When Does Linear Causal Abstraction Work? Mapping the Boundary on the Grassmannian

Elliot Tower
elliot@elliottower.com

Abstract

Distributed Alignment Search (DAS) identifies low-dimensional linear subspaces that mediate model behavior under interchange interventions, implicitly assuming that the relevant causal variable lives on the Grassmannian manifold $\mathrm{Gr}(k, d)$. When does this assumption hold, and what happens when it fails?

We construct an atlas of 14 modular arithmetic operations across four prime moduli ($p = 53, 97, 113, 211$) and multiple model depths (1, 2, and 4 layers). We report four findings. **(1)** Whether a model develops Grassmannian causal variables is governed by grokking, but grokking itself depends on the interaction between operation, prime, and depth: grokking rates range from 8% (p=53) to 71% (p=211), and depth has a non-monotonic effect. Two robust extremes emerge—operations that never grok regardless of conditions (squaring, cubing) and operations that always grok given sufficient data (addition, multiplication)—with remaining operations on a difficulty spectrum whose boundaries shift with dataset size and model capacity. **(2)** Linear DAS returns IIA = 0.0 on fully grokked modular addition, while memorized models achieve IIA ≥ 0.86 without geometric structure, demonstrating that IIA alone cannot distinguish genuine causal variables from lookup tables. **(3)** A Structured pi-SAE—a label-conditional sparse autoencoder with end-to-end intervention training—recovers nonlinear causal variables where DAS fails. Unconstrained nonlinear baselines achieve perfect IIA vacuously via degenerate encoder-decoders; we introduce *intervention faithfulness* metrics (diversity ratio, distributional fidelity) that expose this failure mode. **(4)** Mechanistic analysis reveals that nonlinearity is necessary for grokking but attention is not (DMF + ReLU groks 18× faster); that Fourier structure emerges *after* generalization, not before it; that weight-space SVD predicts causal geometry without optimization (IIA = 0.999 vs. DAS's 0.150); and that the causal subspace identity is degenerate—10 seeds produce orthogonal subspaces that all support equally valid interventions.

1 Introduction

Mechanistic interpretability seeks to identify causally active, human-understandable variables inside neural networks. Causal abstraction formalizes this goal: a model variable is "causal" if interventions on its internal representation behave like interventions on a high-level algorithmic variable (**?**). Distributed Alignment Search (DAS) operationalizes causal abstraction by learning a rotation $Q \in \mathbb{R}^{d \times k}$ that identifies a k-dimensional subspace of the residual stream. Swapping the in-subspace component between two inputs and measuring whether the model produces the corresponding counterfactual output yields Interchange Intervention Accuracy (IIA) (**?**).

DAS has been applied to syntax (**?**), arithmetic (**?**), and board games (**?**), but every application embeds a structural assumption: the causal variable is a *linear* subspace—a point on the Grassmannian manifold $\mathrm{Gr}(k, d)$. When the true causal variable is linear, DAS recovers it. When it is not, DAS may still return a high-IIA subspace that reflects memorization or distributional artifacts rather than genuine causal structure.

We do not ask whether DAS "works" in the narrow sense of achieving high IIA—it almost always does. We ask: **when does the linear assumption hold?** Is the subspace DAS recovers a genuine Grassmannian

point with the geometric properties expected of a linear causal variable, or an artifact of searching in the wrong function class? And when it fails, **can a nonlinear method recover the causal variable?**

To answer these questions, we construct an atlas of 14 operations over $\mathbb{Z}_p$ spanning the boundary between linear and nonlinear causal structure. We train transformers across four prime moduli ($p = 53, 97, 113, 211$) and three depths (1, 2, 4 layers), fit DAS at multiple subspace dimensions, and evaluate the recovered subspace with three diagnostics beyond IIA: equivariance under the operation's symmetry group, circle geometry of the DAS projection, and intrinsic dimension from k-sweeps. When DAS fails, we apply a Structured pi-SAE (**?**)—a label-conditional sparse autoencoder—showing that nonlinear causal variables exist even where linear methods return zero.

The atlas reveals that Grassmannian causal structure is governed by grokking, but grokking itself is not a fixed property of the operation: it depends on dataset size and model depth. At p=53, only 1 of 13 operations groks; at p=211, 10 of 14 do. Depth has a non-monotonic effect: 2-layer models unlock grokking for operations like subtraction and absolute difference that fail at 1 layer, but 4-layer models regress on these same operations while the robust grokkers persist. Two robust extremes persist across all conditions: squaring and cubing never grok, while addition and multiplication always grok at $p \geq 97$. Between these extremes, operations sit on a difficulty spectrum whose position shifts with prime and depth.

Grokked modular addition returns IIA = 0.0 under linear DAS despite having learned the correct algorithm, while memorized models achieve IIA $\geq$ 0.86 without geometric structure—demonstrating that IIA alone cannot distinguish genuine causal variables from lookup tables. A Structured pi-SAE with end-to-end intervention training recovers nonlinear causal variables where DAS fails. To confirm this method generalizes beyond grokking, we validate on six GPT-2 language tasks, achieving strict IIA $\geq$ 0.90 on five of six at k=1 (Table 5).

Beyond the atlas, we ask **how does grokking create causal structure?** Architecture ablations show that nonlinearity is necessary but attention is not. Fine-grained trajectory analysis reveals that Fourier structure emerges *after* generalization, and that weight-space SVD predicts causal geometry without optimization. The causal subspace itself is gauge freedom—the model does not prefer a canonical subspace.

Contributions.

1. **Multi-condition atlas.** We construct an atlas of 14 operations across 4 prime moduli and multiple model depths, mapping the boundary of linear causal abstraction. The boundary is not a fixed property of the operation: grokking rates range from 8% (p=53) to 71% (p=211), depth has a non-monotonic effect (2-layer helps, 4-layer regresses for boundary operations), and only the extremes (squaring/cubing vs. addition/multiplication) are robust across all conditions. Stochastic grokking—seed-dependent outcomes for operations near the boundary—provides the cleanest evidence that generalization, not architecture, governs causal geometry.

2. **IIA insufficiency and faithfulness metrics.** We show that high IIA is compatible with memorization (IIA = 1.0 at k=4 for non-grokking operations) and with degenerate encoder-decoders (NL-DAS achieves IIA = 1.0 via lookup tables). We introduce equivariance testing, intervention diversity ratio, and continuous distributional measures (KL, JS) as necessary complements, organized into a four-level evaluation hierarchy.

3. **Structured pi-SAE.** A label-conditional sparse autoencoder with end-to-end intervention training recovers nonlinear causal variables where DAS fails entirely. The generative structure prevents the degenerate solutions that plague unconstrained nonlinear baselines. Validated on six GPT-2 language tasks (strict IIA $\geq$ 0.90 on five of six at k=1) to confirm generalization beyond grokking.

4. **Mechanistic analysis.** Nonlinearity is necessary for grokking but attention is not (DMF + ReLU groks 18× faster). Fourier structure lags generalization by ∼1,000 epochs. Weight-space SVD achieves IIA = 0.999 without optimization. The causal subspace identity is degenerate (10 seeds, pairwise overlap $\approx$ 0.008).

2 Background

2.1 Causal abstraction and DAS

Let $h(x) \in \mathbb{R}^d$ be an intermediate activation for input x. DAS searches for an orthonormal $Q \in \mathbb{R}^{d \times k}$ whose column span defines a causal subspace. Given a base input x_b and source input x_s, the interchange intervention replaces the in-subspace component:

$$h' = h_b - QQ^\top h_b + QQ^\top h_s. \tag{1}$$

IIA measures the fraction of interventions for which the model's output matches the target counterfactual. High IIA indicates causal sufficiency: the subspace contains enough information to drive behavior under interchange. But sufficiency is not validity. A subspace may score high IIA because it encodes a memorized lookup table, because correlated directions redundantly encode the same label, or because the intervention distribution is too easy. **?** demonstrated this concretely: subspace activation patching can produce interpretability illusions where dormant pathways inflate IIA without reflecting genuine causal structure. Our central claim is that IIA must be paired with geometric diagnostics that test whether the subspace is a genuine linear causal variable rather than an artifact of the linear function class.

2.2 The Grassmannian manifold

The Grassmannian $\mathrm{Gr}(k, d)$ is the set of all k-dimensional linear subspaces of $\mathbb{R}^d$. A DAS solution $Q \in \mathbb{R}^{d \times k}$ represents a point $[Q] = \mathrm{span}(Q)$ on this manifold; any QR for $R \in O(k)$ represents the same point. The geodesic distance between two subspaces is determined by their *principal angles*: if Q_1, Q_2 have orthonormal columns, the singular values of $Q_1^\top Q_2$ are $\cos\theta_i$, and the geodesic distance is

$$d_{\mathrm{Gr}}([Q_1], [Q_2]) = \left(\sum_i \theta_i^2 \right)^{1/2} \tag{2}$$

(**??**). Every application of DAS is implicitly a search on this manifold. The question "does DAS work?" is really "does the true causal variable live on $\mathrm{Gr}(k, d)$?" When it does, DAS recovers a meaningful point. When it does not, DAS still returns a point on $\mathrm{Gr}(k, d)$, but that point may be an artifact.

2.3 Structured disentangled generative models

Semi-supervised deep generative models (**?**) extend the VAE framework (**?**) by partitioning the latent space into supervised and unsupervised factors. The generative model factors as $p(x, y, z) = p(x \mid y, z)\, p(y)\, p(z)$, where y is a labeled (interpretable) variable and z captures unlabeled nuisance variance. The recognition model $q(y, z \mid x) = q(z \mid x, y)\, q(y \mid x)$ uses neural network encoders, allowing nonlinear latent structure.

This framework is a natural generalization of DAS: DAS constrains the causal variable to live on $\mathrm{Gr}(k, d)$ (a linear subspace), while the structured VAE allows it to occupy a nonlinear manifold in activation space. The key advantage is disentanglement: the supervised factor y is trained to predict the operation's output, while the unsupervised factor z absorbs everything else. If the causal variable is linear, the VAE encoder should learn a rotation equivalent to DAS. If the causal variable is nonlinear (e.g., circular structure from Fourier representations), the VAE can recover it where DAS cannot.

The loss function combines the evidence lower bound (ELBO) with a supervised classification term:

$$\mathcal{L} = \underbrace{-\mathbb{E}_q[\log p(x \mid y, z)]}_{\text{reconstruction}} + \beta \underbrace{D_{\mathrm{KL}}(q(y, z \mid x) \,\|\, p(y)p(z))}_{\text{regularization}} + \alpha \underbrace{\mathcal{L}_{\mathrm{CE}}(y, \hat{y})}_{\text{supervision}}. \tag{3}$$

The encoder weights projecting to y provide a linearized approximation of the causal subspace; if this linearization has high overlap with the DAS subspace, it confirms that the causal variable is approximately Grassmannian.

2.4 Identifiability of the causal latent factor

A central concern for any latent-variable method is *identifiability*: does the encoder recover the *true* latent factors, or an arbitrary reparameterization? The Structured pi-SAE architecture is designed to approximate the conditions of iVAE (**?**): (i) the task label y is an observed auxiliary variable; (ii) the label-conditional prior $p(z_{\text{causal}} \mid y) = \mathcal{N}(\mu_y, \sigma_y^2 I)$ is an exponential family with per-label means; (iii) the ELBO reconstruction loss encourages an approximately injective decoder.

Connection to iVAE identifiability. If the iVAE conditions of **?** hold exactly—injective decoder, sufficient variability of per-label parameters, and exponential family conditionals—then Theorem 1 of **?** guarantees recovery of the true latent factors up to a component-wise bijection. In practice, neural network activations are deterministic functions of the input, not samples from a generative model, so these conditions hold only approximately. The ELBO encourages but does not guarantee decoder injectivity, and the rank condition on per-label means is plausible for $p = 113$ classes but may not hold for small datasets (e.g., the 182-class capitals task). We therefore treat iVAE identifiability as a *design motivation* rather than a formal guarantee, and rely on empirical controls (§4.7) to validate that the encoder recovers meaningful structure.

Three consequences of the generative structure. (1) NL-DAS violates the reconstruction constraint: its decoder is not reconstruction-constrained, so multiple latent codes may map to the same activation and identifiability does not hold (§4.7). **(2) The component-wise ambiguity is harmless for IIA**: interchange swaps z_{causal} wholesale, so any bijection ϕ cancels. **(3) Linear DAS is the special case $\phi \in O(k)$.** When the true causal variable is linear, the structured encoder degenerates to a rotation, matching the DAS solution.

2.5 Grokking as a phase transition

Grokking is a training regime in which models memorize the training set long before generalizing (**?**). Modular arithmetic models trained with weight decay exhibit grokking reliably for some operations but not others (**??**). Prior work analyzed grokking through Fourier representations and circular structure in the embedding space. Our perspective is complementary: we ask how the causal subspace recovered by DAS changes across the memorization-to-generalization transition, and whether that transition is deterministic or stochastic across random seeds. **?** frame grokking as compression—moving from a memorized high-complexity solution to a generalizing low-complexity one; our Grassmannian perspective gives this compression a geometric interpretation as convergence to a specific point on $\text{Gr}(k, d)$.

3 Methods

3.1 Operations and models

We study 14 binary operations over $\mathbb{Z}_p$ (Table 1), chosen to span group actions, polynomials, piecewise functions, and non-algebraic maps. We train transformers with $d_{\text{model}} = 128$, four attention heads ($d_{\text{head}} = 32$), and $d_{\text{MLP}} = 512$ on examples of the form $(a, b, =) \mapsto y$. The core sweep varies three axes: prime modulus $p \in \{53, 97, 113, 211\}$, model depth $n_{\text{layers}} \in \{1, 2, 4\}$, and operation (14 types). Training uses the standard grokking setup: 30% train split, AdamW with learning rate 10^{-3}, weight decay 1.0, full-batch, and 25,000–80,000 epochs depending on the operation (1.5× for deeper models). We classify a model as *grokked* if its final test cross-entropy is below 0.1.

3.2 DAS fitting and k-sweeps

For each trained model, we cache residual-stream activations and train a DAS rotation $Q \in \mathbb{R}^{d \times k}$ using interchange interventions for 200 gradient steps (Adam, lr $= 10^{-3}$, batch size 16 interchange pairs). We sweep $k \in \{2, 4, 6, 8, 10, 12, 16, 20, 24, 32\}$ to estimate the intrinsic dimension k^*, defined as the smallest k

Operation	Formula	Category	Seeds tested
Multiplication	$ab \bmod 113$	Group action	42, 2024
Comp. addition	$(a+b) \bmod 91$	Group action	orig, 42
Subtraction	$(a-b) \bmod 113$	Group action	42
Division	$a/b \bmod 113$	Group action	42
Bitwise XOR	$a \oplus b \bmod 113$	Group action	42, 137, 2024
Sum of squares	$(a^2+b^2) \bmod 113$	Polynomial	42
Cubic sum	$(a^3+b^3) \bmod 113$	Polynomial	42
Cubing	$a^3 \bmod 113$	Polynomial	42
Squaring	$a^2 \bmod 113$	Polynomial	42
Polynomial	$(a^2+b) \bmod 113$	Polynomial	42
Affine	$(2a+3b+5) \bmod 113$	Polynomial	42
Max	$\max(a,b) \bmod 113$	Piecewise	42
Abs. difference	$\lvert a-b \rvert \bmod 113$	Piecewise	42
Power	$a^b \bmod 113$	Non-algebraic	42, 137

Table 1: Operation atlas. Group actions provide clean equivariance targets. Polynomials test whether nonlinear maps admit linear causal variables. Piecewise and non-algebraic operations probe the boundary.

achieving IIA ≥ 0.95. Each k value is an independent DAS optimization, not a nested subspace—k-sweep profiles reveal whether the causal variable is inherently low-dimensional or requires many dimensions.

3.3 Hard-example selection for IIA

Standard IIA evaluation includes many "easy" examples where the intervention does not change the model's top-1 prediction, inflating IIA regardless of whether the subspace is causally valid. We define a *hard example* as a (base, source) pair where: (1) the base and source have different correct outputs $y_b \neq y_s$, (2) the model's prediction matches the correct output for both base and source separately, and (3) the interchange intervention at the DAS subspace *can* flip the model's prediction from y_b to y_s. Hard IIA measures the fraction of hard examples where the flip actually occurs. This selection follows the methodology of constrained PCA-initialized DAS, which showed that standard IIA saturates too easily while hard IIA reveals genuine causal structure.

3.4 Equivariance testing

For operations with a natural group action, we test whether the DAS subspace respects the algebraic symmetry. Given input (a, b) with DAS projection $z = Q^\top h(a, b)$, we construct a shifted input $(a+1 \bmod p, b)$ and check whether the DAS projection rotates by $2\pi/p$ radians. The equivariant fraction is the proportion of test inputs satisfying this rotation property. Random orthogonal subspaces of the same dimension serve as controls; across all operations, random-subspace equivariance is below 1%.

3.5 Structured VAE for nonlinear causal variables

For each trained grokking model, we cache residual-stream activations at the post-MLP hook and train a structured VAE with:

- z_{causal}: k-dimensional latent factor ($k \in \{2, 4, 8\}$), semi-supervised with the operation's output label via a classification head.
- z_{nuisance}: m-dimensional unsupervised factor ($m = 15$) to absorb non-causal variance.
- Encoder: 2-layer MLP ($d_{\text{model}} \to 128 \to z$) with ReLU.
- Decoder: 2-layer MLP ($z \to 128 \to d_{\text{model}}$) with ReLU.

Training uses the combined ELBO + classification loss (Eq. 3) for 300 epochs with $\beta = 1.0$, $\alpha = 10.0$, L_1 coefficient $\lambda = 1.0$, and Adam optimizer with learning rate 10^{-3}.

VAE-based IIA. We evaluate the VAE's causal subspace by performing interchange interventions in the *latent* space: given base activation h_b and source activation h_s, we encode both, swap z_{causal} while keeping z_{nuisance} from the base, decode back to activation space, and measure whether the model produces the source's output. This is the nonlinear analogue of Eq. 1: instead of $QQ^\top$ projection, we use the encode-swap-decode path.

VAE equivariance. We test whether the VAE's z_{causal} exhibits equivariance under additive shifts, using the same protocol as for DAS (shift input $a \to a + 1$, check whether z_{causal} rotates consistently). Because the VAE encoder is nonlinear, it can represent circular structure directly in a 2D latent space.

3.6 Nonlinear DAS baseline

To isolate the contribution of the VAE's generative structure from the benefit of nonlinearity alone, we compare against an unconstrained nonlinear featurizer. Given MLP encoder $f : \mathbb{R}^d \to \mathbb{R}^d$ and decoder $g : \mathbb{R}^d \to \mathbb{R}^d$, the nonlinear DAS (NL-DAS) intervention is:

$$h' = g\big(f(h_b) - QQ^\top f(h_b) + QQ^\top f(h_s)\big) \tag{4}$$

where $Q \in \mathbb{R}^{d \times k}$ is learned jointly with f and g by maximizing interchange accuracy. Both f and g use the same architecture as the VAE encoder (2-layer MLP, 256 hidden units, ReLU), matching the VAE's representational capacity. Crucially, NL-DAS has *no reconstruction constraint*: f and g are trained exclusively on interchange loss and need not be (approximately) inverse to each other.

We also evaluate NL-DAS+recon, which adds a reconstruction penalty $\lambda \|g(f(h)) - h\|^2$ to the training loss, forcing f and g to approximate mutual inverses. This tests whether the unconstrained NL-DAS result degrades when the encoder-decoder cannot be degenerate.

3.7 Intervention faithfulness metrics

Binary IIA measures *interchange effectiveness*: does swapping the subspace component change the model's output to match the source? But effectiveness alone does not imply that the intervention is *faithful*—that it modifies only the targeted causal variable while preserving all other information. A method that achieves IIA = 1.0 by overwriting the entire activation with a class-specific template is effective but not faithful.

We introduce four complementary metrics to evaluate intervention faithfulness:

Continuous distributional metrics. Binary IIA saturates at 1.0 whenever the top-1 prediction flips, hiding differences in intervention quality. We measure the full distributional impact:

- **KL divergence**: $D_{\text{KL}}(p_{\text{clean}} \| p_{\text{iv}})$, where p_{clean} and p_{iv} are the softmax distributions before and after intervention. Lower KL means the intervention preserves more of the clean distribution beyond the swapped variable.
- **Jensen-Shannon divergence**: the symmetric alternative $D_{\text{JS}} = \frac{1}{2} D_{\text{KL}}(p \| m) + \frac{1}{2} D_{\text{KL}}(q \| m)$ where $m = (p + q)/2$. Bounded in $[0, \ln 2]$, more stable than KL when distributions have disjoint support.
- **Probability difference**: $p_{\text{iv}}(y_s) - p_{\text{iv}}(y_b)$, a continuous analog of binary IIA that captures the model's confidence margin, not just top-1 agreement.
- **Normalized logit difference**: $\frac{l_{\text{iv}}(y_s) - l_{\text{iv}}(y_b)}{|l_{\text{clean}}(y_s) - l_{\text{clean}}(y_b)|}$, where l denotes logits. A value of 1.0 means the intervention reproduces the clean model's margin exactly; values $\gg 1$ or $\ll 0$ indicate distortion.

Intervention diversity. For each source label y_s, we compute the standard deviation of the intervened activations h' across all base inputs with the same source. If the decoder produces the same h' regardless of which base input was used—varying only with y_s—the intervention is a lookup table: it does not preserve base-specific nuisance information. We define the *diversity ratio*:

$$\rho = \frac{\mathbb{E}_{y_s}\left[\text{std}(h'_{y_s})\right]}{\mathbb{E}_{y_s}\left[\text{std}(h_{y_s})\right]} \tag{5}$$

where the numerator is over intervened activations sharing source label y_s and the denominator is over the corresponding natural source activations. $\rho \approx 0$ indicates a lookup table (all intervened activations for the same y_s collapse to a single point). $\rho \approx 1$ indicates the intervention preserves the natural variation within each class.

Reconstruction fidelity. For encoder-decoder methods (NL-DAS, VAE), we measure $\|g(f(h)) - h\|^2$ on held-out activations. A high reconstruction MSE with high IIA is a warning sign: the decoder may be "hallucinating" the correct output rather than faithfully reconstructing the activation with a modified causal component. The generative constraint is not merely a regularizer—it is a necessary condition for identifiability (Proposition 2.4), as shown empirically by pi-VAE's failure despite having the correct prior direction (Table 4).

3.8 Cross-distribution generalization

Standard train/test splits draw from the same distribution. For IOI on GPT-2, we additionally evaluate under distribution shift: (1) **disjoint names**—train on one set of proper names, evaluate on a completely disjoint set; and (2) **disjoint templates**—train on sentence templates 1–3, evaluate on templates 4–5 plus novel hard templates. A method that generalizes under distribution shift is finding a genuine causal variable; one that fails is exploiting distributional regularities in the training set.

4 Results

4.1 The atlas: a grokking-governed spectrum

At a single prime (p=113, 1-layer), the 14 operations appeared to partition into three classes (Table 2). The multi-modulus sweep reveals that this partition is not fixed: grokking depends on the interaction between operation, prime, and depth (Table 3).

Three patterns emerge from the multi-condition sweep. First, grokking rate scales with dataset size: at p=53 (∼800 training pairs), only bitwise XOR groks; at p=211 (∼13,000 pairs), 10 of 14 operations do. This suggests that many operations classified as "never Grassmannian" at p=113 simply had insufficient data to cross the grokking threshold.

Second, the relationship between depth and grokking is non-monotonic. At p=113 with 2 layers, subtraction, absolute difference, and division all grok, whereas none of these grok with 1 layer at the same prime. But at 4 layers, these same operations *regress*—none grok despite the additional capacity. Only the robust grokkers (addition, multiplication, bitwise XOR, sum of squares) maintain grokking at 4 layers. The 2-layer sweet spot suggests that moderate depth helps by providing compositional capacity, while excessive depth hinders generalization by enabling more complex memorization strategies. This non-monotonic pattern is consistent with the observation that deeper networks have higher effective capacity for memorization (**?**), which competes with the generalization pathway that grokking requires.

Third, two robust extremes persist across all conditions. Squaring and cubing never grok at any prime or depth tested—their output structure (quadratic/cubic residues) appears fundamentally resistant to the generalization transition. Conversely, addition and multiplication grok at every prime ≥ 97 and at every depth, consistent with their clean group-homomorphic algebraic structure.

Operation	Grok	Loss	IIA(k=2)	k^*	Equiv.	Class
Always Grassmannian (at p=113)						
Multiplication	✓	0.000	1.000	2	0.995	Always
Subtraction	✓	0.000	1.000	2	1.000	Always
Division	✓	0.000	1.000	2	1.000	Always
Bitwise XOR	✓	0.003	1.000	2	1.000	Always
Cubic sum	✓	0.000	1.000	2	1.000	Always
Sum of squares	✓	0.005	1.000	2	0.987	Always
Max	✓	0.074	0.960	2	0.957	Always
Stochastic						
Comp. addition*	✓	0.000	—	—	0.998	Stochastic
Comp. addition (s42)	×	10.207	0.800	6	0.110	Stochastic
Power (s137)	✓	0.088	0.940	4	0.965	Stochastic
Power (s42)	×	1.143	0.980	2	0.665	Stochastic
Never Grassmannian (at p=113)						
Cubing	×	9.765	0.857	4	0.509	Never
Squaring	×	7.808	0.857	4	0.314	Never
Abs. difference	×	1.835	0.800	>32	0.189	Never
Polynomial	×	21.550	0.720	>32	0.015	Never
Affine	×	26.209	0.780	>32	0.026	Never

Table 2: Atlas results at p=113 (1-layer). k^* is the smallest k achieving IIA ≥ 0.95. *Original seed (unknown value); k-sweep not recorded. Table 3 shows how this partition shifts across primes.

4.2 Linear DAS fails on grokked modular addition

The starkest evidence that the Grassmannian assumption can fail even for correct, generalizing models comes from modular addition. We train DAS at $k \in \{2, 4, 8, 16\}$ on residual-stream activations from a fully grokked model (test loss $< 10^{-7}$, test accuracy 100%). At *every* subspace dimension, DAS returns IIA = 0.0—not low, but exactly zero.

This is not a failure of DAS optimization. The model has learned the correct modular addition algorithm via Fourier representations (**?**): the "identity" of each number a is encoded as $(\cos 2\pi fa/p, \sin 2\pi fa/p)$ for key frequencies f. This representation lives on S^1 (a circle embedded in $\mathbb{R}^d$), not on a linear subspace of $\mathbb{R}^d$. DAS searches $\mathrm{Gr}(k, d)$ for a flat plane that captures a curved manifold—a category error, not an optimization failure.

This motivates the Structured pi-SAE approach (§4.6): a nonlinear sparse encoder can map the circular representation to a structured latent space where interchange interventions succeed. On grokked addition, the Structured pi-SAE achieves IIA = 1.0 at k=2, confirming that the causal variable exists but is nonlinear.

4.3 Stochastic grokking: the same operation, opposite outcomes

The multi-modulus sweep reveals that stochastic grokking is not an anomaly confined to two operations—it is a pervasive feature of the boundary region. At p=113, power ($a^b \bmod 113$) grokked at seed 137 (test loss 0.088, equivariance 96.5%) but not at seed 42 (test loss 1.14, equivariance 66.5%). Composite addition ($a + b \bmod 91$) grokked at p=97 but not at p=113 or p=211 (at seed 42)—the same operation exhibits different grokking behavior across primes, confirming that the boundary is genuinely stochastic rather than a deterministic function of the operation's algebraic structure.

The multi-modulus data also reveals that operations previously classified as "never Grassmannian" were misclassified. Bitwise XOR groks at all four primes tested, including p=53 where almost nothing else groks. Sum of squares groks at three of four primes. These operations were classified as "never" based on a single prime (p=113) where they happened not to grok—a classification error that the multi-condition sweep

Operation	p=53	p=97	p=113	p=211	2L	4L
Robust grokkers						
Addition	—	✓	✓	✓	✓	✓
Multiplication	—	✓	✓	✓	✓	✓
Bitwise XOR	✓	✓	✓	✓	✓	✓
Sum of squares	∼	✓	✓	✓	✓	✓
Difficulty spectrum (condition-dependent)						
Division	—	✓	—	✓	✓	—
Subtraction	—	—	—	✓	✓	—
Max	∼	∼	✓	✓	∼	—
Min	∼	∼	✓	✓	∼	∼
Abs. difference	—	—	—	✓	✓	—
Comp. addition	—	✓	—	—	✓*	—
Power	—	—	—	✓	—	—
Affine	—	—	—	—	—	—
Robust non-grokkers						
Squaring	—	—	—	—	—	—
Cubing	—	—	—	—	—	—
Aggregate grokking rate						
	1/13 (8%)	6/14 (43%)	6/14 (43%)	10/14 (71%)		

Table 3: Grokking across primes and depth. 1-layer results across four primes; 2L and 4L columns at p=113. ✓= grokked (test CE < 0.1); ∼ = near-grokked (acc > 0.97 but CE > 0.1); — = did not grok. *Comp. addition groks at p=97 (1L and 2L) but not p=113 or p=211 at seed 42—consistent with stochastic classification. Note that 4-layer models do *not* uniformly improve over 2-layer: subtraction, abs. difference, and division grok at 2L but regress at 4L.

corrects. The lesson is methodological: single-condition classifications of grokking behavior are unreliable for operations near the boundary.

4.4 Memorization artifacts: high IIA, no structure

Squaring and cubing achieve IIA = 0.86 at $k = 2$ and IIA = 1.0 by $k = 4$ without grokking (test losses 7.8 and 9.8). Equivariance exposes the artifact: 31.4% for squaring and 50.9% for cubing, far below the >95% threshold. The lesson: **high IIA with low-k saturation is compatible with a lookup table**. "Works for interchange" and "is a Grassmannian causal variable" are different claims.

4.5 k-sweep profiles correlate with equivariance

Grassmannian variables saturate at IIA ≈ 1.0 by $k = 4$: their causal structure is inherently 2-dimensional. Non-Grassmannian variables show gradual IIA growth that may not reach 1.0 even at $k = 32$. Polynomial reaches IIA = 0.72 at $k = 2$ and requires $k \geq 8$ to approach 0.90.

4.6 Structured pi-SAE recovers nonlinear causal variables

We upgrade the vanilla structured VAE to a **Structured pi-SAE**: a label-conditional sparse autoencoder that partitions the latent space into z_{causal} (supervised, L_1-penalized) and z_{nuisance} (unsupervised). The "pi" denotes the label-conditional prior from **?**; the "SAE" denotes L_1 sparsity on z_{causal} with an expansion factor of 8× (so k causal dimensions expand to $8k$ sparse features). This combines the identifiability guarantees of auxiliary supervision (**?**) with the interpretability benefits of sparse dictionary learning (**?**).

The 2×2 ablation. Both components—structured prior and sparsity—are necessary (Table 4). On grokking tasks: (1) the plain SAE (no label-conditional prior) achieves IIA = 1.0 on addition but degrades

Operation	Plain prior $\mathcal{N}(0, I)$ VAE	SAE	Structured prior $p(z \mid y)$ pi-VAE	**Str. pi-SAE**
Grokked				
Addition	0.02	1.00	0.00	**1.00**
Multiplication	0.00	0.72	0.00	**1.00**
Quartic sum	0.02	0.32	0.00	**1.00**
IOI (GPT-2)	0.56	0.83	0.95	**0.98**
Non-grokked (correct answer: ≈ 0)				
Mixed product	0.03	0.01	0.01	0.04
Squaring	0.00	0.00	0.00	0.00

Table 4: 2×2 ablation: IIA at k=2. Plain SAE works on addition but degrades on harder operations (multiplication 0.72, quartic sum 0.32) because without the structured prior, the sparse encoder has no target direction. pi-VAE (structured prior, no sparsity) achieves 0.00 on grokked operations—the prior provides the right direction but without sparsity the encoder spreads the causal signal across all dimensions, which cannot be cleanly swapped. Only Structured pi-SAE achieves 1.00 uniformly on all grokked operations and ≤ 0.04 on all non-grokked operations.

Task	DAS	DAS (strict)	E2E	E2E (strict)
IOI	0.19	0.19	**1.00**	**1.00**
Gender bias	0.62	0.52	**1.00**	**1.00**
Greater-than	0.93	0.28	**1.00**	**1.00**
SVA	0.92	0.00	**1.00**	**1.00**
Hypernymy	0.07	0.02	**0.97**	**0.90**
Capitals[†]	0.12	—	**0.51**	**0.26**

Table 5: Standard and strict IIA at k=1 on six GPT-2 language tasks (layer 8, additive intervention). Structured pi-SAE with E2E training dominates DAS on all six tasks. [†]Capitals has 182 classes with only 190 examples (≈1 per class), limiting classifier training.

on harder operations (multiplication 0.72, quartic sum 0.32); (2) the pi-VAE (structured prior, no sparsity) achieves IIA ≈ 0 across all operations; (3) *only* the Structured pi-SAE achieves IIA = 1.0 uniformly across all grokked operations and IIA ≈ 0 on all non-grokked operations.

Reconstruction MSE as a falsifiability criterion. For non-grokked operations, reconstruction MSE is 11–24; for grokked operations, 0.6–1.2. High MSE with near-zero IIA means the model has *no structured causal variable to recover*—not that the method failed. This diagnostic is unavailable for NL-DAS, which returns IIA = 0.6–0.8 on non-grokked operations (false positives) because it has no reconstruction constraint.

End-to-end intervention training. For language model tasks where the causal structure is more complex, we add an **end-to-end (E2E) intervention CE loss**: after swapping z_{causal} and decoding, we run the intervened activation through the remaining layers and compute $-\log p(y_s \mid h')$. This differentiable objective directly optimizes for intervention success, analogous to how DAS trains on interchange accuracy. Training proceeds in two phases: 200 epochs of reconstruction + classification warmup, followed by E2E intervention training with an additive intervention $h' = h_b + \text{dec}(z_{\text{swap}}) - \text{dec}(z_{\text{orig}})$ that cancels reconstruction error.

Validation on GPT-2 language tasks. The Structured pi-SAE was developed to recover nonlinear causal variables in grokking models. A natural concern is whether it overfits to the specific geometry of modular arithmetic. To test generalization, we evaluate on six tasks from the MIB benchmark (**?**) at layer 8 of GPT-2-small, where DAS is the established baseline (Table 5).

The Structured pi-SAE with E2E training matches or exceeds DAS on all six GPT-2 tasks at k=1, confirming that the method is not specific to grokking geometry. On five of six tasks, E2E achieves strict IIA ≥ 0.90; on IOI, gender bias, greater-than, and SVA it reaches 1.00. The sole exception is capitals, where the dataset

Task	k	DAS	NL-DAS	NL-DAS+r	VAE	ρ_{NL}	ρ_{VAE}
IOI	1	0.183	1.000	0.867	1.000	0.05	0.83
IOI	2	0.322	1.000	0.700	0.989	0.05	0.89
IOI	4	0.572	1.000	0.811	0.989	0.05	0.92
Addition	1	0.006	0.083	0.028	1.000	0.12	1.00
Addition	2	0.022	0.422	0.039	1.000	0.29	1.00
Addition	4	0.283	0.689	0.033	1.000	0.43	1.00
Mult.	1	0.028	0.239	0.028	1.000	0.65	1.00
Mult.	2	0.072	0.544	0.111	1.000	1.00	1.00
Mult.	4	0.283	0.933	0.372	1.000	1.12	0.99
Squaring	†	*did not grok (acc = 0.00)*					

Table 6: NL-DAS achieves high IIA by learning degenerate encoder-decoders. The VAE column reports Structured pi-SAE results. ρ is the intervention diversity ratio (Eq. 5); NL-DAS+r adds a reconstruction penalty ($\lambda_{\text{recon}} = 0.1$). †Squaring ($x^2$ mod 113) failed to grok within 80,000 epochs at all tested k. All controls (random labels, reconstruction-only, untrained, unconstrained plain VAE) achieve IIA < 0.09, confirming the Structured pi-SAE is not cheating.

contains 182 classes with only 190 examples, limiting classifier training. On hypernymy, E2E achieves IIA = 0.97 at k=1—14× better than DAS (0.07)—and DAS's strict IIA on SVA is exactly 0.00 despite standard IIA of 0.92, revealing that DAS shifts probability mass toward the correct token without ever making it the argmax. These results serve as a sanity check: the Structured pi-SAE is a strict generalization of DAS (matching it where linear structure suffices, exceeding it where nonlinear structure is needed), not a special-case method.

On IOI (**?**), NL-DAS achieves IIA = 1.00, but the diversity ratio $\rho \approx 0.05$ (vs. $\rho \approx 0.89$ for the pi-SAE) exposes it as vacuous (Table 6).

4.7 Unconstrained nonlinear DAS is vacuous

A natural response to DAS's linearity constraint is to prepend an MLP featurizer: learn a nonlinear coordinate system in which the causal variable *becomes* linear, then apply standard DAS. This approach—which we call NL-DAS (§3.6)—achieves IIA = 1.000 on IOI at $k = 1$, compared to DAS's 0.183. This looks like a dramatic improvement. It is not.

The lookup-table failure mode. NL-DAS is trained end-to-end on interchange loss with no reconstruction constraint. The encoder f and decoder g need not be approximate inverses, and the training objective provides no incentive for them to be. This creates a degenerate solution: f maps every activation to a space where the first coordinate encodes the class label and the remaining coordinates are arbitrary; g ignores the base-specific content entirely and produces a "canonical activation for class y_s" regardless of which base input was used.

The intervention diversity ratio (§3.7) exposes this failure. Table 6 reports results for IOI and two grokking operations (addition, multiplication mod 113) across $k \in \{1, 2, 4\}$; squaring did not grok within 80,000 epochs.

Two lines of evidence support this:

1. **Controls rule out the structured VAE cheating.** Random-label VAE (IIA ≤ 0.02), reconstruction-only VAE with $\alpha = 0$ (IIA ≤ 0.005), untrained VAE (IIA = 0.000), and an unconstrained plain VAE with no causal/nuisance split (IIA ≤ 0.08) all fail. The structured VAE's success requires correct supervision *and* the causal-nuisance partition. NL-DAS lacks the reconstruction constraint that prevents degenerate solutions; without inductive biases, unsupervised disentanglement is impossible (**?**).

2. **The structured prior and reconstruction constraint are necessary, not optional.** The ELBO forces the decoder to produce activations close to the original, the KL regularization prevents encoder collapse, and the causal-nuisance partition ensures that base-specific information is preserved in z_{nuisance}. NL-DAS has none of these constraints.

Why the VAE is not vacuous. The structured VAE's ELBO provides three constraints that prevent the lookup-table failure mode: (1) the reconstruction loss forces the decoder to produce activations close to the original, not class-specific templates; (2) the KL regularization prevents the encoder from collapsing the latent space to a few discrete points; and (3) the causal-nuisance partition ensures that z_{nuisance} absorbs base-specific information, so the decoder must use it. Together, these constraints mean that swapping z_{causal} while preserving z_{nuisance} genuinely modifies only the causal component.

Implications. Any nonlinear method for causal abstraction must be evaluated with intervention faithfulness metrics, not just IIA. The NL-DAS failure mode is not specific to our architecture—it applies to any unconstrained encoder-decoder trained exclusively on interchange loss. Prior work proposing neural network featurizers for causal abstraction (**?**) should be re-evaluated with diversity ratio and reconstruction checks.

4.8 Surprising positive and negative cases

Cubic sum ($a^3 + b^3 \bmod 113$) achieves perfect equivariance (100%) despite cubing alone reaching only 50.9%. The boundary is not "group action versus polynomial" but whether the operation's binary structure provides sufficient additive symmetry.

Affine ($2a + 3b + 5 \bmod 113$) is a linear function that *fails* to produce a Grassmannian variable. Under an additive shift $a \mapsto a + g$, the output shifts by $2g$, not g. Equivariance requires the operation to be a group action, not merely a linear function.

Max ($\max(a, b) \bmod 113$) achieves 95.7% equivariance despite not being a group action. It is piecewise equivariant: when the argmax is stable under the shift, max behaves like a shifted identity.

5 The Mechanism: How Grokking Creates Causal Structure

The atlas (§4) established *what*: causal structure emerges through grokking, with the boundary depending on operation, dataset size, and model depth. This section asks *how* and *why*. We trace the emergence of causal structure through five lenses: the minimal architecture required, the fine-grained dynamics of Fourier crystallization, spectral compression during the phase transition, the relationship between weight-space structure and activation-space causal variables, and the degeneracy of the causal subspace itself.

5.1 Minimal architecture for grokking

What is the simplest model that groks? We compare a depth-2 deep matrix factorization (DMF) $y = W_2 W_1 x$ against variants with nonlinearity (DMF + ReLU), a linearized transformer (identity attention + ReLU MLP), a softmax-only transformer (no MLP), and the full single-layer transformer. All are trained on modular multiplication ($\mathbb{Z}_{113}$) with weight decay 1.0 for 50,000 epochs.

Three findings emerge. First, **nonlinearity is necessary**: the linear DMF never generalizes despite 50,000 epochs. Its nuclear norm drops from 7,340 to 3,877—spectral compression occurs—but test accuracy remains $\sim$0%. Compression alone is not sufficient for grokking; the nonlinear activation function that enables the Fourier representation is essential.

Second, **attention is not necessary for grokking on modular multiplication**: DMF + ReLU groks 18$\times$ faster than the full transformer. The softmax attention mechanism adds enormous overhead to the grokking process without being required for the underlying algorithm on this operation. Whether this extends to other operations or multi-task settings remains open.

Third, the **linearized transformer groks without Fourier structure**: it achieves 100% test accuracy by epoch 14,000 but its Fourier alignment remains at 0.000 throughout. Grokking and Fourier representation

Architecture	Grok epoch	Fourier align.	Groks?
DMF + ReLU	∼2,000	0.095	Yes (18× faster)
MLP-only	∼2,500	0.99	Yes
Linearized TF	∼14,000	0.000	Yes (no Fourier)
Full transformer	∼37,000	0.10→0.54	Yes
Linear DMF	—	0.078	**No**
Softmax-only	—	0.103	**No**

Table 7: Grokking speed hierarchy. Nonlinearity is necessary (linear DMF never groks despite nuclear norm dropping from 7,340 to 3,877). Attention is not (DMF + ReLU groks 18× faster). Softmax alone is insufficient.

Epoch	Test accuracy	Fourier alignment
37,000	65.6%	0.070
37,500	91.3%	0.093
38,000	98.6%	0.143
38,500	99.7%	0.260
39,500	99.7%	0.498
40,000	99.8%	0.543

Table 8: Fine-grained Fourier trajectory for modular multiplication. Test accuracy crosses 90% at epoch 37,500; Fourier alignment crosses 0.30 approximately 1,000 epochs later. The model generalizes *before* its weights crystallize into the Fourier basis.

are dissociable—grokking can occur via non-Fourier algorithms. The Fourier representation is one route to generalization, not the only one.

5.2 Fourier structure emerges after generalization

Prior work identified Fourier representations as the mechanism underlying grokking in modular arithmetic (**??**). A natural assumption is that Fourier structure *drives* grokking: the model discovers the Fourier basis, then generalizes. Our fine-grained trajectory analysis (1,002 checkpoints at 100-epoch resolution) reveals the opposite: **generalization precedes Fourier crystallization**.

Test accuracy crosses 90% at epoch 37,500, but Fourier alignment does not cross 0.30 until approximately epoch 38,500—a lag of ∼1,000 epochs. The model generalizes before its weights fully organize into the Fourier basis.

The AGOP double-peak. Tracking the Average Gradient Outer Product (AGOP) Fourier alignment through training reveals an unexpected non-monotonic pattern. For the full transformer:

The AGOP alignment peaks at 0.894, then *dips* to 0.662 during the actual grokking transition (epoch 35,000), before crystallizing at 0.974 post-grokking. The dip coincides with the period of rapid generalization—the phase transition temporarily disrupts the gradient's Fourier structure before it stabilizes. By contrast, the DMF's AGOP alignment remains flat at 0.06–0.08 throughout 50,000 epochs, confirming that the linear model never discovers Fourier structure even in its gradients.

The key implication: the gradient "knows" the Fourier basis ∼10,000 epochs before the weights do (AGOP FA is 0.894 at epoch 30,000 while weight FA is still 0.075). Grokking may be the process by which weight-space catches up to gradient-space structure.

Fourier structure lives in attention, not MLPs. In a grokked 1-layer model ($p = 113$), we measure Fourier selectivity across all components: 0 of 512 MLP neurons are Fourier-selective, and the embedding has a nearly flat Fourier spectrum (participation ratio = 3.93). The Fourier structure resides entirely in attention: the QK circuit selects individual frequency pairs (rank-1 structure), while the OV circuit operates

Epoch	Test acc.	AGOP FA	Weight FA	AGOP erank
5,000	0.4%	0.108	0.080	10.0
25,000	6.3%	0.712	0.081	9.8
30,000	19.1%	**0.894**	0.075	8.7
35,000	95.8%	0.662 (dip)	0.270	5.4
40,000	100%	**0.974**	0.864	4.7
45,000	100%	0.888	0.977	4.4

Table 9: AGOP trajectory for modular multiplication. The Fourier alignment of the gradient outer product shows a double-peak: 0.894 at epoch 30,000, followed by a dip to 0.662 during the grokking transition (epoch 35,000), then crystallization at 0.974 (epoch 40,000). The *gradient* discovers Fourier structure before the *weights* do (weight FA is 0.075 when AGOP FA is 0.894).

Method	Site	Best k	IIA
SVD(FF)	attn_out	32	**0.999**
DAS	attn_out	8	0.150
DAS	mlp_out	8	0.631
SVD(FF)	mlp_out	64	0.444

Table 10: SVD of the feedforward product matrix vs. DAS on a grokked multiplication model. At the attn_out site, SVD achieves IIA = 0.999 *with no optimization* (DAS: 0.150 after 200 gradient steps). The Grassmannian distance between SVD and DAS subspaces is 1.5–8.2 radians (near-orthogonal).

in a broader k-dimensional subspace of Fourier modes. This decomposition is consistent with the "clock" interpretation of **?**.

5.3 Spectral compression during grokking

The factored feedforward product matrix $W_{\text{FF}} = W_{\text{in}}W_{\text{out}}$ undergoes dramatic rank collapse during grokking. Tracking the effective rank (exponential of the Shannon entropy of the normalized singular value spectrum) through training:

Epoch	FF effective rank	Stage
34,000	48	Pre-grokking
37,500	14	Transition
41,000	9	Post-grokking

The 5.3× compression from effective rank 48 to 9 confirms that grokking compresses the model's representational capacity into a low-dimensional subspace. This compression is necessary but not sufficient: the linear DMF also achieves spectral compression (nuclear norm 7,340 → 3,877) without grokking (§5.1). What distinguishes successful grokking is that the compressed subspace acquires *structure*—specifically, equivariance under the operation's symmetry group.

5.4 Weight-space SVD predicts causal geometry

If grokking compresses the solution into a low-rank subspace, can we read the causal variable directly from the weight matrices without fitting DAS? We compare SVD of the feedforward product matrix against DAS fitted to activations, measuring IIA for both.

At the attention output site, SVD(FF) achieves IIA = 0.999 at $k = 32$ with *zero optimization*—no gradient steps, no interchange training—while DAS achieves only 0.150 at $k = 8$ after 200 gradient steps. **Caveat**: the comparison is not dimension-matched. SVD requires 4× the subspace dimension, reflecting that the top singular vectors of W_{FF} include directions relevant to the causal variable but also spectral directions that are not causally active. A fair evaluation requires comparing both methods at matched k values, which we leave

for future work. The current result shows that the causal variable is *readable* from weight-space structure without optimization, not that SVD is a superior method to DAS.

The Grassmannian distance between the SVD and DAS subspaces ranges from 1.5 to 8.2 radians—they are near-orthogonal. This is consistent with the subspace degeneracy finding (§5.5): many different subspaces carry the same causal information. The SVD subspace is one valid parameterization; the DAS subspace is another. Neither is "correct"—both are equivalent points in the orbit of the gauge group acting on $\mathrm{Gr}(k, d)$.

The site dependence is notable: SVD(FF) is most informative at attn_out (where the attention output has been computed) but weaker at mlp_out (where the MLP has further mixed the representation). DAS shows the opposite pattern, achieving 0.631 at mlp_out. This suggests that weight-space and activation-space methods have complementary strengths: SVD captures the structural potential of the weight matrices, while DAS captures the functional use of that structure under the data distribution.

5.5 Subspace identity is degenerate

If grokking produces a specific causal subspace, we might expect different random seeds to converge to the *same* point on $\mathrm{Gr}(k, d)$. They do not. Ten seeds of modular multiplication all grok successfully (IIA 0.94–1.00), but the recovered DAS subspaces are nearly orthogonal: pairwise overlap ≈ 0.008, where random k-dimensional subspaces of $\mathbb{R}^{128}$ would have expected overlap $k/d = 0.016$. The grokked subspaces are *less* aligned than chance—consistent with the null distribution of random subspaces on $\mathrm{Gr}(2, 128)$, where pairwise geodesic distances concentrate around $\pi/2 \cdot \sqrt{k} \approx 2.22$ radians (the observed range is $[1.93, 2.19]$).

The basin of attraction is enormous. Perturbing the DAS subspace by rotating it 1+ radians on $\mathrm{Gr}(k, d)$ barely drops IIA. The causal variable is not a fragile low-dimensional needle in a high-dimensional haystack—it is a robust feature of the representation that can be read out from exponentially many subspaces.

Spectral graph structure. Computing the pairwise Grassmannian distance between all 10 seeds' subspaces yields a nearly uniform distance matrix: $d_{\mathrm{Gr}} \in [1.93, 2.19]$. The subspaces are equidistant from each other, forming a "spectral simplex" on the Grassmannian rather than clustering around a preferred direction.

This finding has a concrete implication for mechanistic interpretability: **the identity of a causal subspace is not a meaningful property**. What matters is the equivariance of the learned function *within* the subspace, not the subspace itself. Two analyses of the same model at different random seeds would find orthogonal DAS solutions—but both solutions are equally valid. The model's computation defines a large equivalence class on $\mathrm{Gr}(k, d)$: many subspaces carry the same causal information, analogous to the rotational degeneracy in attention circuits where $W_Q \to W_Q R$, $W_K \to W_K R$ leaves $W_Q W_K^\top$ invariant.

6 Discussion

6.1 A hierarchy of causal validity

Our results establish a four-level hierarchy for evaluating causal variable claims, where each level adds a stronger requirement:

1. **Interchange effectiveness** (IIA; necessary, not sufficient): High IIA means the method supports interchange but does not distinguish genuine causal structure from memorization (squaring achieves IIA = 1.0 at $k = 4$ without grokking) or from degenerate encoder-decoders (NL-DAS achieves IIA = 1.0 at $k = 1$ via lookup tables).

2. **Interchange faithfulness** (diversity ratio, reconstruction; necessary for nonlinear methods): A faithful intervention modifies the causal variable while preserving all other information. The diversity ratio ρ (Eq. 5) and reconstruction MSE detect lookup-table failure modes. Without faithfulness, "high IIA" means only "the decoder can produce the right output," not "the encoder identified the right variable." Linear DAS is faithful by construction (the projection $QQ^\top$ preserves the orthogonal complement), but nonlinear methods must demonstrate faithfulness empirically.

3. **Distributional fidelity** (KL, JS, continuous metrics; distinguishes quality among effective methods): Among methods achieving IIA ≈ 1.0, continuous distributional metrics reveal how much of the clean distribution is preserved versus distorted. A method that flips the top-1 prediction (high IIA) while scrambling the rest of the distribution (high KL) is less valid than one that reproduces the full counterfactual distribution.

4. **Structural diagnostics** (equivariance, cross-distribution generalization): High equivariance means the recovered variable transforms coherently under the operation's symmetry group—the strongest evidence that the method has found a genuine causal variable rather than exploiting distributional artifacts. Cross-distribution generalization (disjoint names, novel templates) provides the analogous test for natural-language tasks that lack algebraic symmetries.

The hierarchy has an important asymmetry. Linear DAS automatically satisfies level 2 (faithfulness) because the orthogonal projection modifies only the subspace component, but it may fail level 1 when the causal variable is nonlinear. Unconstrained nonlinear methods easily satisfy level 1 but fail level 2—they can achieve perfect IIA by learning degenerate solutions. The Structured pi-SAE satisfies both because the ELBO provides the reconstruction and regularization constraints that prevent degeneracy.

6.2 Grokking enables causal structure—linear or nonlinear

The multi-condition atlas shows that grokking is the process by which a model's internal representation transitions from an unstructured memorization encoding to a structured causal encoding. The stochastic grokking discovery—where the same operation groks at one prime or seed but not another—provides the cleanest evidence for this claim, because it controls for everything except the generalization transition itself. The multi-modulus sweep strengthens this evidence: an operation can be "never Grassmannian" at p=53 and "always Grassmannian" at p=211, with the transition governed by whether sufficient data exists to drive grokking. For operations with linear algebraic structure, the resulting encoding lives on $\mathrm{Gr}(k, d)$ and DAS recovers it. For operations with nonlinear structure, the encoding lives on a curved manifold and requires nonlinear methods to access.

The mechanism section (§5) reveals that this transition involves three dissociable processes: spectral compression (effective rank $48 \to 9$), Fourier crystallization (which *lags* generalization by ∼1,000 epochs), and the selection of one among exponentially many equivalent causal subspaces (subspace degeneracy). The AGOP analysis (§5.2) provides a concrete timeline: the gradient discovers Fourier structure ∼10,000 epochs before the weights do, and the actual phase transition is marked by a temporary *dip* in gradient-space Fourier alignment before crystallization.

The minimal architecture experiments (§5.1) further clarify: nonlinearity is necessary for grokking, but attention is not. DMF + ReLU groks 18× faster than the full transformer, and the linearized transformer groks without any Fourier structure at all. These findings suggest that Fourier representations are the dominant grokking mechanism for transformers with softmax attention, but grokking itself is a more general phenomenon—the transition from memorization to a structured causal encoding—that can occur through multiple algorithmic routes.

The key insight is that grokking is about *causal structure*, not *linear* causal structure. The Grassmannian perspective captures the linear case precisely, but the broader phenomenon is the emergence of any structured causal variable—linear or nonlinear—during the generalization transition.

6.3 Implications for language model interpretability

For practitioners applying DAS or nonlinear causal abstraction to language models:

1. **Never report IIA alone.** IIA is necessary but insufficient at every level: linear DAS on memorized models (squaring, IIA = 1.0) and unconstrained nonlinear methods (NL-DAS, IIA = 1.0) both achieve perfect IIA vacuously. Always report at least one faithfulness metric (diversity ratio for nonlinear methods, equivariance or cross-distribution generalization for any method).

2. **Use continuous distributional metrics.** KL divergence and normalized logit difference distinguish among methods that all achieve IIA ≈ 1.0. Binary IIA treats a 51% probability shift the same as 99%; continuous metrics capture this difference.

3. **Be suspicious of unconstrained nonlinear featurizers.** Any encoder-decoder trained end-to-end on interchange loss without reconstruction or regularization constraints can learn a lookup table. The ELBO (or an equivalent generative constraint) is necessary to prevent this failure mode.

4. **Test cross-distribution generalization.** For natural-language tasks, evaluate on disjoint entity sets or novel templates. A method that fails under distribution shift is exploiting surface regularities, not recovering causal structure.

5. **Consider nonlinear methods** when DAS IIA is low. Low IIA does not mean the causal variable does not exist—it may mean DAS is searching the wrong function class. But validate with faithfulness metrics.

6.4 Weight-space vs. activation-space methods

The SVD result (§5.4) reveals a surprising complementarity. Weight-space methods (SVD of $W_{\text{in}}W_{\text{out}}$) achieve near-perfect IIA at attn_out *without any optimization*, outperforming 1,000 steps of DAS gradient descent. This suggests that for grokking models, the causal variable is directly legible in the weight matrices—DAS's optimization is recovering structure that was already "written down" in the weights.

The practical implication is that weight-space analysis should precede activation-space methods: if SVD of the relevant weight product achieves high IIA, DAS optimization may be unnecessary. If SVD fails but DAS succeeds, the discrepancy suggests the causal variable involves computation beyond simple linear readout from the weight matrices.

6.5 Limitations

While our multi-condition sweep covers four primes and three depths, the grokking experiments remain on synthetic modular arithmetic with small transformers ($d_{\text{model}} = 128$). Whether the grokking–Grassmannian link extends to pretrained language models is an open question, though the k-sweep and equivariance diagnostics apply regardless. The Structured pi-SAE introduces a model-selection problem (architecture, β, α) that DAS avoids. The stochastic grokking finding rests on two seeds per operation at p=113; multi-seed experiments (10+ seeds) across primes are needed to estimate grokking probabilities precisely. The equivariance metric in the multi-modulus sweep had a measurement bug (tolerance too tight for larger primes); grokking status and IIA are valid, but equivariance values across primes require re-measurement. The SVD vs. DAS comparison is not dimension-matched ($k = 32$ vs. $k = 8$); matched-dimension experiments are needed for fair comparison. The GPT-2 evaluation compares methods at different capacity levels (linear DAS vs. nonlinear pi-SAE with E2E training); capacity-controlled baselines would strengthen the comparison.

7 Related Work

Causal abstraction. Causal abstraction provides the theoretical foundation for DAS (**??**). **?** scaled DAS via Boundless DAS. **?** identified interpretability illusions where dormant pathways inflate IIA. Our geometric diagnostics address a complementary problem: even within the linear intervention class, high IIA does not imply genuine linear structure.

Grokking. Grokking was introduced by **?** and mechanistically analyzed through Fourier representations (**??**). **?** frame grokking as compression; our spectral analysis (§5.3) gives this compression concrete numbers (effective rank 48 $\rightarrow$ 9, 5.3×) and shows it is necessary but not sufficient—linear DMFs compress without grokking. Our AGOP analysis (§5.2) adds temporal resolution absent from prior work: the gradient discovers Fourier structure ∼10,000 epochs before the weights, and generalization precedes Fourier crystallization by ∼1,000 epochs. Our contribution is to show that grokking has a specific consequence for causal abstraction: it

is the process that converts non-Grassmannian representations into Grassmannian ones (for linear operations) or into structured nonlinear ones (for nonlinear operations). Stochastic grokking has been observed via the slingshot mechanism (**?**) but not connected to causal geometry.

Spectral analysis of neural networks. Random matrix theory has been applied to weight matrices for understanding training dynamics (**?**). Preliminary stratum classification (classifying heads by spectral gap and effective rank) has discriminative power for single-task grokking models but loses all discriminative power under superposition in pretrained models. Task-conditioned spectral analysis may bridge this gap but remains untested.

Disentangled representations and causal variables. **?** introduced semi-supervised VAEs for disentanglement. Recent work on identifiable representations (**?**) establishes conditions under which latent factors can be recovered. **?** proved that unsupervised disentanglement is impossible without inductive biases; our structured prior is precisely the auxiliary supervision their theorem requires. Our application differs in that we evaluate disentanglement through causal interchange (IIA), not reconstruction quality, connecting the VAE literature to causal abstraction.

Representation geometry. Grassmannian methods have been applied to subspace tracking and domain adaptation (**??**). Linear representation hypotheses in interpretability (**???**) argue that many concepts are encoded in linear subspaces; our work tests this hypothesis causally rather than correlationally.

Nonlinear causal abstraction. Concurrent work on nonlinear interchange interventions (**?**) uses neural network featurizers before applying DAS. Our results show that this approach is vulnerable to the lookup-table failure mode (§4.7): unconstrained featurizers achieve perfect IIA by learning degenerate encoder-decoders. Our Structured pi-SAE approach uses the ELBO to provide the reconstruction and regularization constraints that prevent this degeneracy, connecting to the identifiability literature (**??**) which establishes that auxiliary supervision is necessary for recovering latent factors. The NL-DAS failure mode and our faithfulness metrics provide empirical evidence for this theoretical requirement in the causal abstraction setting.

8 Conclusion

DAS searches for causal variables in a specific function class: linear subspaces on the Grassmannian. Our atlas of 14 operations across four primes and three depths maps where this search succeeds and where it fails. The boundary is not a fixed property of the operation: it depends on dataset size, model depth, and random initialization, with only the extremes (squaring/cubing vs. addition/multiplication) robust across all conditions.

Four implications follow. First, IIA should never be reported alone: IIA = 1.000 is compatible with a genuine causal variable, a memorized lookup table, and a degenerate encoder-decoder. The four-level hierarchy we propose—interchange effectiveness, interchange faithfulness, distributional fidelity, and structural diagnostics—provides a complete evaluation framework.

Second, nonlinear extensions of causal abstraction require generative structure. Unconstrained nonlinear featurizers achieve perfect IIA vacuously; the Structured pi-SAE's combination of ELBO, label-conditional prior, and L_1 sparsity prevents this failure mode while generalizing to standard GPT-2 benchmarks (strict IIA ≥ 0.90 on 5/6 tasks).

Third, single-condition evaluations of grokking behavior are unreliable. Operations classified as "never Grassmannian" at one prime may grok at another (bitwise XOR groks at all four primes; sum of squares at three of four). The multi-condition atlas is essential for distinguishing robust non-grokkers from operations that simply need more data or capacity.

Fourth, grokking's creation of causal structure is more nuanced than previously understood. Nonlinearity is necessary but attention is not (DMF + ReLU groks 18× faster). Fourier structure *follows* generalization rather than driving it, and the gradient discovers this structure ∼10,000 epochs before the weights crystallize.

The causal subspace itself is degenerate—10 seeds produce orthogonal subspaces that all support equally valid causal interventions—suggesting that *functional* diagnostics (equivariance, intervention faithfulness) are more robust than *structural* diagnostics for evaluating causal geometry.

The linearity assumption underlying DAS is not a benign default. It is a testable hypothesis whose validity depends on the dataset, model, and operation—not just the operation alone. The natural fix—adding nonlinearity—introduces a new failure mode that is invisible to the standard evaluation metric. This paper provides the diagnostics to detect both problems, the mechanistic analysis to understand why they arise, and the structured nonlinear method that avoids them.

Data and code. All code for model training, DAS fitting, structured VAE training, geometric diagnostics, and figure generation is available at `[repositoryURL]`.