0.810
	16	0.703	0.699	0.655	0.593	0.710	**0.784**	0.721
	32	**0.740**	0.728	0.651	0.629	0.697	0.685	0.653
k-smallest sum	8	0.116	0.109	0.034	0.083	0.108	**0.152**	0.097
	16	0.009	0.009	−0.005	0.005	0.007	**0.010**	−0.039
	32	−0.010	−0.024	−0.003	−0.011	−0.017	**0.014**	0.001
furthest-first	8	0.422	0.478	0.275	0.298	0.456	**0.604**	0.531
	16	0.396	0.391	0.163	0.249	0.365	**0.438**	0.205
	32	**0.315**	0.276	0.100	0.209	0.252	0.225	0.206

These tasks are classical combinatorial-geometry quantities computed by standard algorithms

(pairwise-distance matrix, k-nearest-neighbor sort, greedy farthest-first traversal, k-smallest pair enumeration). None has an exploitable group symmetry; targets depend on the magnitudes of specific pairwise distances, not on the global coordinate frame, so equivariant baselines gain no free advantage. The pattern in Table 24 has three regimes: **(i) Low dimension (d=8)**: MZAttn-Full wins on all four tasks with margins +0.036 (min-sum) to +0.126 (furthest-first) over the best context-rigid baseline. This is the regime where the sparse-combinatorial structure is most difficult for context-rigid attention, and MZAttn's per-sample operator adapts to identify the relevant subset (extremal pair for diameter, k-th-NN for knn-radius, farthest cluster center for k-center). **(ii) Moderate dimension (d=16)**: MZAttn-Full retains an advantage on 3 of 4 tasks (+0.04 to +0.08), though the gap narrows. **(iii) High dimension (d=32)**: MZAttn is overtaken by Standard or ST-Large on 3 of 4 tasks. We attribute this to concentration of measure: in high dimension, pairwise distances between isotropic points concentrate tightly around their mean, so sparse-combinatorial targets (diameter, k-th-NN, k-center) become near-deterministic functions of simple aggregate statistics (mean pairwise distance, set variance) that context-rigid attention readily extracts. The per-sample-operator advantage of MZAttn no longer applies because all samples are, statistically, nearly identical.

This d-dependent crossover is the mirror image of the MST and rotation-matching scaling pattern (Tables 23, 4), where MZAttn's advantage grows with d. The two regimes correspond to distinct mechanisms: MST and rotation involve $O(d^2)$-parameter operators (pairwise-distance tree, off-diagonal rotation) whose complexity scales with d; SGSP involves $O(1)$-parameter selectors (pick 2 extreme points, pick k boundary points) whose intrinsic complexity is dimension-independent and whose d-scaling concentrates rather than diversifies. The unified theoretical prediction is that MZAttn helps when TDOF grows with d relative to the information content of the aggregate statistics: a prediction confirmed by both scaling directions.

D.4.6 Large-Scale Set Task Generalization

To verify that MZAttn's advantage persists at larger set sizes n and element dimensions d than those used in the main experiments, we evaluate on three synthetic tasks (max, sum, range) across a sweep of scales: $d \in \{16, 64, 128\}$ and n ranging from 10–50 (small) up to 500–1000 (large). Each configuration is trained for 150 epochs with 3 seeds.

Table 25: Large-scale generalization (R^2, mean over 3 seeds). MZAttn's advantage on the sum task grows with scale: at d=64, n=200–500, MZ reaches $R^2 = 0.998$ while ST-Large reaches only 0.946, a 30× MSE reduction. On the simpler max task, architectural differences diminish as capacity becomes the bottleneck; on range, all models struggle regardless of architecture.

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

Table 25 reveals three scaling regimes: **(i)** Sum at scale: MZAttn retains its large-margin advantage ($R^2 = 0.998$ vs. 0.946 for ST-Large at d=64, n=200-500), showing that the zonotope structure's inductive bias for aggregation tasks holds at scale. **(ii)** Max at scale: while MZAttn leads at small scale (d=16: $R^2 = 0.843$ vs. 0.748 for Standard), at larger scales ($n \geq 200$) the max task becomes capacity-bottlenecked and all models converge to similar performance ($R^2 \approx 0.74$ at medium, 0.63 at large). **(iii)** Range at scale: a known-hard task where all models struggle

uniformly ($R^2 < 0.1$ at $n \geq 200$). Range over $n \geq 200$ i.i.d. samples concentrates tightly around its expectation (the empirical max minus min has variance scaling as $O(\log n/n)$), so the regression target is nearly constant with respect to typical input variations; the low R^2 across architectures reflects this concentration ceiling rather than a relational-capacity gap, and is a property of the task at this scale. These findings support the theoretical predictions of Theorems 1 and 2: MZ's advantage is selective, tied to the relational complexity (TDOF) of the target function, and persists at scale when the task demands high-TDOF operators (as in sum, which requires a full linear combination of elements).

D.4.7 Early-Stopping Patience Ablation (ResNet-50 d=1024)

The main-text Table 33 reports ResNet-50 (d=1024) results with early-stopping patience set to 300, a departure from our default patience of 20 used elsewhere. To verify that this is not a setting where the patience choice selectively favours MZAttn, we sweep the patience value $\in \{20, 100, 300\}$ for the four top contenders (MZ-Full, ST-Large, HyperNet, LinearAttn) on ResNet-50 k-center furthest (k=5, n=16, d=1024) with 5 seeds each; all other hyperparameters are held fixed. Results are in Table 26.

Table 26: Patience sweep on ResNet-50 k-center furthest (R^2 mean ± std over 5 seeds, 80 epochs). Across all three patience values $\{20, 100, 300\}$ every model converges to within 0.002 R^2 of its patience-300 plateau, and the ranking stabilises to MZAttn-Full > ST-Large ∼ HyperNet > LinearAttn. Patience choice does not selectively favour any architecture: MZAttn-Full's lead over the next-best architecture is 0.003–0.005 R^2 across the full patience range. Per-seed JSONs in `real_geom_cifar_resnet50_furthest_diag_p20_*.json`, `_sweep_p100_*.json`, and `_diag_p300_*.json`.

Model	patience = 20	patience = 100	patience = 300
MZAttn-Full	**0.795±0.001** (1st)	**0.794±0.002** (1st)	**0.796±0.001** (1st)
ST-Large	0.790±0.001 (3rd)	0.791±0.001 (2nd)	0.791±0.002 (2nd)
HyperNet	0.790±0.003 (2nd)	0.790±0.003 (3rd)	0.789±0.003 (3rd)
LinearAttn	0.788±0.002 (4th)	0.788±0.003 (4th)	0.789±0.003 (4th)

Interpretation. In a fresh 5-seed sweep at the three patience values $\{20, 100, 300\}$ (epochs 80, batch size 128, learning rate 10^{-4}, identical across runs), every model converges to within ±0.002 R^2 of its patience-300 plateau. The MZ-Full > ST-Large ∼ HyperNet > LinearAttn ranking is therefore stable across the full patience range, with MZAttn-Full's lead over the next-best architecture remaining at 0.003–0.005 R^2 throughout. The main-text claim that MZAttn-Full ranks first on ResNet-50 k-center furthest does not depend on the specific patience value chosen for early stopping.

D.5 Real-world experiments and equivariant comparison

Real-data and meta-learning evaluations beyond the synthetic high-TDOF regime, plus a side-by-side comparison with E(n)-equivariant graph networks on tasks where strict symmetry is or is not present. These delineate where MZAttn's architectural advantage transfers and where it does not.

D.5.1 Few-shot Meta-Regression: a 7-way Parameter-Fair Comparison

The main-body experiments test whether MZAttn can fit a fixed target whose intrinsic complexity requires a high-TDOF operator. A natural follow-up question is whether MZAttn also excels when

the operator must be inferred from a support set at inference time. We design a few-shot meta-regression task that is the meta-learning analogue of the quadratic TDOF benchmark in Section 3.2 and report a parameter-fair 7-way comparison against every architecture used in the main results table, addressing the concern that a 2-method (Standard vs. MZAttn) baseline is not parameter-fair.

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 27: Few-shot meta-regression at L=32 (highest-TDOF setting), parameter-fair 7-way comparison (MSE ↓, mean±std over 3 seeds). At shallow depth (n=2), most non-MZ baselines fail to learn the task (MSE ≈ 30); Standard narrowly wins against MZAttn. At matched depth (n=4), ST-Large is the clear winner and MZAttn-Large places second, beating Standard, HyperNet, Perceiver, and LinearAttn. Perceiver fails to converge at either depth.

Depth	Standard	ST-Large	HyperNet	Perceiver	LinearAttn	MZAttn-Full	MZAttn-Large
n=2	**0.79**	0.83	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.

Result interpretation. The 7-way comparison shows that generic meta-regression, where the target is a random linear operator in an L-dimensional subspace, is not a setting where MZAttn dominates. ST-Large, a wider Set Transformer, wins at L=32 (MSE 0.17 vs. ≥ 0.36 for every other method), and at L=4–8 it wins or ties with LinearAttn. MZAttn-Large ranks second at L=32 with 4 layers (MSE 0.36), beating Standard, HyperNet, Perceiver, and LinearAttn. Previous versions of this paper reported a 38–51% MSE reduction of MZ over Standard; that comparison was against a single 77K-parameter vanilla Standard Transformer and did not include parameter-matched ST-Large (255K) or LinearAttn (380K), both of which in fact outperform MZ here.

What the corrected result implies. The mismatch between this meta-regression outcome and the main-body wins on MEB, SGSP, MST, rotation matching, and convex hull is informative. Those main-body tasks share a feature that meta-regression does not: their target operators are algorithmic (boundary-point selection for MEB and hull, minimum-spanning-tree edge selection for MST, rotation for rotation matching), and therefore have sharp combinatorial structure that an expressive but unstructured operator class (such as wide attention's effective W_V) struggles to approximate with smooth gradient updates. Meta-regression by contrast has dense, smooth A_τ operators drawn from a Gaussian, which a high-capacity standard attention layer can fit directly. We therefore scope the empirical claim precisely: MZAttn is the architecture of choice when the target operator has sharp algorithmic and combinatorial structure and a matching TDOF (MEB, SGSP, MST, rotation, convex hull), not as a universal improvement to attention on any high-TDOF-seeming task.

Connection to data-driven reachability. The algebraic machinery of MZAttn has a natural interpretation through the lens of data-driven reachability analysis [2]. In that framework, a matrix zonotope represents the set of system matrices consistent with noisy observations; in ours, the learned MZ represents the family of "plausible relational mappings" consistent with the observed context. MZ–zonotope multiplication propagates this set-valued information through the network, maintaining a structured envelope of possible transformations at each layer, an inductive bias that is absent from pointwise architectures.

D.5.2 QM9 Molecular Property Prediction

To evaluate on a real-world scientific benchmark, we test all models on QM9 molecular property prediction [31]. We parse the original `gdb9.sdf` with RDKit [32] 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 predict three properties spanning distinct physical regimes: dipole moment (μ, depends on charge distribution), isotropic polarizability (α, depends on spatial extent), and HOMO–LUMO gap (electronic structure). 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.

Table 28: QM9 molecular property prediction across all 12 targets (R^2, mean over 3 seeds; best in **bold**; the Gap column is unbolded because ST-Large 0.879 and MZ-Large 0.874 tie within 0.005 R^2). Parameter-matched attention models (ST-Large, Perceiver, HyperNet, MZAttn) cluster within 0.04 R^2 on the 8 thermodynamic targets (α, R^2, ZPVE, U_0, U_{298}, H_{298}, G_{298}, C_v); on the 4 electronic-structure / frontier-orbital targets (μ, HOMO, LUMO, Gap) the spread is wider (HyperNet trailing by 0.08–0.15 R^2), reflecting their higher relational complexity. MZ-Large (n_g=16, L=8) is evaluated under batch size 64 on all 12 targets to keep the protocol consistent across the row (the same per-target budget MZ-Large was originally validated under for HOMO/LUMO); MZ-Large tops the original ST-Large on HOMO (+0.033 R^2) and LUMO (+0.074 R^2), and matches it on Gap (within 0.005 R^2, where ST-Large 0.879 $\geq$ MZ-Large 0.874); a parameter-matched ST-XLarge subsequently closes the HOMO/LUMO gap (Appendix D.5.2); on the remaining nine targets MZ-Large fails to improve over the default MZAttn and degrades by 0.11–0.26 R^2, consistent with the TDOF prediction that scaling generator capacity helps only on targets that admit a non-trivial context-adaptive solution. Specialised equivariant models (SchNet [35], DimeNet [10]) achieve higher absolute accuracy with hand-designed features.

Model	Params	μ	α	HOMO	LUMO	Gap	R^2	ZPVE	U_0	U_{298}	H_{298}	G_{298}	C_v
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	0.954	0.856	0.856	0.856	0.859	0.940
ST-Large	361K	**0.795**	0.914	0.842	0.852	0.879	**0.967**	**0.958**	0.864	0.862	0.863	0.858	**0.950**
ST-XLarge	443K	–	–	**0.883**	**0.937**	–	–	–	–	–	–	–	–
Perceiver	341K	0.766	**0.920**	0.774	0.819	0.824	**0.967**	0.955	**0.891**	**0.886**	**0.883**	**0.888**	0.948
HyperNet	360K	0.641	0.902	0.722	0.773	0.786	0.955	0.934	0.871	0.863	0.869	0.876	0.922
MZAttn	368K	0.761	0.911	0.802	0.830	0.855	0.947	0.952	0.877	0.875	0.871	0.867	0.940
MZAttn-Large	417K	0.500	0.698	0.875	0.926	0.874	0.823	0.702	0.759	0.760	0.722	0.757	0.676

Table 28 reports all 12 QM9 targets. With the default generator budget (n_g=8, L=4), MZAttn is competitive with the best parameter-matched attention baseline on every target (within 0.05 R^2),

confirming the TDOF prediction that adding context-adaptive capacity does not regress on low-to-moderate-TDOF molecular targets. MZAttn matches or marginally exceeds Standard attention on 10 of 12 targets (the exceptions are LUMO, where Standard leads by 0.008 R^2, and the spatial-extent target R^2 where Standard leads by 0.013); on μ and α MZAttn edges Standard by 0.009 and 0.007 R^2 respectively. MZAttn is not the best on any single target. On the hardest electronic-structure targets (HOMO, LUMO), where baselines vary most, we also test a larger variant MZAttn-Large with n_g=16 zonotope generators and L=8 matrix-zonotope generators (417K parameters, 1.16× ST-Large). MZAttn-Large outperforms the original ST-Large on HOMO (0.875 vs. 0.842) and LUMO (0.926 vs. 0.852); however, when we additionally test ST-XLarge, a Set Transformer scaled to 443K parameters by widening d from 80 to 92, ST-XLarge attains $R^2 = 0.883 \pm 0.005$ on HOMO and $0.937 \pm <0.001$ on LUMO (n=3 seeds; `qm9_xlarge_results_homo.json`, `_lumo.json`), exceeding MZAttn-Large on both. The HOMO–LUMO gap of MZAttn-Large over ST-Large therefore reflects a parameter-count advantage rather than a structural one, and a parameter-matched context-rigid baseline closes the gap. The complementary observation is that MZAttn-Large improves over the default MZAttn on three frontier-orbital / electronic-structure targets (HOMO +0.073, LUMO +0.096, Gap +0.019 R^2) and underperforms it on the remaining nine targets by 0.11–0.26 R^2 (Table 28, last row), with the largest degradation on the heat capacity (C_v: 0.676 vs. MZAttn's 0.940). The TDOF-based capacity bound is again confirmed: scaling generator budget pays off only on the small subset of targets (HOMO, LUMO, Gap) that admit a non-trivial context-adaptive solution, and adds variance without bias correction on aggregate-statistic targets where capacity is already saturated. This is consistent with the broader scope claim of the paper: on low-to-moderate TDOF real-data targets like QM9, where the per-sample operator does not vary in a high-rank subspace, scaling context-rigid attention is a competitive alternative to structural augmentation. The complementary regime is the high-TDOF synthetic suite (MEB, MST, rotation matching, SGSP), on which an MZ-Slim variant with L=1 and only 1.6× Standard's parameter count already outperforms parameter-matched ST-Large (Appendix D.1.6); structural inductive bias dominates there, and capacity scaling does not catch up. The QM9 ST-XLarge equivalence and the synthetic MZ-Slim parameter-efficiency win together demarcate the boundary precisely. Deep Sets lags every attention-based model on dipole moment by ≥ 0.20 R^2, consistent with the higher relational complexity of that target; Slot Attention is competitive with the weakest attention baselines (within 0.04 R^2 on μ and α). Specialized equivariant models (SchNet [35], DimeNet [10]) achieve higher absolute accuracy, but require hand-designed radial basis functions and spherical harmonics, whereas all models compared here use the same generic set-input interface.

D.5.3 Slot Attention Baseline Comparison

To compare against the modern Slot Attention baseline [22] on tasks where our existing models have established results, we evaluate on four synthetic set tasks: max, sum, median, and range. These tasks span a range of relational complexity from low-TDOF (max, median) to moderate-TDOF (sum, range).

Table 29 shows Slot Attention underperforming every other architecture, including Deep Sets. Iterative slot competition was designed for object-centric image decomposition and does not appear to transfer to set-to-scalar regression. MZAttn is best on max ($R^2 = 0.847$ vs. 0.817 for Standard) and reaches $R^2 = 0.999$ on sum compared to 0.960 for Set Transformer, a 40× residual-MSE reduction relative to the Set Transformer baseline. On the low-TDOF median task, all models perform similarly ($R^2 \approx 0.65$), matching the theory's prediction.

Table 29: Slot Attention baseline comparison on synthetic set tasks (R^2, mean ± std over 3 seeds). Models are reported at canonical sizes (Slot Attention 59K, Deep Sets 36K, Set Transformer 208K, MZAttn ∼ 365K) rather than parameter-matched; the parameter-matched comparison against ST-Large (362K) and HyperNet (360K) is in Table 2. Slot Attention's iterative routing trails the standard and matrix-zonotope attention variants on all four tasks; MZAttn attains the highest R^2 on the sum task (0.999).

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

D.5.4 Comparison with E(n)-Equivariant GNNs

We compare MZAttn to E(n)-equivariant graph neural networks (EGNN) [33], a symmetry-enforcing baseline that uses pairwise distances $\|x_i - x_j\|$ as invariant features. All runs use 3 seeds; EGNN is parameter-matched at ∼303K (close to MZAttn's 368K) with h_{dim}=140 and 4 message-passing layers.

The full numerical content is reported in main-body Table 7. The pattern is architecturally interpretable: **(i) Where EGNN wins.** On every task where the label is invariant under rotations and translations of the input (rotation matching, convex hull volume, MEB radius, covariance Frobenius norm, Mahalanobis distance, PCA reconstruction error), EGNN outperforms MZAttn. This confirms the expected pattern: hard-coding the correct symmetry beats learning it. **(ii) Where MZAttn wins.** The quadratic TDOF target (Eq. 4) depends on specific Frobenius-orthogonal matrices $\{M_l\}$ in a fixed basis; rotating the input changes the target. EGNN, whose output is rotation-invariant by construction, cannot express this target and collapses to $R^2 \approx 0$ across $L \in \{4, 8, 16\}$. MZAttn handles this regime because its matrix-zonotope operator is not tied to any group action. **(iii) Target scale sensitivity.** Our default baselines train on raw regression targets (see Appendix D.2.4); the collapse of Standard and ST-Large at high d on MST is partly an optimization artifact that target normalization can mitigate. The MZAttn vs. ST-Large gap narrows under normalization on MST and Hull, but persists on MEB (where target magnitudes are small) and on the quadratic TDOF task where the issue is architectural rather than numerical. EGNN, being a different architecture family, is unaffected by this confound.

The comparison delineates MZAttn's scope: MZAttn is the right tool for context-adaptive linear operators that are not reducible to a known group action; EGNN is the right tool when the target's symmetry is known and can be enforced by construction.

D.5.5 Computational Cost

Table 30: Computational overhead (ratio MZ/Standard). RTX 5090, batch 64.

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×

D.5.6 Negative Transfer: Neural-Process Image Completion and Tabular Meta-Regression

To test whether MZAttn's advantage generalizes beyond synthetic high-TDOF relational tasks, we evaluated it on two families of real-world benchmarks where the structural prerequisites (per-sample operator variability in a high-dimensional subspace) are weak or absent. Both yielded ties or slight losses rather than wins; we report these honestly as they sharpen the scope of the claim.

Neural-Process image completion (MNIST, CIFAR-100). Following the Attentive Neural Process setup [17], 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 (K=50, T=200, 5 seeds), ST-Large achieved the best test NLL of -1.657; MZAttn-Large reached -1.599 at 3.4× the parameters of Standard (-1.617). On CIFAR-100 (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× the parameters. This pattern aligns with the theory: image pixel prediction is dominated by spatial-locality inductive bias, not per-image operator variance; the high-TDOF prerequisite for MZAttn's gain is absent, so extra capacity in the form of matrix-zonotope value projections yields no architectural advantage.

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, with comparable cross-seed standard deviations of ± 6.06 for Standard and ± 5.45 for MZAttn-Full at 10 seeds with patience 12), indicating that small-K in-context operators do not identify the true per-dataset regression operator when the test datasets differ structurally from training. The two architectures are within seed variance on NLL (MZAttn-Full $+20.02$ vs. Standard $+18.66$; gap 1.36, pooled std 5.76) 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, and the apparent five-fold seed-variance ratio reported in earlier 5-seed pilots is an artefact of small-sample standard-deviation estimation; the 10-seed run reports comparable spreads.

Systematic real-world transfer study. To empirically characterise which real-data configurations inherit the synthetic gain of MZAttn and which do not, we ran the geometric-target family (MEB, convex hull, MST weight, k-center furthest cost) on a systematic grid of point-cloud constructions across MNIST, Fashion-MNIST, and CIFAR-100, organised into three tiers of per-sample rank. Tier 1: raw 2D pixel coordinates of on-pixels, with optional per-sample 2×2 affine perturbations (TDOF-4) or 3-digit multi-cluster mixtures. Tier 2: 8D feature-lifted pixels (position + intensity + gradient + polynomial features) and CIFAR-100 4×4 image patches encoded as 8D patch-feature vectors (mean and std RGB, intensity, colorfulness), with variable-size and class-mixture variants. Tier 3: 256D ImageNet-pretrained ResNet-18 layer-3 features of CIFAR-100 (adaptively average-pooled to n=16 points per image). Together the tiers span a 128× increase in per-sample dimension while

holding the target family fixed, yielding a controlled ablation of the structural prerequisites the theory predicts. Table 31 reports MZAttn-Full's rank among 7 parameter-comparable baselines in each (construction, target) cell; the full numerical R^2 values and per-method seed statistics are in the supplementary JSON dumps.

Table 31: Systematic transfer study: MZAttn-Full rank (out of 7 parameter-matched methods) per (construction, target) cell, grouped by per-sample rank tier. 3 seeds per cell; dashes mark cells not in the grid. Mean ranks across populated cells: 3.9 on Tier-1 (low-rank 2D pixel, n=11 cells), 3.9 on Tier-2 (medium-rank 8D lift, n=5), 3.0 on Tier-3 (high-rank deep feature, n=2); the Tier-3 mean is descriptive only given the small cell count, and the rank=1 Furthest cell drives it. Expected rank under the no-advantage null is 4. All printed ranks reproduce from the per-cell JSONs via `Base/collect_transfer_matrix.py`.

Construction	d	MEB	Hull	MST	Furthest
Tier 1: low-rank 2D pixel coordinates					
MNIST raw pixel	2	4	5	–	–
MNIST + affine (TDOF-4)	2	5	–	–	–
MNIST 3-digit mixture	2	4	4	4	–
Fashion-MNIST pixel	2	–	4	–	**2**
Fashion-MNIST + affine	2	–	3	–	–
Fashion-MNIST 3-digit mix	2	–	–	4	–
CIFAR-100 raw pixel	2	–	3	–	5
Tier 2: medium-rank 8D feature-lift					
MNIST 8D polynomial lift	8	3	3	–	–
CIFAR-100 4×4 patch features	8	5	**2**	–	–
CIFAR-100 variable-size patches	8	5	–	4	–
CIFAR-100 class-mixture patches	8	5	–	–	–
Tier 3: high-rank deep feature					
CIFAR-100 ResNet-18 layer-3	256	(unstable)	–	5	**1**

The rank improvement of MZAttn-Full at Tier-3 (deep-feature space), relative to Tier-1 and Tier-2 where ranks are tied within seed noise, is the empirical signature the theory predicts: MZAttn's advantage requires data to have jointly (i) genuinely high-rank per-sample structure (not a 2D manifold embedded via a low-rank projection), (ii) sparse combinatorial targets (such as the 3–9 extreme points that determine MEB or the k=5 centers determining k-center cost), and (iii) the absence of a stronger spatial or equivariance prior that standard attention can exploit. On every cell where any one of these is absent, MZAttn-Full ranks 3rd–5th, not catastrophically worse, but also not ahead. The rank-1 cell appears only when all three conditions are constructively met. The synthetic reference point (mixture-of-Gaussians MEB at d=8, n=10–30, where MZAttn-Full reaches R^2=0.684 against baselines stuck at $R^2 \leq 0.023$; Table 3) reproduces exactly in the same code. The systematic grid therefore serves as the empirical scaffold on which the theoretical scope of the claim rests: it is not a set of failed transfer attempts but a controlled verification that the architectural advantage tracks the structural prerequisites of Theorem 1.

Positive transfer: CIFAR-100 ResNet-18 features. Guided by the diagnostic above, we construct a real-data setting satisfying the three prerequisites: (1) ResNet-18 (ImageNet-pretrained) layer-3 feature maps of CIFAR-100 images, adaptively average-pooled to spatial 4×4, giving n=16 genuinely-high-rank points in $\mathbb{R}^{256}$ per image (not a low-rank projection of pixel coordinates); (2) the

sparse combinatorial SGSP-family targets (MEB radius, hull volume, MST weight, diameter, k=5 k-center furthest cost, k=5 median k-NN radius); (3) no imposed symmetry. Table 32 summarises the 7-way comparison at 5 seeds per method per task.

Table 32: Real-world SGSP on CIFAR-100 ResNet-18 layer-3 features (n=16, d=256, R^2 mean ± std, bold = best). Seed counts differ by row: k-center furthest uses 5 seeds (file: `real_geom_cifar_resnet_furthest_5seed.json`); Diameter, k-NN radius, MST weight, MEB radius use 3 seeds (default seed pool $\{42, 123, 456\}$). The MEB cell at n=16 $\leq d$+1=257 is structurally degenerate (no sparse-combinatorial selector available); we mark it "unstable" and exclude it from the ranking claim. On the remaining four tasks, methods tie within ≤ 0.03 R^2, and MZAttn-Full ranks 1st on k-center furthest cost.

Task	Standard	ST-Large	HyperNet	Perceiver	LinearAttn	MZAttn-Full
Furthest (k=5)	0.789	0.793	0.792	0.746	0.797	**0.801**
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
MEB radius	0.800	0.815	0.783	0.728	**0.817**	0.542 (unstable)

Across the full SGSP family on real ResNet-18 features, LinearAttn ranks first on three of the four valid cells (diameter, k-NN radius, MST weight; the MEB cell is structurally degenerate at n=16 $\leq d$+1 and is excluded from the ranking claim), and the gap between the top and bottom attention-based method is narrow (≤ 0.03 R^2) on these dense-statistical targets. Aggregating many pairwise distances reduces these targets to dense statistics that context-rigid attention readily extracts, so the absence of an MZ advantage there is consistent with the TDOF theory rather than evidence against it. Only the sparsest combinatorial target, k-center furthest cost, yields a ranking on which MZAttn-Full is first among the seven parameter-matched baselines in both a 5-seed and an independent 10-seed verification run. We are explicit that the ranking, not the magnitude, is the evidence at the smaller backbone: 5-seed 0.801 vs. LinearAttn 0.797 and 10-seed 0.797 vs. 0.796 are both within pooled seed standard deviation. The cleaner statistical separation arrives at the larger backbone (Section 5, ResNet-50 d=1024, Welch's $t \approx 5.4$, $p < 0.005$); the ResNet-18 result establishes that the rank ordering is stable across two independent seed pools, a necessary but weaker form of evidence.

Dimension scaling: ResNet-50 features (d=1024). Table 33 reports the seven-baseline dimension-scaling comparison. The reported ResNet-50 numbers use patience 300 to give every model ample budget to converge at d=1024, and the patience sweep in Appendix D.4.7 confirms the MZAttn-Full > ST-Large ∼ HyperNet > LinearAttn ranking is stable across patience $\{20, 100, 300\}$ at 80 training epochs. Per-method standard deviations are: MZAttn-Full 0.7955 ± 0.0009, MZAttn-Large 0.7930 ± 0.0024, ST-Large 0.7909 ± 0.0017, with MZAttn-Full also the most stable method in the comparison. The absolute magnitude of the d=1024 gap is small (+0.0046 R^2, equivalent to a 5.5% reduction in $1-R^2$ residual error) but reproduces across 5 seeds at Welch $p \approx 0.0018$.

What these results show. MZAttn's gain is structural: it bites when the target operator genuinely varies per sample in a high-rank subspace and depends on a sparse combinatorial subset of the input. When the target is governed by spatial locality (image pixels), smooth interpolation (Neural-Process regression), or a known symmetry (equivariant tasks), MZAttn's per-sample operator prior is not the right tool. Real high-rank deep-feature point clouds with a sparse combinatorial target

Table 33: Dimension scaling on the k-center furthest target (k=5, n=16, R^2 mean over 5 seeds). ResNet-50 numbers use early-stopping patience 300 for full convergence at d=1024; a patience sweep $\{20, 100, 300\}$ confirms ranking stability for any patience ≥ 100 (Appendix D.4.7).

Feature	Standard	ST-Large	HyperNet	Perceiver	LinearAttn	MZAttn-Large	MZAttn-Full
ResNet-18 (d=256)	0.789	0.793	0.792	0.746	0.797	0.789	**0.801**
ResNet-50 (d=1024)	0.789	0.791	0.789	0.784	0.789	0.793	**0.796**

(CIFAR-100 ResNet features, k-center furthest cost) do satisfy the prerequisites, and MZAttn-Full wins there. The real-world transfer is therefore architecturally precise rather than universal, in line with the TDOF theory.