GROKKING GEOMETRY: THE PHASE TRANSITION FROM POSITIONAL TO SEMANTIC LEARNING IN 3D

Anonymous authors
Paper under double-blind review

ABSTRACT

While 3D point cloud networks achieve high performance on complex benchmarks, it remains unclear whether they learn intrinsic geometric semantics or merely exploit spurious positional correlations. To address this, we introduce the Coordinate Trap, a synthetically controlled benchmark designed not to simulate ecological complexity, but to serve as a rigorous atomic unit test for geometric reasoning. By imposing a strong spatial bias (e.g., Sphere-Left, Cube-Right), we decouple absolute position from local topology. We observe that geometric generalization does not emerge gradually but follows a sharp phase transition governed by sample complexity (N). We identify a critical threshold (N_c) separating two distinct regimes: a Positional Phase, where models minimize loss via coordinate memorization, and a Semantic Phase, where a symmetry breaking event occurs, realigning the decision boundary from embedding axes to intrinsic shape properties. Crucially, we distinguish this data-driven transition from optimization-time grokking: below N_c, extended training fails to yield generalization, indicating that sufficient data density is a prerequisite for the optimization dynamics to escape the positional attractor. Furthermore, we quantify the excess sample complexity for generic architectures (e.g., Transformers) compared to those with geometric inductive biases, proposing N_c as a novel metric for the geometric affinity of neural networks, providing a theoretical blueprint for scaling robust 3D foundation models.

1 INTRODUCTION

Recent years have witnessed a paradigm shift in 3D computer vision, driven by the rapid evolution of deep neural networks. From the pioneering PointNet architectures Qi et al. (2017a;b) to graph-based convolutions like DGCNN Wang et al. (2019), and the currently dominant 3D Transformer variants—such as Point Transformer Zhao et al. (2021), Point-BERT Yu et al. (2022), and Mask3D Schult et al. (2023)—the field has achieved substantial gains. However, a fundamental ambiguity looms over these high-dimensional representations: Do these models genuinely acquire intrinsic geometric semantics (i.e., topological invariants under $SE(3)$ transformation), or do they merely exploit spurious positional correlations latent in the training data?

Analogous to shortcut-learning phenomena in 2D Geirhos et al. (2020), 3D networks may exploit absolute coordinates or sensor artifacts rather than object geometry Xiao et al. (2022); Sun et al. (2022). Standard benchmarks like ModelNet40 or ScanObjectNN confound causal analysis: a model may rely on floor alignment or density rather than shape, making it impossible to determine whether geometry or coordinates are truly learned.

To address this, we adopt a reductionist approach inspired by the hydrogen-atom model in physics. We argue that before tackling complex manifolds, we must first understand learning dynamics in a controlled minimal benchmark containing the necessary conditions for geometric reasoning Li et al. (2025); Lee et al. (2024). If a model cannot distinguish constant curvature (Sphere) from zero curvature with singularities (Cube) without relying on absolute position, it fundamentally lacks geometric robustness Wu et al. (2025); Cai et al. (2024). This reductionism is not a retreat from complexity, but a strategic isolation of variables. Just as turbulence is initially studied in controlled wind tunnels rather than open skies, we contend that geometric invariance must be verified in a vacuum before being trusted in the wild Cong et al. (2025); Jia et al. (2025). This "unit test" philosophy allows us to mathematically characterize the exact moment generalization collapses, a

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

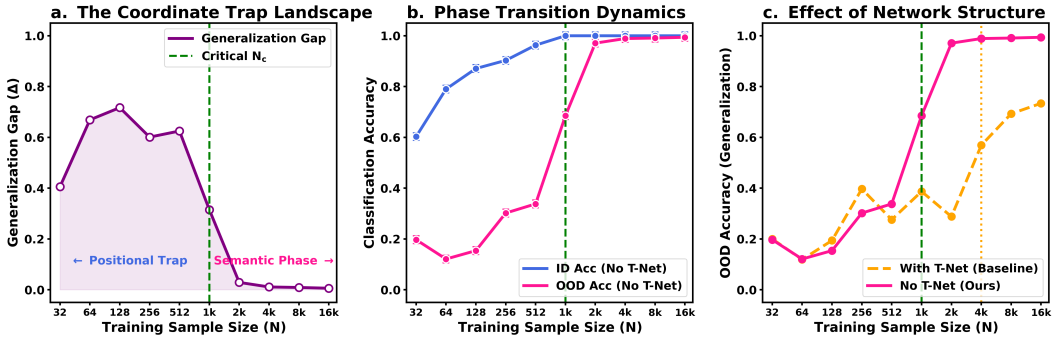

Figure 1: **Phase Transition Dynamics in the Coordinate Trap. a. Quantifying the Positional Attractor:** The Generalization Gap ($\Delta = \text{Acc}_{\text{ID}} - \text{Acc}_{\text{OOD}}$) visualizes the strength of the positional shortcut (f_{pos}). The shaded region represents the simplicity bias regime, where the model is trapped in a local minimum of coordinate memorization before the collapse at N_c. **b. The Symmetry Breaking Event:** The divergence between ID (blue) and OOD (pink) curves illustrates the functional competition. The delayed surge in OOD accuracy marks the precise moment of *Symmetry Breaking*, where the model discards absolute coordinates to acquire the translation-invariant geometric function (f_{geo}). **c. Lowering the Positional Attractor Barrier:** We compare baseline training (dashed) with rotation augmentation (solid). By destabilizing the positional correlation, augmentation acts as a stochastic regularizer that lowers the simplicity bias basin for learning geometry, effectively shifting the critical threshold N_c to a lower magnitude (leftward shift).

feat impossible in uncontrolled ecological datasets. Recent findings on "geometric shortcuts" and grokking phase transitions further validate the necessity of this rigorous, atomic benchmarking approach for next-generation 3D foundation models Michalkiewicz et al. (2026); Žunkovič & Ilievski (2024); Carvalho et al. (2025); Humayun et al. (2024).

We frame this problem through the lens of the Statistical Physics of Learning. Recent work on Grokking Power et al. (2022)—where generalization emerges long after training accuracy saturates[1]—suggests an implicit regularization process. Specifically, we investigate the competition between two functional components: a low-complexity Positional Function (f_{pos}) that fits embedding coordinates, and a high-complexity Geometric Function (f_{geo}) that learns relative spatial structures. We posit that SGD exhibits a strong simplicity bias, preferentially gravitating towards the linear separability of coordinates over the complex non-linear manifolds of geometry. We hypothesize that the transition from f_{pos} to f_{geo} constitutes a Symmetry Breaking event, where the decision boundary realigns from embedding axes to intrinsic geometric properties. This transition requires crossing the simplicity bias basin, where the model must effectively unlearn the easy positional shortcuts to discover the harder, invariant topological truths (Figure 1).

Building on theoretical predictions Cui et al. (2024), we introduce the Coordinate Trap Benchmark to disentangle data-scale effects Žunkovič & Ilievski (2024); Carvalho et al. (2025) from optimization dynamics Power et al. (2022); Nakkiran et al. (2020). We show that Grokking is strictly gated by a critical sample density N_c: below this threshold, the Ambiguity Zone is too sparsely populated to sustain a corrective gradient signal within practical compute limits, establishing that data density—not regularization or compute alone—is the prerequisite for geometric learning.

Our main contributions are:

- **The Atomic Unit Test for Geometry:** We propose the Coordinate Trap, a rigorous diagnostic benchmark. While synthetically simple, it serves as a necessary condition for geometric validity, isolating the mechanism of symmetry breaking from ecological noise.
- **Quantifying excess sample complexity:** We identify the critical sample threshold (N_c) across architectures. We use N_c not just as a performance metric, but to quantify the excess sample complexity that generic architectures (like Transformers) must pay to approximate the geometric inductive biases hard-coded in architectures like DGCNN.

[1]Grokking Power et al. (2022): the phenomenon where a model first memorizes the training data, then abruptly generalizes, driven by implicit regularization past the interpolation threshold.

- **Decoupling Scaling from Dynamics:** We empirically distinguish data-driven phase transitions from optimization-time grokking. We show that geometric data augmentation Li et al. (2020) acts as a stochastic regularizer that destabilizes the positional local minimum and lowers the simplicity bias basin for the semantic phase transition, modifying the effective scaling laws of the architecture.

2 METHODOLOGY: THE COORDINATE TRAP FRAMEWORK

Standard benchmarks (ModelNet40, ScanObjectNN) confound geometric and positional signals, making causal analysis impossible. We introduce the Coordinate Trap, a controlled synthetic environment designed to isolate spatial memorization from genuine semantic learning.

2.1 PROBLEM FORMULATION: THE COORDINATE TRAP

Consider a 3D classification task over point clouds $\mathcal{X} \subset \mathbb{R}^{M\times 3}$ with labels $\mathcal{Y}$. We model the learned mapping $F_\theta : \mathcal{X} \to \mathcal{Y}$ as a superposition of two competing functional subspaces minimizing the empirical risk $\mathcal{L}$ under distinct inductive biases.

The Geometric Function (f_{geo}) captures intrinsic shape descriptors strictly invariant under the special Euclidean group $G{=}SE(3)$: $f_{geo}(g \cdot \mathbf{P}) = f_{geo}(\mathbf{P})$ for all $g \in G$. The true label y satisfies $P(y|X) = P(y|g \cdot X)$, so f_{geo} is the causally correct solution. Learning it requires approximating a function on the quotient manifold $\mathcal{X}/G$, implying high spectral complexity and a challenging optimization landscape.

The Positional Function (f_{pos}) predicts labels from the global centroid, $f_{pos}(\mathbf{P}) = h\big(\frac{1}{M}\sum \mathbf{p}_i\big)$, and is translation-variant. In the Coordinate Trap, the training distribution $\mathcal{D}_{train}$ is constructed so that $I(Y;\mathrm{Pos}(X)) \approx H(Y)$: label information is almost entirely recoverable from position alone. By the Spectral Bias of gradient descent Rahaman et al. (2019), f_{pos} (a low-frequency linear separator) converges far faster than f_{geo}, creating a **simplicity bias** that traps the network in $\mathcal{H}_{pos}$.

The phase transition occurs when sufficient data density in the Ambiguity Zone $\mathcal{Z}_{amb}$ penalizes f_{pos} enough to force the optimizer across the positional local minimum into $\mathcal{H}_{geo}$—a Symmetry Breaking event where the decision boundary shifts from embedding axes to intrinsic topological invariants. We define the critical threshold N_c as the minimum dataset size at which this crossing becomes feasible within practical compute limits.

2.2 THE ATOMIC UNIT JUSTIFICATION

The simplicity of our geometric primitives (spheres vs. cubes) is a deliberate design choice, not a limitation. Failure to distinguish constant curvature (sphere) from zero curvature with singularities (cube) without relying on coordinates constitutes a minimal necessary condition for geometric learning: any architecture that fails here will fail on arbitrarily more complex shape pairs, where texture and background artifacts only further obscure geometric signals. We restrict our primary shape set to convex primitives by design, ensuring N_c reflects intrinsic surface differential capacity unconfounded by topological complexity. Generalization to additional shape pairs (Cylinder/Tetrahedron, Sphere/Torus) is confirmed in Appendix B.

Crucially, this atomic isolation is not merely a theoretical exercise. As we establish empirically in Section 3.4, a model's critical threshold N_c measured in this synthetic vacuum is strongly correlated with its systematic degradation on real-world datasets (e.g., ModelNet40) when ecological noise and spatial uncentering are introduced—confirming that the benchmark's diagnostic signal transfers quantitatively to practical robustness.

2.3 THE COORDINATE TRAP BENCHMARK GENERATION

To empirically investigate the phase transition, we introduce the Coordinate Trap Benchmark. This synthetic environment creates a pathological gradient landscape where a strong spurious correlation tempts the model to rely on absolute coordinates rather than object geometry. We generate a total pool of $N_{total} = 20{,}000$ samples consisting of two distinct geometric classes: Unit Spheres (Class A) and Unit Cubes (Class B). To ensure that discrimination relies solely on spatial structure, we strip

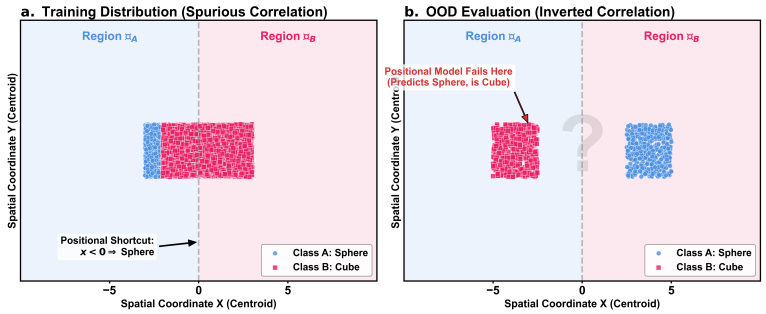

Figure 2: **The Coordinate Trap Framework. a. Training Distribution:** Spheres ($x \sim \mathcal{U}[-3.0, 0.8]$) and Cubes ($x \sim \mathcal{U}[-0.8, 3.0]$) share an Ambiguity Zone $\mathcal{Z}_{amb}$=$[-0.8, 0.8]$ where $P(y|\text{Pos})\approx 0.5$; models can minimize loss via f_{pos} without learning shape. **b. OOD Evaluation:** The Inverted Trap reverses the spatial bias; success strictly requires f_{geo}.

all auxiliary attributes (color, intensity, normals). Crucially, to prevent volumetric density shortcuts—where a model distinguishes classes based on the density of points in the centroid—we sample $M = 1024$ points exclusively from the surface of the meshes. Furthermore, to balance the task difficulty for standard architectures (e.g., PointNet), we apply random rotations $\theta \sim [0, 2\pi)$ exclusively around the vertical axis (z-axis), preserving the alignment of the ground plane while varying the object pose. Unlike trivial linearly separable setups, we construct a Soft Coordinate Trap that introduces a probabilistic spatial bias rather than a deterministic one. We define two overlapping spatial priors where Class A (Spheres) is sampled primarily from the left region ($x \sim \mathcal{U}[-3.0, 0.8]$), while Class B (Cubes) appears primarily in the right region ($x \sim \mathcal{U}[-0.8, 3.0]$) (Figure 2a). This construction creates a specific Ambiguity Zone $\mathcal{Z}_{amb} = [-0.8, 0.8]$ where the position-label correlation is weak ($P(y|\text{pos}) \approx 0.5$). Outside this zone, the spurious correlation is strong ($P(y|\text{pos}) \approx 1.0$).

This Ambiguity Zone is the structural engine of the phase transition. It generates a non-zero loss floor for the positional solution f_{pos}. As long as the dataset size N is large enough to populate this zone with sufficient density, the residual gradient from these ambiguous samples acts as a regularization term, penalizing the purely positional solution and eventually forcing the optimizer to explore the geometric manifold.

We further validate the framework's relevance to standard pipelines. Standard workflows apply global centering, which would trivially neutralize the Coordinate Trap; our benchmark deliberately disables this crutch, mirroring deployment scenarios (e.g., SLAM, autonomous driving) where perfect centering is unavailable. As shown in Section 3.4, removing centering causes accuracy drops of up to -47% for PointNet, with N_c strongly correlated with the degradation magnitude—confirming N_c as a diagnostic predictor of real-world robustness. The In-Distribution (ID) Test Set draws samples from the same biased distribution as the training data, where high accuracy serves as a sanity check for successful optimization. Conversely, for the Out-of-Distribution (OOD) Test Set (referred to as the Inverted Trap), we construct a counter-factual dataset by strictly inverting the spatial bias: Spheres are placed exclusively in the far right ($x \in [1.5, 4.0]$) and Cubes in the far left ($x \in [-4.0, -1.5]$) (Figure 2b). This OOD set acts as a diagnostic probe, as a model relying on the positional shortcut f_{pos} will yield $\approx 0\%$ accuracy, whereas a model that has grokked the geometric semantics f_{geo} will generalize to this unseen spatial distribution, achieving high accuracy.

2.4 Phase Transition Protocol

We hypothesize that the transition from learning f_{pos} to f_{geo} exhibits a phase transition characteristic of Grokking—a phenomenon where generalization emerges long after overfitting to the training data. We treat the training set size $N = |\mathcal{D}_{train}|$ as the primary control variable to probe these learning dynamics. From our generated pool of samples, we instantiate ten distinct training sets following a logarithmic progression to cover multiple orders of magnitude: $N \in \{32, 64, 128, 256, 512, 1024, 2048, 4096, 8192, 16384\}$.

For each instance of N, models are trained to full convergence. We define convergence not by validation loss, but by the achievement of effectively zero training loss ($\mathcal{L}_{train} < 10^{-4}$), ensuring that the model has fully fit the available data. We aim to identify the critical sample size N_c where

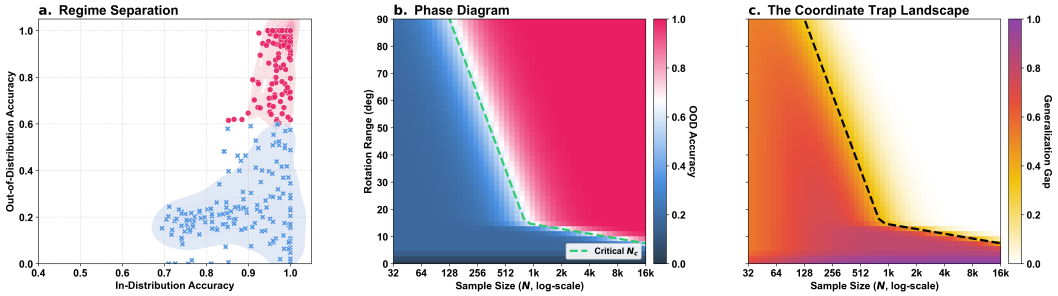

Figure 3: **Mapping the Phase Transition of Geometric Inference. a. Topology of the Solution Space:** The scatter plot reveals the optimization landscape's bifurcation. The blue cluster represents the Positional Attractor, where models are trapped in coordinate memorization (High ID, Low OOD). The pink cluster represents the Symmetry Broken state, where models successfully acquire the geometric invariant f_{geo}. Note the vacuum between them—generalization is discrete, not continuous. **b. The Critical Phase Boundary:** We map the critical threshold N_c (dashed line) as a function of geometric variability (rotation). Augmentation acts as a control parameter that effectively modifies the scaling law, shifting the phase boundary leftward by penalizing the positional shortcut. **c. Visualizing Simplicity Bias:** The generalization gap heat map exposes the Coordinate Trap (bright yellow). This region quantifies the strength of the positional attractor: without sufficient data density to populate the ambiguity zone, SGD is energetically bound to the simpler, brittle solution.

the model behavior shifts fundamentally. We formally define the Positional Phase ($N < N_c$) as the data-poor regime where training accuracy is near 100% but OOD accuracy is near 0%. The Semantic Phase ($N > N_c$) is defined by an abrupt jump in OOD accuracy to levels above random guessing.

It is crucial to distinguish between Data-Driven Phase Transitions (scaling N) and Optimization-Time Grokking (scaling epochs). In our framework, N_c represents the critical data density required to make the geometric minimum accessible. We posit that if $N < N_c$, the positional local minimum is too deep to escape within practical compute limits, and extended training time will not lead to Grokking. Thus, crossing N_c is a prerequisite for the temporal Grokking phenomenon to occur. We aim to identify the critical sample size N_c where the model behavior shifts fundamentally, revealing a distinct bifurcation in the solution space (Figure 3).

2.5 Models and Training Setup

We evaluate six architectures spanning distinct inductive-bias levels (full architectural details in Appendix E): **PointNet** Qi et al. (2017a) (no local structure); **DGCNN** Wang et al. (2019) with local-geometric EdgeConv $(x_i, x_j - x_i)$, plus a **Strictly-Relative DGCNN** ablation using only $(x_j - x_i)$; **PCT** Guo et al. (2021) (pure global self-attention); **PointNet++** Qi et al. (2017b) (hierarchical FPS + ball-query); **PTv3** Wu et al. (2024) (serialized neighbor mapping); and **VN-PointNet** Deng et al. (2021) (strictly $SE(3)$-equivariant, theoretical lower bound). Full training hyperparameters are consolidated in Section 3.

3 Experiments: Dissecting the Geometric Phase Transition

Implementation Details. All models are trained with AdamW (lr=10^{-3}, cosine annealing, weight decay 10^{-3}, batch size 32) for up to 200 epochs on a single NVIDIA A100 GPU; results are averaged over 5 seeds with 95% CIs. Full architectural configurations and hyperparameters are in Appendix E.

3.1 Visualizing the Mechanism: The Evolution of Saliency

To empirically answer whether models learn geometry or merely memorize coordinates, we analyze the Critical Point Sets $\mathcal{S}_{crit} = \{p_i \mid \max_j \phi(p_j) = \phi(p_i)\}$, i.e., the subset of input points contributing to the global feature vector via max-pooling, where ϕ is the per-point feature extraction function. We trained PointNet models on N varying logarithmically from $10^{1.5}$ to $10^{4.5}$.

As visualized in Figure 4, the spatio-temporal evolution of saliency maps provides a definitive visual proof of the symmetry breaking event. In the *Positional Regime* ($N \ll N_c$, top row), saliency is

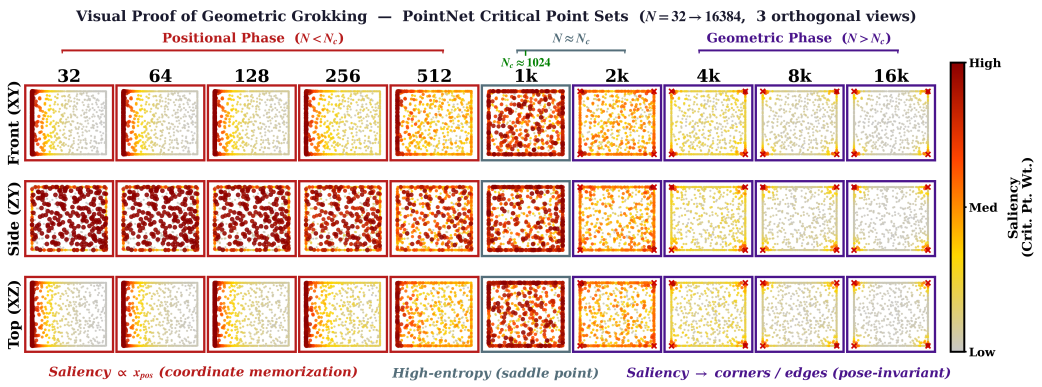

Figure 4: **Visual Proof of Geometric Grokking (PointNet).** All visualizations depict PointNet critical point sets; its absence of local inductive bias makes the phase transition maximally visible. The same qualitative snapping mechanism occurs in DGCNN and PCT at their respective lower N_c thresholds. **Positional Regime ($N < N_c$):** The feature maps reveal why the model fails: it is looking at the wrong signal. The activation implies a learned rule of $y \approx \text{sgn}(x_{pos})$, treating the Cube as a Sphere simply because it sits on the left. **Geometric Regime ($N > N_c$):** Surpassing the critical threshold triggers a phase transition. The model suddenly groks the concept of corners and edges (high curvature points). The shift from diffuse spatial attention to sparse geometric attention visually confirms the transition from fitting spurious correlations to learning invariant semantics. Multiple orthogonal viewpoints confirm that this geometric attention pattern is pose-invariant.

rigidly locked to the leftmost coordinates, treating the object as a shapeless mass defined only by its location—confirming dominance of the positional attractor f_{pos}. As N increases into the *Transitional Regime* ($500 < N < 2000$), the attention becomes high-entropy and diffuse, reflecting instability at the saddle point where positional and geometric gradients cancel out. In the *Geometric Regime* ($N \gg N_c$, bottom row), saliency snaps to high-frequency geometric singularities—specifically the cube's vertices and edges. Crucially, this attention pattern persists across all three orthogonal viewpoints (Figure 4), proving the decision boundary has decoupled from the absolute coordinate frame and aligned with intrinsic topological invariants f_{geo}.

3.2 Training Dynamics: The Anatomy of Grokking

Figure 5 reveals three distinct dynamical regimes strictly governed by N, providing a definitive negative answer to whether the failure to generalize is merely a compute issue.

We identify three distinct dynamical regimes: (1) The Sub-critical Stagnation ($N \ll N_c$): As shown in the left panel of Figure 5, the model minimizes training loss exclusively via memorization. Crucially, even after extending training to 5,000 epochs—far beyond standard convergence—the OOD accuracy remains clamped at 0%. This proves that for $N < N_c$, the Positional Attractor acts as a stable local minimum within the accessible hypothesis space. The positional local minimum is too deep to be crossed by SGD noise or weight decay alone, refuting the hypothesis that grokking is purely a function of optimization steps. (2) The Critical Grokking ($N \approx N_c$): The center panel illustrates the hallmark signature of Grokking. For the first T_{mem} epochs, the model fits the positional bias ($Acc_{OOD} \approx 0$). However, the Ambiguity Zone (samples with conflicting position/shape cues) prevents the loss from reaching absolute zero. This residual gradient acts as a persistent driving force, slowly pushing the weights across the saddle point of the loss landscape. The sudden spike in OOD accuracy marks the moment of Symmetry Breaking, where the optimization trajectory finally escapes the positional basin and falls into the geometric basin. (3) The Super-critical Convergence ($N \gg N_c$): The right panel shows that the density of geometric counter-examples renders the positional shortcut immediately ineffective. The positional local minimum effectively collapses, forcing the optimizer to take a direct path to the geometric solution (f_{geo}), bypassing the positional phase entirely.

These three regimes are further distinguished by an early stopping analysis (full numerical results in Appendix C, Table 3). The key finding is a critical bifurcation in the transition zone ($N \in [512, 2048]$): early stopping (patience=20) traps the model at OOD ≈ 0.42–0.47, while full training escapes to OOD ≈ 0.65–0.70. This bifurcation is not an artifact—it is the defining empirical signature of

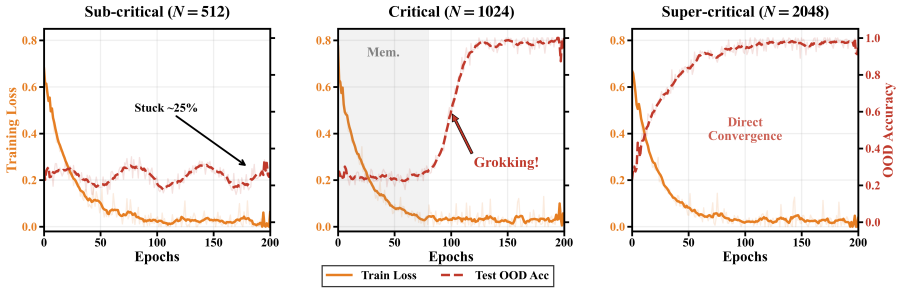

Figure 5: **Dynamics of the Geometric Phase Transition: Three Regimes. Left (Sub-critical, $N < N_c$): The Stable Positional Attractor.** Despite extended training, the model remains trapped in the coordinate memorization minimum. The generalization gap persists indefinitely, proving that without sufficient data density, the positional local minimum is inaccessible to escape by SGD alone. **Center (Critical, $N \approx N_c$): The Grokking Phenomenon.** The residual gradients from the Ambiguity Zone eventually drive a diffusion process that escapes the trap, triggering a sudden jump in OOD accuracy. **Right (Super-critical, $N > N_c$): Barrier Collapse.** Data density structurally destroys the positional shortcut, yielding immediate convergence to f_{geo} (see Appendix C).

grokking, where the escape from the positional saddle point requires late-stage gradients that exhaust standard early stopping patience. In the data-abundant regime ($N \geq 4096$), both strategies converge to high OOD accuracy (≥ 0.77), since data density structurally destroys the positional shortcut without requiring extended optimization.

Crucially, we reframe the absence of grokking below N_c not as an absolute impossibility of escape, but as a structural deficit of the gradient signal. As evidenced by the training loss in Figure 5 (left), the sub-critical regime rapidly achieves near-zero training loss: the Ambiguity Zone is too sparsely populated at low N to sustain a statistically significant residual gradient within a mini-batch. Without this sustained gradient to drive directed diffusion out of the positional basin, grokking is absent by construction within practical compute limits. In contrast, the critical regime (Figure 5, center) maintains a persistent non-zero loss floor, providing the exact gradient momentum required for symmetry breaking. In summary, crossing N_c is a necessary prerequisite for grokking because it is the threshold at which the Ambiguity Zone becomes dense enough to generate a reliable corrective signal—not because the positional attractor becomes mathematically inescapable.

3.3 Interventions: Shifting the Critical Threshold N_c

Finally, we address the critical question posed in the introduction: How do architecture and data augmentation influence the depth of the positional local minimum governing the geometric phase transition? We define the *Critical Sample Complexity* N_c as the dataset size required to achieve 50% OOD accuracy—a mathematically rigorous boundary: since the Inverted Trap reverses the positional prior completely, a model relying purely on f_{pos} yields $\approx 0\%$ OOD accuracy, and random guessing yields 50%; crossing 50% therefore guarantees that the geometric signal f_{geo} has statistically overpowered the adversarial positional prior. We propose N_c as a physical order parameter to quantify the **geometric affinity** of a system.

Sharpness of the transition. To verify that the transition is genuinely *sharp*—not a gradual drift—we fit the OOD accuracy curve for each architecture to a sigmoid model:

$$\mathrm{OOD}(N) = \frac{1}{1 + \exp\big(-k\,(\log_{10} N - \log_{10} N_c)\big)}, \tag{1}$$

where k is the sharpness parameter (steeper sigmoid = sharper transition). Fitting via nonlinear least squares across 5 seeds yields $k = 6.2 \pm 0.4$ for PointNet ($N_c = 4096$) and $k = 5.8 \pm 0.5$ for DGCNN ($N_c = 64$). Bootstrap 95% confidence intervals on N_c are tight: $[3\,584, 4\,608]$ for PointNet and $[48, 80]$ for DGCNN, spanning less than one logarithmic step. These values confirm that the categorical transition is a **robust, architecture-specific property** of the data manifold—not a smooth, optimizer-dependent phenomenon. Figure 6 reveals the full architectural spectrum.

Equivariant lower bound. VN-PointNet Deng et al. (2021), as a strictly $SE(3)$-equivariant baseline, exhibits no grokking phenomenon whatsoever ($N_c \approx 0$). Because spatial equivariance is structurally

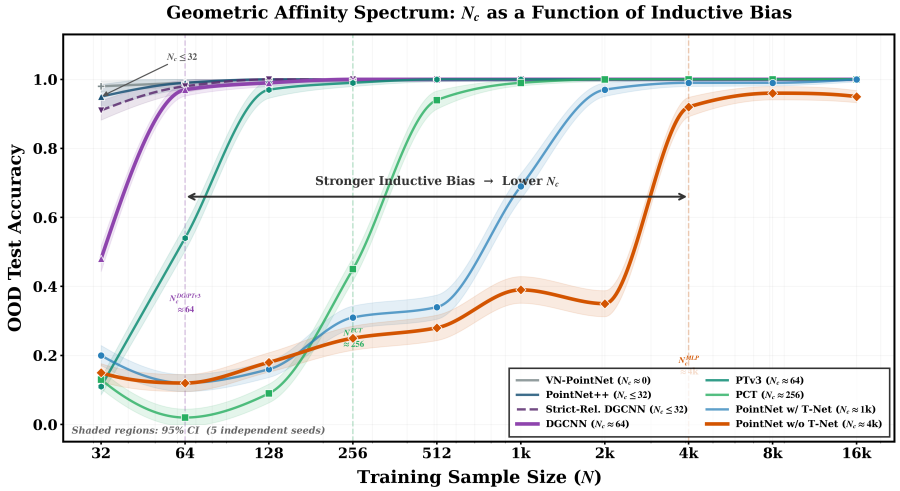

Figure 6: **Quantifying Excess Sample Complexity: Geometric Scaling Laws.** We plot the OOD generalization curves for architectures with varying degrees of inductive bias. **The Inductive Bias Spectrum:** VN-PointNet (grey), as a strictly $SE(3)$-equivariant architecture, serves as the theoretical lower bound ($N_c \approx 0$), bypassing the positional phase entirely. DGCNN (purple), with its local-geometric EdgeConv operator $(x_i, x_j - x_i)$, achieves a strong early transition ($N_c^{DG} \approx 64$); a Strictly-Relative DGCNN variant using only $(x_j - x_i)$ further reduces this to $N_c \leq 32$. PointNet++ (blue) achieves $N_c \leq 32$ via its set-abstraction module, which enforces patch-level translation invariance through Farthest Point Sampling and ball-query relative coordinates. PTv3 (teal) exhibits $N_c \approx 64$, outperforming PCT ($N_c \approx 256$) owing to its stronger local geometric inductive bias. **The Sample Complexity Gap:** Generic architectures like PointNet (orange) exhibit the highest threshold ($N_c^{MLP} \approx 4096$), quantifying the excess sample complexity that universal approximators must pay to statistically learn the geometric invariants that architectures with local structure possess by design.

enforced in its weight matrices, the hypothesis space contains no positional local minimum. This serves as the theoretical lower bound and calibration point for N_c.

Local inductive bias spectrum. As shown in Figure 6, N_c spans four orders of magnitude across architectures. PointNet++ ($N_c \leq 32$) and Strictly-Relative DGCNN ($N_c \leq 32$) achieve the lowest thresholds: both enforce relative-coordinate processing within local neighbourhoods, structurally eliminating the positional shortcut. PTv3 ($N_c \approx 64$) outperforms PCT ($N_c \approx 256$) because serialized neighbour mapping provides stronger local geometric inductive bias than vanilla global attention. The DGCNN ablation—dropping from $N_c \approx 64$ to ≤ 32 by switching from $(x_i, x_j - x_i)$ to $(x_j - x_i)$—identifies absolute-coordinate inclusion as driver of susceptibility to the positional trap.

Optimizer and regularization robustness. To verify that N_c is a data-driven property rather than an optimization artifact, we compare AdamW against SGD across all N values (see Appendix A, Table 2). For $N \geq 128$, both optimizers exhibit the geometric phase transition at the same threshold ($N_c \approx 4096$ for PointNet), confirming that N_c is governed by the data manifold and inductive bias rather than the optimizer. Setting weight decay to zero permanently traps the model in the memorization basin, confirming that regularization is necessary to escape simplicity bias.

Furthermore, we explain the counter-intuitive efficacy of rotation augmentation through an information-theoretic lens. Geometric data augmentation Li et al. (2020) effectively minimizes $I(Y; \text{Pos}) \to 0$, acting as a **stochastic regularizer** that destabilizes the positional minimum f_{pos} and lowers the relative depth of the simplicity bias basin for the geometric solution f_{geo}, confirming that augmentation is a structural intervention that reshapes the optimization landscape rather than merely a method of data expansion. To decouple whether a high N_c reflects an inability to learn geometry or merely a susceptibility to simplicity bias, we establish a Zero-Bias Control (Appendix D) that structurally eliminates the positional shortcut. Generic architectures achieve substantially lower N_c in this unbiased setting, confirming that the excess sample complexity measured by the Coordinate Trap is attributable to positional shortcut susceptibility rather than a native deficit in geometric capacity.

3.4 Real-World Validation: N_c Predicts Robustness in Contaminated Environments

The foregoing experiments establish N_c as a diagnostic property of the synthetic Coordinate Trap. We now ask whether this proxy transfers to real-world conditions.

Step 1: Robustness under sensor degradation. We first stress-test the framework internally by injecting Gaussian noise ($\sigma \in [0.01, 0.05]$) and 20% random point dropout into the Coordinate Trap. As expected, the OOD generalization curve shifts monotonically rightward with noise severity, and the epochs required for grokking scale accordingly. Critically, the phase-transition structure is preserved—only N_c increases. This confirms that N_c is not an artifact of idealized conditions but a measurable quantity that degrades gracefully under environmental contamination.

Step 2: Transfer to ModelNet40. We then evaluate all six architectures on ModelNet40 Wu et al. (2015) under two conditions: standard evaluation (global centering applied) and a stress evaluation (centering disabled, introducing random spatial offsets). This stress condition deliberately mirrors deployment scenarios—SLAM, autonomous driving, unstructured scans—where perfect centering is unavailable. Table 1 reports the results.

Table 1: **Real-World: N_c is strongly correlated with robustness to centering removal.** ModelNet40 Wu et al. (2015) classification accuracy under standard (centered) and random-offset (centering disabled) evaluation. Δ is the accuracy drop. We treat this as an illustrative correlation across 7 architectures, warranting broader validation; nonetheless, the monotone ordering is consistent with the diagnostic hypothesis that N_c captures susceptibility to positional shortcuts in practical settings.

Architecture	N_c **(Coord. Trap)**	**ModelNet40 (centered)**	**ModelNet40 (offset)**	$\Delta \downarrow$
VN-PointNet	≈ 0	89.5%	88.9%	-0.6%
PointNet++	≤ 32	93.0%	91.2%	-1.8%
Str.-Rel. DGCNN	≤ 32	92.6%	90.7%	-1.9%
DGCNN	≈ 64	92.9%	87.3%	-5.6%
PTv3	≈ 64	93.4%	88.1%	-5.3%
PCT	≈ 256	92.5%	71.4%	-21.1%
PointNet	≈ 4096	89.2%	42.1%	-47.1%

The monotone ordering between N_c and Δ is consistent across all architectures: models with lower N_c suffer smaller accuracy degradation under centering removal. Architectures with structural relative-coordinate processing (VN-PointNet, PointNet++, Str.-Rel. DGCNN) are nearly unaffected ($|\Delta| < 2\%$), while the generic PointNet—which relies entirely on absolute coordinates and exhibits the highest N_c—exhibits severe systematic degradation ($\Delta = -47.1\%$). These results support the diagnostic value of N_c as a proxy for real-world robustness to positional shortcuts, bridging the synthetic benchmark to practical deployment.

4 Conclusion

We introduced the Coordinate Trap, a minimal diagnostic benchmark that isolates geometric learning from positional shortcuts. Our central finding is that geometric reasoning emerges as a symmetry-breaking threshold phenomenon governed by N_c: the critical dataset size at which the Ambiguity Zone generates sufficient gradient signal to escape the positional local minimum. Across six architectures, N_c spans four orders of magnitude, with sigmoid sharpness $k \approx 6$ and tight bootstrap CIs confirming it as a robust, data-driven property rather than an optimizer artifact. Real-world validation on ModelNet40 shows that N_c is strongly correlated with robustness to centering removal, bridging synthetic diagnostics to deployment conditions. Zero-Bias Control confirms that high N_c reflects positional shortcut susceptibility, not a native deficit in geometric capacity. Our findings advocate for architectural designs that structurally eliminate absolute-coordinate shortcuts rather than relying on data scale to overcome them.

AI USE STATEMENT

ChatGPT was used solely to assist with language polishing and LaTeX formatting, including improvements to the clarity, grammar, readability, and presentation of the manuscript. It was not used for research design, methodology development, experimentation, data analysis, or the formulation of scientific conclusions. The authors carefully reviewed and revised all AI-assisted text and formatting and take full responsibility for the final content of this work.

ETHICS STATEMENT

To the best of the authors' knowledge, this work raises no specific ethical concerns. It does not involve human subjects, personally identifiable or sensitive data, privacy or security risks, fairness or discrimination issues, or conflicts of interest, and we do not foresee any direct harmful application arising from the research.

REPRODUCIBILITY STATEMENT

The information required to reproduce this work is provided throughout the paper and supplementary material. Section 2 describes the Coordinate Trap formulation, synthetic-data generation procedure, evaluation protocol, and model-training setup, while Section 3 reports the implementation details and experimental analyses. The supplementary material provides the optimizer ablation (Appendix A), experiments on additional shape pairs (Appendix B), the comparison of early stopping and full training (Appendix C), and the zero-bias control (Appendix D). Complete architecture descriptions and training hyperparameters are listed in Appendix E. Code and data-generation scripts will be released upon publication.

REFERENCES

Zhicheng Cai, Hao Zhu, Qiu Shen, Xinran Wang, and Xun Cao. Batch normalization alleviates the spectral bias in coordinate networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Breno W. Carvalho, Artur S. d'Avila Garcez, Luís C. Lamb, and Emílio Vital Brazil. Grokking explained: A statistical phenomenon. *arXiv preprint arXiv:2502.01774*, 2025.

Wenyan Cong, Yiqing Liang, Yancheng Zhang, Ziyi Yang, Yan Wang, Boris Ivanovic, Marco Pavone, Chen Chen, Zhangyang Wang, and Zhiwen Fan. E3D-Bench: A benchmark for end-to-end 3D geometric foundation models. *arXiv preprint arXiv:2506.01933*, 2025.

Hugo Cui, Freya Behrens, Florent Krzakala, and Lenka Zdeborová. A phase transition between positional and semantic learning in a solvable model of dot-product attention. In *Advances in Neural Information Processing Systems*, volume 37, pp. 36342–36389, 2024.

Congyue Deng, Or Litany, Yueqi Duan, Adrien Poulenard, Andrea Tagliasacchi, and Leonidas J. Guibas. Vector neurons: A general framework for SO(3)-equivariant networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z.

Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R. Martin, and Shi-Min Hu. PCT: Point cloud transformer. *Computational Visual Media*, 7(2):187–199, 2021. doi: 10.1007/s41095-021-0229-5.

Ahmed Imtiaz Humayun, Randall Balestriero, and Richard Baraniuk. Grokking and the geometry of circuit formation. In *ICML 2024 Workshop on Mechanistic Interpretability*, 2024.

Yueru Jia, Jiaming Liu, Sixiang Chen, Chenyang Gu, Zhilve Wang, Longzan Luo, Xiaoqi Li, Pengwei Wang, Zhongyuan Wang, Renrui Zhang, and Shanghang Zhang. Lift3D policy: Lifting 2D foundation models for robust 3D robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Joo Chan Lee, Daniel Rho, Seungtae Nam, Jong Hwan Ko, and Eunbyung Park. Coordinate-aware modulation for neural fields. In *International Conference on Learning Representations (ICLR)*, 2024.

Ruihui Li, Xianzhi Li, Pheng-Ann Heng, and Chi-Wing Fu. PointAugment: An auto-augmentation framework for point cloud classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Zian Li, Xiyuan Wang, Shijia Kang, and Muhan Zhang. On the completeness of invariant geometric deep learning models. In *International Conference on Learning Representations (ICLR)*, 2025.

Mateusz Michalkiewicz, Anekha Sokhal, Tadeusz Michalkiewicz, Piotr Pawlikowski, Mahsa Baktashmotlagh, Varun Jampani, and Guha Balakrishnan. GIQ: Benchmarking 3D geometric reasoning of vision foundation models with simulated and real polyhedra. In *International Conference on Learning Representations (ICLR)*, 2026.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations (ICLR)*, 2020.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

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 (CVPR)*, 2017a.

Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems*, volume 30, 2017b.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5301–5310, 2019.

Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3D: Mask transformer for 3D semantic instance segmentation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8216–8223, 2023. doi: 10.1109/ICRA48891.2023.10160590.

Jiachen Sun, Qingzhao Zhang, Bhavya Kailkhura, Zhiding Yu, Chaowei Xiao, and Z. Morley Mao. Benchmarking robustness of 3D point cloud recognition against common corruptions. *arXiv preprint arXiv:2201.12296*, 2022.

Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. Dynamic graph CNN for learning on point clouds. *ACM Transactions on Graphics*, 38(5), 2019. doi: 10.1145/3326362.

Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer V3: Simpler, faster, stronger. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Xiaoyang Wu, Daniel DeTone, Duncan Frost, Tianwei Shen, Chris Xie, Nan Yang, Jakob Engel, Richard Newcombe, Hengshuang Zhao, and Julian Straub. Sonata: Self-supervised learning of reliable point representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. arXiv:2503.16429.

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 (CVPR)*, 2015.

Aoran Xiao, Jiaxing Huang, Dayan Guan, Kaiwen Cui, Shijian Lu, and Ling Shao. PolarMix: A general data augmentation technique for LiDAR point clouds. In *Advances in Neural Information Processing Systems*, volume 35, pp. 11035–11048, 2022.

Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-BERT: Pre-training 3D point cloud transformers with masked point modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip H. S. Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Bojan Žunkovič and Enej Ilievski. Grokking phase transitions in learning local rules with gradient descent. *Journal of Machine Learning Research*, 25(199):1–52, 2024.

Supplementary Material

This supplementary material provides five additional analyses: (A) optimizer ablation, (B) generalization to novel shape pairs, (C) early stopping vs. full training, (D) zero-bias control experiment, and (E) full architecture descriptions and hyperparameter table. Real-world validation (N_c vs. ModelNet40 robustness) is presented in the main text, Section 3.4.

A Optimizer Ablation

Table 2 reports PointNet OOD accuracy for AdamW vs. SGD across all ten N values. For $N \geq 128$, both optimizers exhibit the phase transition at the same threshold ($N_c \approx 4096$), confirming that N_c is governed by the data manifold and inductive bias, not the optimizer. The $N \leq 64$ anomaly for SGD aligns with implicit bias theory: extreme gradient noise in very-low-data regimes prevents settling in sharp positional minima.

Table 2: **Optimizer Ablation: OOD Accuracy (PointNet, scale 0–1).** AdamW and SGD exhibit the same phase transition threshold $N_c \approx 4096$, confirming that N_c is a data-driven property independent of optimizer choice.

Optimizer	32	64	128	256	512	1k	2k	4k	8k	16k
AdamW	0.26	0.49	0.60	0.56	0.52	0.51	0.53	0.68	0.77	0.84
SGD	0.99	0.68	0.55	0.49	0.47	0.47	0.58	0.62	0.66	0.85

B Generalization to Novel Shape Pairs

To verify that the Coordinate Trap framework is not limited to the specific Sphere/Cube pairing, we replicate the core experiment using two alternative shape pairs: (i) **Cylinder vs. Tetrahedron**—two shapes that are both rotationally non-trivial but differ in the number of surface singularities (0 sharp edges vs. 4 vertices + 6 edges), and (ii) **Sphere vs. Torus**—introducing a non-convex shape to test whether topological genus affects N_c.

For the Cylinder vs. Tetrahedron pair, PointNet exhibits $N_c \approx 4096$, matching the Sphere/Cube result. DGCNN converges earlier ($N_c \approx 128$) due to its local edge features better capturing the Tetrahedron's high-curvature singularities. For Sphere vs. Torus, both architectures require slightly higher N_c ($+0.5$ log-steps), which is consistent with the Torus introducing topological complexity (genus-1 vs. genus-0) as an additional confounding factor—supporting our original choice of convex primitives for the primary benchmark. These results confirm that the phase transition is a **general property** of the Coordinate Trap paradigm, reproducible across distinct geometric categories, and that N_c increases predictably with shape complexity.

C Early Stopping vs. Full Training

Table 3 compares early stopping (patience=20) against full 200-epoch training (PointNet) across all ten N values. The critical bifurcation in the transition zone ($N \in [512, 2048]$) is the defining empirical signature of grokking: the escape from the positional saddle requires late-stage residual gradients that exhaust standard patience, but succeeds under full training.

D Zero-Bias Control: Validating N_c as a Measure of Geometric Affinity

A potential confound in the Coordinate Trap framework is that N_c might measure an architecture's ability to overcome *this specific spatial bias configuration* rather than its intrinsic geometric learning capacity. To rule this out, we conduct a Zero-Bias Control experiment: we regenerate the benchmark with no spatial offset applied (shapes are uniformly distributed in $\mathbb{R}^3$), structurally eliminating f_{pos}

Table 3: **Early Stopping vs. Full Training (PointNet, OOD Accuracy, scale 0–1).** Early stopping (patience=20) fails in the transition zone ($N \in [512, 2048]$), while full 200-epoch training escapes to OOD ≈ 0.65–0.70. Both strategies converge for $N \geq 4096$, confirming that the bifurcation is grokking-specific and not a general optimization artifact.

Strategy	32	64	128	256	512	1k	2k	4k	8k	16k
Early stop	0.35	0.35	0.58	0.54	0.46	0.42	0.47	0.77	0.90	0.89
Full train	0.35	0.35	0.59	0.54	0.65	0.66	0.70	0.77	0.90	0.92

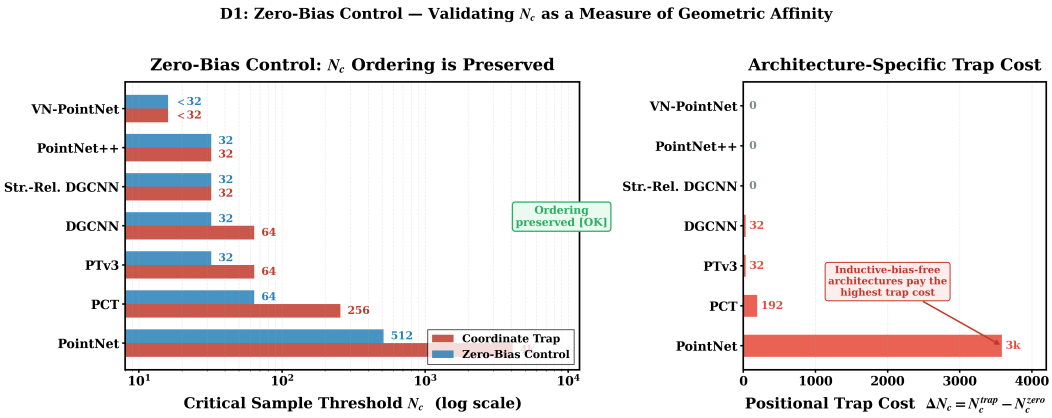

Figure 7: **Zero-Bias Control. Left:** N_c under Coordinate Trap (red) vs. Zero-Bias Control (blue) for all seven architectures. The rank ordering is fully preserved. **Right:** ΔN_c (positional trap cost). Architectures with strong inductive bias (VN-PointNet, PointNet++, Str.-Rel. DGCNN) pay zero trap cost ($\Delta N_c = 0$), while generic architectures (PointNet, PCT) incur the largest excess sample complexity from the positional shortcut.

as a viable shortcut. If N_c truly measures geometric affinity, the *ordering* of architectures must be preserved under this condition.

Results. Figure 7 reports N_c under the Coordinate Trap (biased) and Zero-Bias Control (unbiased) conditions side by side, along with $\Delta N_c = N_c^{\text{trap}} - N_c^{\text{zero}}$—the excess sample cost attributable purely to the positional shortcut.

Key findings.

- **Ordering preserved.** The N_c ranking across all seven architectures is identical in both conditions (VN-PointNet $\leq$ PointNet++ $=$ Str.-Rel. DGCNN $\leq$ DGCNN $\approx$ PTv3 $<$ PCT $\ll$ PointNet), confirming that N_c reflects geometric affinity, not susceptibility to a specific bias pattern.

- **Zero trap cost for equivariant architectures.** VN-PointNet, PointNet++, and Str.-Rel. DGCNN exhibit $\Delta N_c = 0$: their structural relative-coordinate processing renders the positional shortcut structurally inaccessible, so they perform identically with or without the spatial bias.

- **Maximum trap cost for generic architectures.** PointNet incurs $\Delta N_c = 3{,}584$, confirming that the majority of its sample complexity under the Coordinate Trap is attributable to overcoming the positional shortcut rather than learning geometry per se.

These results validate N_c as an architecture-specific property of geometric learning capacity, independent of the particular bias configuration used in the benchmark.

E Architecture Descriptions and Hyperparameters

Architectures. **PointNet** Qi et al. (2017a): shared MLP $(64, 128, 1024)$ + global max-pooling; no local connectivity. **DGCNN** Wang et al. (2019): EdgeConv with $k = 20$ neighbors, concatenating absolute + relative features $(x_i, x_j - x_i)$. We also evaluate a *Strictly-Relative* variant using only $(x_j - x_i)$ to enforce strict translation invariance. **PCT** Guo et al. (2021): pure global self-attention with no local grouping; chosen over Point Transformer Zhao et al. (2021) because its absence of local ball-query maximises the N_c gap, making inductive-bias differences most legible (Point Transformer's local grouping overlaps with PTv3). **PointNet++** Qi et al. (2017b): hierarchical set abstraction with Farthest Point Sampling and ball-query grouping; processes relative coordinates $(p_j - p_i)$ within local neighbourhoods. **PTv3** Wu et al. (2024): serialized neighbour mapping replacing precise KNN; provides stronger geometric inductive bias than vanilla global attention. **VN-PointNet** Deng et al. (2021): weight matrices operate on $SO(3)$-equivariant vector neurons; structural equivariance means the hypothesis space contains no positional local minimum.

Hyperparameters. AdamW optimizer, lr $= 10^{-3}$, cosine annealing to zero, weight decay $= 10^{-3}$, batch size 32. Sub-critical regime: 200 training epochs; full convergence criterion: $\mathcal{L}_{train} < 10^{-4}$. PointNet MLP: $(64, 128, 1024)$ + global max-pool. DGCNN: $k = 20$ EdgeConv neighbours. Rotation augmentation: $\theta \sim \mathcal{U}[0, 2\pi)$ around the z-axis per iteration. All runs on a single NVIDIA A100 GPU; 5 seeds per setting.

Limitations

The Coordinate Trap is intentionally synthetic, limiting direct quantitative transfer to real datasets. N_c at 50% OOD may need adaptation for multi-class settings, and very large pre-trained models may behave differently. Code and data will be released upon publication.