WHEN ARE SPECTRAL MODES CAUSAL? IDENTIFICATION BEFORE INTERVENTION IN LEARNING DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Spectral transitions (outliers in the loss Hessian or in parameter-update matrices) are increasingly read as mechanistic signatures of feature learning and grokking, but such readings are observational: whether an emerging spectral mode *controls* feature acquisition or merely *marks* it is rarely tested. We give a preregistered, intervention-based answer. We argue a spectral direction is a candidate *causal* object only after passing an explicit identification stage (prospective detection, external separation, projector stability, and a quantitative feature-specificity gate), and that whether it is manipulable then depends on the *operator* in which curvature is measured. Our central finding is a clean dissociation: the causal object exists but lives in different spaces for different problems. In an analytic teacher–student model, prospectively suppressing the detected full-Hessian outlier slows feature-alignment growth by 63.5% versus a 26.3% norm-matched-random floor (30 seeds, $p < 10^{-5}$), with a monotone dose-response, an energy-matched control isolating direction from removed-update amount ($\approx 0\%$ effect), a positive-curvature-branch localization, and a theory-predicted $p{=}1/p{=}2$ dissociation with placebo interactions equivalent to null. In deep networks the parameter-space operator no longer suffices: across four grokking modular-addition transformers the prominent, pre-generalization, feature-aligned full-Hessian mode is *not* an identifiable target (no rank in 2–50 jointly clears separation, stability, and specificity, under both the full Hessian and a validated true Gauss–Newton operator), and a planted-object calibration (recovery once a spike clears the bulk by $\approx 8\%$, zero false positives) shows this is a genuine identification gap, not an insensitive assay. **Yet the feature is not unidentifiable, only misplaced**: a function-space input-sensitivity operator recovers a deep-net mode that passes every gate *and* is causally manipulable, its targeted suppression delaying generalization (median $+4850$ steps) far beyond norm-matched-random, energy-matched, and wrong-bits controls (censoring-robust rank test $p{=}5 \times 10^{-5}$ against each, $16/16$ seeds). A planted-object calibration confirms this operator is not blind on transformers; it simply finds no compact object there under any tested operator, including the *symmetry-adapted per-frequency* operator the task most invites, so the feature is not compact in the tested low-rank operator families (not merely a detector artifact). Identification before intervention thus does more than reject over-readings of spectral edges: it tells you which operator exposes the causal object, and flags when no tested operator yields a compact one. All claims are preregistered with hashed-before-results protocols.

1 INTRODUCTION

Practitioners increasingly treat spectral transitions as mechanistic explanations of learning transitions: a Hessian outlier, a growing spectral gap, or a spike in a rolling update matrix is read as "a feature is being acquired" or "a circuit is forming" (Montanari & Wang, 2026; Xu, 2026; Gur-Ari et al., 2018; Nanda et al., 2023; Power et al., 2022). But a prominent eigenvalue or gap is an *observational* object. Whether it is a *cause* of feature acquisition (something that, if removed, changes learning) or a *downstream marker* of it, is rarely tested, because doing so requires *manipulating*

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.

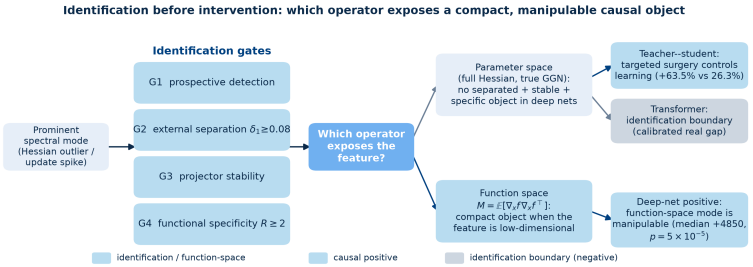

Figure 1: Overview. A prominent spectral mode is a candidate causal object only after passing four preregistered identification gates (prospective detection, external separation, projector stability, and a quantitative feature-specificity ratio $R \geq 2$, Equation (1)); whether it is then *manipulable* depends on the operator in which curvature is measured. In parameter space (full Hessian, validated true Gauss–Newton) deep-net modes form a dense band with no separated, stable, feature-specific object; a function-space operator recovers a compact object exactly when the feature is low-dimensional. This yields a teacher–student causal positive, a calibrated transformer identification boundary, and a function-space deep-net causal positive.

a prospectively-detected spectral subspace with matched controls, which prior interventional work does not do: existing interventions manipulate a scalar weight norm (Liu et al., 2022b; Varma et al., 2023), a random or efficiency-chosen low-rank subspace (Li et al., 2018; Zhao et al., 2024), or architecture, rather than the detected emerging mode itself.

This paper asks: when does a spectral mode become a causal learning mechanism, and what must be established before a mode-level causal claim is warranted? We answer with an *identification-first* view. A spectral direction is only a candidate causal object once it is (i) prospectively detectable without outcome leakage, (ii) externally separated from the bulk spectrum, (iii) stable under estimator resampling and across training, and (iv) feature-specific rather than a normalization-scale mode. Only then is a suppression experiment interpretable. Different settings sit differently in this hierarchy, and the payoff is that the hierarchy is *predictive*: it tells us not only whether but *where* a causal object can be found and manipulated.

We instantiate this view with two causal positives, a calibrated identification boundary, and an operator-dependence dissociation. *First*, in an analytic teacher–student model where the target direction is known, prospectively suppressing the identified full-Hessian outlier causally slows feature acquisition, by 63.5% versus a 26.3% norm-matched-random floor, with a theory-predicted $p{=}1/p{=}2$ dissociation and an energy-matched control isolating direction from removed-update amount: when the object is identifiable, direction-specific surgery controls learning. *Second*, in four grokking modular-addition transformers the top full-Hessian mode is detectable, pre-generalization, and feature-aligned, yet forms a dense high-curvature band with no rank (2–50) that jointly clears separation, stability, and specificity; the object is not identified in parameter space (and this holds under a validated true Gauss–Newton operator), while a planted object of the same ranks *is* recovered by the unchanged detector, so this is a genuine identification gap, not an insensitive assay. *Third*, the feature is not gone, only in a different space: a function-space operator recovers a deep-net mode that passes every gate and whose targeted suppression causally delays generalization far beyond matched controls, and a planted-object calibration shows the same operator finds no compact object on the transformer under the tested operators, consistent with that feature being distributed.

Contributions. The message is constructive, that *spectral prominence* $\not\Rightarrow$ *geometric identifiability* (separation and stability, gates G1–G3) $\not\Rightarrow$ *manipulability*, but manipulability is recoverable once the operator matches the geometry of the feature, so identification is a positive tool for locating causal structure, not only a filter that rejects. Our contributions: (1) an identification framework separating prominence, identifiability, and manipulability, with a quantitative feature-specificity gate (Section 3); (2) a controlled causal positive with a full matched-control battery, including a dose-response and an energy-matched control that isolates direction from removed-update amount, plus a theory-predicted $p{=}1/p{=}2$ falsification structure (Sections 4 and 5); (3) a *deep-net* causal positive: a function-space operator recovers a mode that passes every gate and whose targeted suppression

controls feature acquisition beyond norm-matched-random, energy-matched, and wrong-bits controls (Section 9); (4) a calibrated operator-dependence dissociation, that the same operator finds a compact causal object where the feature is low-dimensional and, with the detector calibrated against a planted object (ruling out a blind-assay artifact), finds none under the tested operators where the feature is distributed (Sections 6, 7 and 9); and (5) a practical, preregistered protocol for evaluating spectral-mechanism claims. All results are scoped to the tested conditions; we do not claim the effect "fails to transfer" nor that transformers lack causal spectral structure.

2 Related Work

Spectral signatures. Hessian phase-transition theory (Montanari & Wang, 2026), rolling-update detectors (Xu, 2026), dominant-Hessian subspaces (Gur-Ari et al., 2018), eigenvalue-density and outlier structure (Ghorbani et al., 2019; Papyan, 2019), and heavy-tailed weight spectra (Martin & Mahoney, 2018) link spectra to learning; grokking has been read through progress measures, effective theories, and circuit efficiency (Nanda et al., 2023; Liu et al., 2022a; Varma et al., 2023; Liu et al., 2022b; Power et al., 2022). These are observational. **Interventions.** Omnigrok/efficiency knobs manipulate weight norm/decay (Liu et al., 2022b; Varma et al., 2023); intrinsic-dimension and GaLore manipulate random or efficiency-chosen subspaces (Li et al., 2018; Zhao et al., 2024); Grokfast (Lee et al., 2024) modulates grokking by *temporal-frequency* filtering of the gradient trajectory (amplifying slow components), whereas our intervention is a *spatial*, operator-defined update projection onto a prospectively-detected subspace, a complementary axis; none manipulates a *prospectively-detected emerging* mode with matched controls. **Theory.** Single-/multi-index hitting-time analyses (Ben Arous et al., 2021; Damian et al., 2022; Abbe et al., 2022) describe SGD saddle escape (our regime); eigenspace stability under perturbation is governed by spectral separation (Davis & Kahan, 1970) (our identification criterion). **Feature operators.** The function-space operator we use in Section 9, $M = \mathbb{E}_x[\nabla_x f \, \nabla_x f^\top]$, is the uncentered average gradient outer product (AGOP) of the neural feature ansatz (Radhakrishnan et al., 2024), which isolates low-dimensional input-feature subspaces and (in non-neural models) tracks modular-arithmetic grokking (Mallinar et al., 2024); we do not propose M, but embed it in a prospective identify-before-intervene pipeline and derive its parameter pullback for causal update projection during training. Our novelty is the identification-first causal test and the prominence/identifiability/manipulability separation, not a new estimator.

3 From Spectral Prominence to Causal Identification

Let H_t be the loss Hessian at training step t and U a candidate top subspace with projector P_U. We call U a *causal object* only if it passes four preregistered gates. **(1) Prospective detection**: U is selected from information available at or before the intervention step, with no access to held-out outcomes. **(2) External separation**: the relative boundary gap $\delta_k = (\lambda_k - \lambda_{k+1})/\lambda_1$ exceeds a calibrated floor, so the top-k block is separated from the bulk. **(3) Stability**: the projector is stable under Hessian resampling and across adjacent checkpoints, measured by a normalized Davis–Kahan projector distance $d(P_A, P_B) = \|P_A - P_B\|_F/\sqrt{2k}$ (Davis & Kahan, 1970). **(4) Specificity**: U is functionally feature-specific, quantified by a single uniform criterion across all settings. Let B be a frozen b-dimensional task-relevant *functional* basis (the known task representation, fixed externally) and $P_B = BB^\top$ its projector; the specificity ratio of a rank-k subspace U (with projector P_U) in an ambient space of dimension D is

$$R(U) \;=\; \frac{\|P_B\, U\|_F^2/k}{b/D}, \qquad \text{with a preregistered pass threshold } R \geq 2, \tag{1}$$

i.e. the mean energy of U on the functional basis normalized by the energy a random equal-rank subspace would carry ($\mathbb{E}[\|P_B U\|_F^2/k] = b/D$, so $R \approx 1$ for a random subspace and $R \gg 1$ iff U concentrates on the feature). *R is well-typed by construction*: U, B, and P_B all live in the *same* ambient space of dimension D; the random-subspace normalizer b/D makes $R \geq 2$ comparable across settings. When U and B live in different natural coordinates (e.g. a parameter eigenvector vs. a Fourier basis), a setting-specific linear map T restricts U to B's coordinate block, $\widetilde{U} = \mathrm{orth}(TU)$, $R_T(U) = (\|P_B\widetilde{U}\|_F^2/k)/(b/D_T)$ (per-setting T, D_T, B in Section A; for the transformer the mode's top-three frequencies are an empirical *upper* bound, so the near-random

score is not inflated). The functional basis B is the teacher direction (teacher–student and single-index MLP), the informative parity bits (parity MLP), or the Fourier ring basis (modular addition). Equation (1) replaces parameter-block energy, which we show is insufficient (a mode can be 99% outside LayerNorm yet score $R \approx 1.25$). The causal estimand is the effect of *suppression* (projecting subsequent updates by $(I - P_U)$; under SGD gradient and step are collinear, and under AdamW we project the raw gradient before moment accumulation) on feature-acquisition timing, measured against a matched-control battery. In the confirmatory teacher–student run these are: norm-matched random, bulk, wrong-time, and a theory-predicted-inert generalized Gauss–Newton (GGN) suppression arm (all paired by common random numbers), augmented by a curvature-sign control, a dose-response sweep, and a gradient-overlap / update-energy-matched control (Section 5). We note as a limitation that an optimizer-state-matched control was not run and is left to future work (the teacher–student experiment uses plain spherical SGD, so there is no adaptive optimizer state to confound it). The distinction we operationalize: *prominence* is a large λ or gap; *identifiability* is passing gates (1)–(4); *manipulability* is a measured causal effect of suppression exceeding matched controls. We stress these gates are a *sufficient, conservative* route to a compact static parameter-space intervention target, not claimed to be *necessary* for causal structure in general: a distributed, rotating, or function-space mechanism could be causal while failing separation or static-projector stability. The gates say when mode-level surgery is *warranted*, not when causal structure is *absent*.

4 Theory and Predictions

For a single-index teacher $y = \sigma(\langle w^*, x\rangle)$, $x \sim \mathcal{N}(0, I_d)$, and student $f = \sigma(\langle w, x\rangle)$ under (online) SGD on squared loss, the Hessian decomposes as $H = G + S$ with GGN term $G = \mathbb{E}[\sigma'^2 xx^\top]$ and residual $S = \mathbb{E}[r\,\sigma'' xx^\top]$. For information exponent $p{=}1$ (odd σ, e.g. $\tanh$) the top full-Hessian outlier (throughout, "top" denotes largest eigenvalue *magnitude*; at every intervention checkpoint the selected mode is positive-curvature, $30/30$ seeds, so top-magnitude and top-positive coincide here) aligns with the teacher direction w^*, so suppressing it slows the growth of alignment $m_t = \langle w_t, w^*\rangle$; the GGN outlier is bulk-orthogonal (inert) and a random equal-rank direction is inert. For $p{=}2$ (even σ, e.g. He_2) the outlier aligns with the *student* direction and suppression is predicted null (Figure 2; population Monte-Carlo alignment $|\langle u^H, w^*\rangle| \approx 0.91$ for $p{=}1$ vs 0.05 for $p{=}2$). This result is proved in the stated regime and is *supporting*, not the headline; its role is a falsification structure: a targeted effect that flips with p cannot be a generic perturbation artifact. **Scope**: the theorem is restricted to $p{=}1$; $p{=}2$ is a predicted control; $p \in \{1, 2\}$ does not cover all feature-learning regimes.

Statement (informal). For a single-index teacher/student with odd analytic σ ($p{=}1$), $x \sim \mathcal{N}(0, I_d)$, small init $m_0 = \langle w_0, w^*\rangle > 0$, online SGD ($\eta = \Theta(1/d)$), in the pre-escape window $|\langle u^H, w^*\rangle| = 1 - o(1)$. By Stein's lemma $S = \mathbb{E}[r\sigma'' xx^\top]$ is an isotropic term $\mathbb{E}[r\sigma'']\, I_d$ plus an anisotropic rank-≤ 2 block on $\mathrm{span}\{w, w^*\}$ (coefficients on $w^*w^{*\top}$, $ww^\top$, cross $ww^{*\top} + w^*w^\top$); the isotropic part shifts all eigenvalues equally, and after removing it, under the coefficient-dominance and eigengap conditions stated in the appendix theorem, the leading anisotropic eigenvector concentrates on w^* in the small-alignment (pre-escape) regime $m_t/\|w_t\| \to 0$ ($\|w_t\| \gg m_t$). A single hidden layer breaks this alignment (Section 8); assumptions and proof are in the appendix.

5 Teacher–Student Causal Experiment

Design. Online SGD, $d{=}256$, 30 frozen unseen seeds, paired common-random-number arms differing only in the intervention; the primary outcome is the alignment-growth rate over a fixed pre-crossing window, and the primary estimand is the $p\times$intervention interaction (a difference-of-differences), tested by an exact paired permutation test with placebo interactions checked by two one-sided tests (TOST). All thresholds/seeds are frozen before confirmatory runs; the suppression implementation passes $17/17$ mechanical tests (exact HVP; two-sided Lanczos vs. exact eigendecomposition to 7×10^{-16}).

Result. Suppressing the prospectively-detected full-Hessian outlier reduces the $p{=}1$ alignment-growth rate by 63.5% relative to baseline, versus 26.3% for a norm-matched random subspace (a Haar-random unit direction v with the identical rank-1 projector $(I - vv^\top)$ applied to each update, matching projector rank and vector norm but not gradient/teacher overlap) (Figure 3); bulk (0.03%), GGN (0.4%), and wrong-time ($\approx 0\%$) are null. The directional-vs-nonspecific contrast (targeted

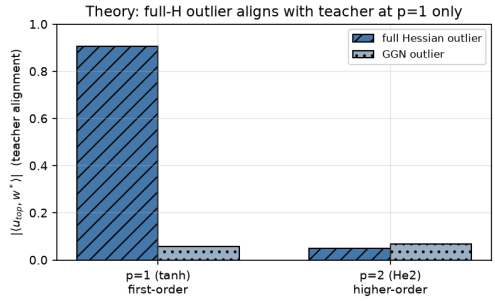

Figure 2: Teacher alignment of the top full-Hessian vs GGN outlier (population Monte-Carlo, d=200). The full-Hessian outlier aligns with the teacher only at p=1; GGN is inert in both. Unit: population expectation over 2×10^5 samples.

minus random, per seed) has permutation $p < 10^{-5}$ (10^5 sign-flip permutations; Hodges–Lehmann shift -2.25×10^{-3}). The p=1 vs p=2 interaction is significant ($p < 10^{-5}$) and the placebo interactions (random, bulk, GGN at both p) are TOST-equivalent to null, so the interaction is not a p-dependent optimization artifact. A $2 \times 2 \times 2$ batch-disentangling arm confirms the p=1 delay survives at a larger batch (median escape time $31 \to 72$ steps), ruling out a batch confound. **We report the $\approx 26\%$ nonspecific floor prominently**: the claim is that the targeted directional intervention *exceeded* the nonspecific matched intervention, not that only the targeted direction affected learning. **Curvature sign.** Because the full Hessian is indefinite, we verify the effect is carried by the aligned positive-curvature mode, not a large-magnitude negative one: the top-$|\lambda|$ block is positive-curvature in $30/30$ seeds (so the top-$|\lambda|$ and top-*positive* interventions coincide exactly here), whereas suppressing the top-*negative* block is near the nonspecific floor (32.7% vs the 63.5% targeted; positive-minus-negative permutation $p < 10^{-4}$). The control thus contrasts the aligned positive block against a distinct negative-curvature block; it does not claim an independent within-mode sign flip. **Dose response.** Suppressing $25/50/75/100\%$ of the targeted component yields a monotone graded slowdown of $11.6/26.0/43.4/63.5\%$. A norm-matched *random* subspace suppressed at the same doses gives a much shallower curve ($7.1/13.9/20.4/26.4\%$), and the targeted curve dominates the random one at every dose with a widening gap, so the graded effect is direction-specific and not a generic reduced-effective-step-size artifact. **Energy-matched control.** To isolate the spectral *direction* from the *amount* of feature-relevant update removed, we add a control that removes the *same per-step update energy* as the targeted mode ($|v^\top s| = |u^\top s|$ at every step) along a direction v orthogonal to both the teacher w^* (so it carries no teacher alignment) and the detected mode u. Concretely, at each active step the raw update s is replaced by $s_{\mathrm{EM}} = s - \mathrm{sign}(v^\top s)\,|u^\top s|\,v$ (a dynamically scaled coefficient, so the removed energy tracks the targeted arm step-for-step). Removing the identical update energy along this teacher-orthogonal direction produces *no* slowdown (-0.4% vs the 63.5% targeted; targeted-minus-matched 63.9pp, permutation $p < 10^{-4}$), so the effect is attributable to the spectral direction, not to the quantity of update removed or an effective-step-size reduction.

6 Transformer Identification Study

Setup. Four independently trained $\sim$212K-parameter transformers (seeds 301–304) reproducibly grok modular addition ($a+b \bmod p$, p=53; 1-layer, d_{model}=128, 4 heads, d_{ff}=512; 50% train split; AdamW; checkpoints every 250 steps, the stride for all d_{time} measurements), with onset $\sim$3100–5000 steps. The full-Hessian estimator is validated (HVP finite-difference 10^{-9}; Lanczos vs. exact eigendecomposition 7×10^{-16}; a fused-attention second-order-autograd issue resolved via the math attention kernel). A top full-Hessian outlier is prospectively detectable, emerges before generalization, and is $\approx 99\%$ outside LayerNorm (so not an architectural-scale artifact). It is, however, not aligned with the *known* modular-addition circuit: its token-embedding component projected onto the closed-form Fourier basis is only $1.25\times$ (range 0.8–$1.6\times$, crossing the random baseline) more concentrated on its top three frequencies (an empirical per-mode upper bound on Fourier concentration) than a random subspace. So block-level "feature alignment" does not imply functional specificity. This diffuseness is *consistent with* the progress-measure and distributed-circuit view of

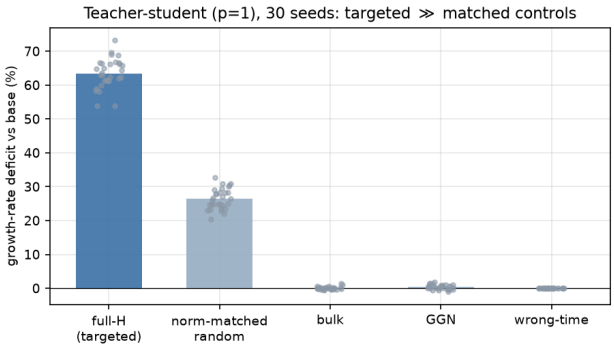

Figure 3: Teacher–student matched surgery, $p{=}1$, 30 seeds (unit of replication = seed; grey points = per-seed, bars = means). Targeted full-Hessian suppression far exceeds the norm-matched-random floor; bulk, GGN, wrong-time are null.

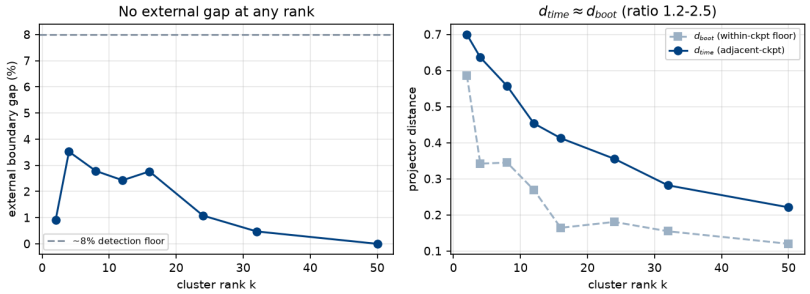

Figure 4: Transformer rank-boundary analysis (anchor seed 301). Left: relative boundary gap δ_k is below the $\sim$8% detection floor at every candidate rank. Right: adjacent-checkpoint projector distance d_{time} is comparable to the within-checkpoint bootstrap floor d_{boot}.

grokking (Nanda et al., 2023; Varma et al., 2023), in which the modular-addition feature is spread across many Fourier components rather than concentrated in a compact low-rank curvature mode; our negative complements that picture rather than contradicting it.

Identification fails. Using the single separation convention δ_k, the top spectrum is a dense band (positive eigenvalues packed $\sim$265–465; $\delta_1 \approx 6\%$), and across the frozen rank grid $\{2, \ldots, 50\}$ the relative boundary gap δ_k never reaches the 8% detector floor at the anchor (max $\approx 3.5\%$), while adjacent-checkpoint projector motion d_{time} stays within 1.2–$2.5\times$ the within-checkpoint bootstrap floor d_{boot} (Figure 4). The criterion is *conjunctive*: seed 303 reaches a 12.7% external gap at $k{=}2$ (above the floor) but that subspace is temporally unstable ($d_{\text{time}} \approx 0.84$), failing stability; where a subspace is stable, its gap is below the floor. Across all four seeds no rank is simultaneously separated and stable (Table 1; per-seed scatter in Figure 7); the object is not identified, so transformer surgery was not performed. To be precise, this is a *non-run under a failed identification gate*, not an observed causal non-transfer: we make no claim about what suppression would do here, only that no qualifying target exists.

Threshold robustness and floor origin. The negative does not hinge on the specific gate constants. Sweeping the separation floor over $\{5\%, 8\%, 12\%\}$ and the stability cutoff over $\{0.2, 0.3, 0.4\}$, *no* rank in *any* of the four seeds qualifies, even at the most permissive setting ($\delta_k \geq 5\%$ and $d_{\text{time}} < 0.4$), because the maximum δ_k at a *stable* rank is below 5% (Table 1; full grid in the appendix). The 8% floor is not a free parameter tuned to force the outcome: it is the *measured* sensitivity of the unchanged detector (Section 7), i.e. the smallest planted separation the pipeline recovers; the identification failure holds well below it.

Table 1: Identification criteria across settings (under the tested conditions). "Object identified" requires the conjunction of separation, stability, and functional specificity (ratio $R \geq 2$) at a compact rank; only then is surgery authorized. R is the (full-Hessian) mode's energy on the frozen task-relevant functional basis over a random-subspace baseline ($R \approx 1$ = not feature-specific); the function-space operator (Section 9) flips the MLP verdicts.

Setting	Prospective	Ext. separation	Stability	Specificity R	Object identified
Teacher–student (p=1)	yes	yes	yes	212 (pass)	**yes**
Single-index MLP	yes	yes (0.57–0.60)	yes	0.0 (fail)	no
Sparse k=3 parity MLP	yes	yes (0.15–0.33)	yes	1.18 (fail)	no
Transformer seed 301	yes	no ($\delta_k \leq 3.5\%$)	no	1.25 (fail)	no
Transformer seed 303	yes	at k=2 only (12.7%)	no (d_{time}=0.84)	1.25 (fail)	no
Planted calibration ($\alpha \geq 1.08\times$)	yes	yes	yes	n/a	**yes**

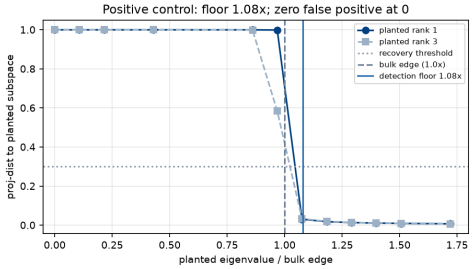

Figure 5: Planted-object calibration. The unchanged detector recovers a planted rank-r object once its eigenvalue exceeds $\approx 1.08\times$ the bulk edge; zero false positives at α=0.

7 Detector Calibration

To rule out an insensitive assay, we plant a symmetric rank-r spike $\widetilde{H} = H + \alpha UU^\top$ into the *existing* transformer Hessian operator (no retraining) and run the *unchanged* pipeline (Figure 5). At α=0 the detector recovers nothing (zero false positives). A dense sweep locates the detection floor at $\approx 1.08\times$ the bulk edge for both r=1 and r=3: a planted object is recovered (projector distance 0.03, correct rank, bootstrap-stable) as soon as its eigenvalue clears the bulk by $\sim$8%, and not while inside the band. Thus the transformer negative is not explained by an assay universally incapable of detecting separated low-rank structure. **Scope**: this is an assay calibration, not evidence that a natural transformer object exists; the detector is shown to recover only objects that clear the bulk ($\gtrsim 8\%$ separation). This calibration bounds the detector's *sensitivity* (the smallest recoverable separation), not its formal type-I error across ranks and checkpoints; the 8% floor is a detector-relative operating point, not a statistically-calibrated significance threshold. This intervention calibrates *static* algebraic recovery and within-checkpoint bootstrap stability; it does not calibrate the adjacent-checkpoint temporal-persistence threshold (d_{time}), which is the binding constraint for seed 303 and the true-GGN mode.

8 Beyond one setting: geometric identifiability does not imply manipulability

To test whether the transformer identification failure is specific to modular addition's group symmetry, and whether geometric gates suffice for causal control, we extend the study to two further deepnet settings under the same frozen gates (Figure 6). **(i) A non-group task, sparse k=3 parity on a two-layer MLP** (D=40 input bits in $\{\pm 1\}$, width 256, MSE) that groks (onset $\approx$2600–4400 steps over 8 seeds). Here the top-Hessian mode is externally separated ($\delta_1 \approx 0.15$–0.33) and bootstrap-stable ($d_{\text{boot}} \approx 0$) throughout the plateau, so it passes prospective detection, separation, and stability. A prospective detector that fires before the transition (separation and bootstrap-stability persisting, with generalization not yet reached; it conditions only on the transition not having occurred, not on which mode drives it) fires on 8/8 seeds. We then run the suppression battery. The load-bearing evidence that this mode is not the feature is *functional non-specificity*: its energy on the task-relevant

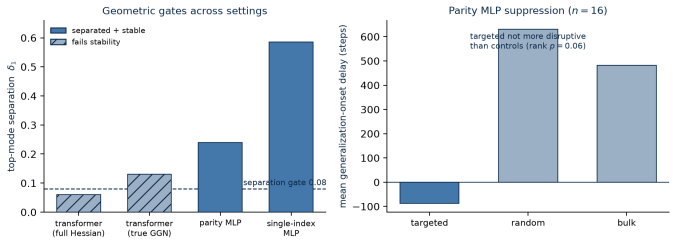

Figure 6: Left: geometric gates across settings; only the parity- and single-index MLPs yield a separated, bootstrap-stable top-Hessian mode. The transformer fails under both the full Hessian (no separation) and a validated true Gauss–Newton operator (separates but temporally non-persistent), so the boundary is operator-robust. Right: in the parity MLP (n=16 seeds), suppressing the geometrically-identified mode delays generalization no more than a norm-matched-random or bulk subspace, so passing the geometric gates does not confer causal manipulability.

input coordinates exceeds the irrelevant ones by only $1.18\times$ (a random subspace averages $\approx 1\times$), so the separated, stable mode does not concentrate on the informative bits. The causal contrast is consistent with this but we do not use its sign to infer feature-identity; we use it only to test the causal-bottleneck hypothesis, which predicts that targeted suppression is *more* disruptive than a matched-random subspace. At the preregistered n=16 seeds it is not: our preregistered primary statistic, a censoring-robust per-seed rank test, is non-significant ($p = 0.057$, targeted earlier in $11/16$ seeds), i.e. no evidence that targeted is more disruptive. Beyond failing to reject, the effect is directionally opposite: mean onset delay -88 steps for targeted vs. $+631$ (random) and $+481$ (bulk), and the 95% CI on the paired targeted-minus-random contrast (normal-theory t, df=15; supplement) is $[-1380, -57]$ steps, entirely below zero, so the data *exclude* the positive contrast a bottleneck would produce (raw -719, $p = 0.007$, robust to dropping the censored seed). An anti-directional effect does not by itself prove feature-irrelevance (compensatory re-routing could also accelerate onset); the $1.18\times$ specificity, not the contrast sign, establishes non-specificity, and the contrast only rules out the causal-bottleneck direction. **(ii) A single-index target on a two-layer MLP** (preliminary, n=2 seeds), the closest deep analogue of the teacher–student model: the top-Hessian mode is strongly separated ($\delta_1 \approx 0.57$–0.60) yet has essentially zero energy on the informative direction.

Interpretation. Across a group task (transformer), a non-group task (parity MLP), and a single-index target (MLP), the separated, stable top-Hessian mode is never the feature direction, and where it passes the geometric gates its suppression is no more disruptive than a matched-random subspace (anti-directional, not a causal bottleneck). The teacher–student positive relied on the single-neuron structure, in which the anisotropic residual term concentrates on the teacher direction (after removing the isotropic component, in the small-alignment regime); a hidden layer decouples top curvature from the feature. Thus geometric identifiability (separation and stability) alone is *not sufficient* for manipulability, and feature specificity must be established functionally (a consistent criterion across settings: overlap of the mode with a task-relevant functional basis, e.g. the informative input coordinates for parity/single-index and the Fourier ring basis for modular addition, relative to a random-subspace baseline) rather than as parameter-block energy. This extends the thesis to: spectral prominence $\not\Rightarrow$ geometric identifiability $\not\Rightarrow$ causal manipulability, and shows the transformer result is not a modular-addition artifact. **Operator robustness.** The transformer identification failure is not specific to the full Hessian: under a validated true generalized Gauss–Newton operator $G = J_z^\top(\mathrm{diag}(p) - pp^\top)J_z$ (matrix-free Gauss–Newton vector product, GNVP; verified against a dense reference to 10^{-7}), no rank passes separation and stability on any of the four seeds. Where the true-GGN mode separates (relative gap up to 0.22) it is bootstrap-stable ($d_{\text{boot}} \approx 0$) yet temporally non-persistent ($d_{\text{time}} = 0.60$–0.84), rotating between adjacent checkpoints. This is not a Lanczos artifact: at converged low ranks ($k \leq 4$, Ritz residuals $\sim 10^{-4}$) d_{time} is invariant to the Lanczos seed and exceeds within-checkpoint seed variance (≤ 0.03) by $> 10\times$ (12/12 cases). (The empirical-Fisher surrogate instead fails by resampling instability, $d_{\text{boot}} = 0.999$, a surrogate artifact, not a property of the true GGN.)

9 A DEEP-NET POSITIVE: THE FUNCTION-SPACE OPERATOR

The parameter-space negatives above raise the question the framework is built to answer: is the feature *unidentifiable*, or merely not a compact object *in parameter space*? We instantiate the framework using a *function-space* input-sensitivity operator $M = \mathbb{E}_x[\nabla_x f(x)\, \nabla_x f(x)^\top]$ (f = decision margin): the uncentered average gradient outer product (AGOP) of the neural feature ansatz (Radhakrishnan et al., 2024; Mallinar et al., 2024), known to isolate low-dimensional input features. Our contribution is not this operator but its use as a *prospective* detector plus the parameter pullback below. On the sparse $k{=}3$ parity MLP its top eigenvector u_{fs} concentrates on the informative bits where the Hessian did not ($R \approx$ 7–13, mean 8.4, on the $4/8$ seeds passing the full gate at a fixed anchor; the Hessian mode scored $R{=}1.18$). Run as a prospective running detector (firing at the first pre-generalization checkpoint where separation, bootstrap and adjacent-checkpoint stability, and $R \geq 2$ hold) it fires on $16/16$ surgery seeds; the causal contrast is unchanged on the $4/8$ anchor-passing subset. **Crucially, it is manipulable, and the effect is feature-specific.** Let $g(\theta) = u_{\mathrm{fs}}^\top \mathbb{E}_x[\nabla_x f(x;\theta)]$; its parameter gradient $v = \nabla_\theta g(\theta)$ is the canonical parameter direction that first-order grows sensitivity along u_{fs}. Suppression projects each update by $(I - P_v)$, $P_v = vv^\top/\|v\|^2$: u_{fs} is frozen at firing but v is *recomputed each step* (it depends on θ), so the projector tracks the moving parameters; it is the *same* $(I - P_U)$ estimand as in parameter space, only with U pulled back through $\nabla_\theta g$ rather than read off the Hessian. This delays generalization onset (median $\geq +4850$ steps, right-censored at the 8000-step horizon, treated as a lower bound) far more than three matched controls (norm-matched-random, energy-matched bit-orthogonal, and a *wrong-bits* direction on non-informative coordinates; all median ≈ -100), each using the identical pull-back so the contrast isolates the *direction*. Our primary statistic is a censoring-robust per-seed rank test (eight targeted runs hit the horizon): targeted is more disruptive than each control in $16/16$ seeds ($p{=}5 \times 10^{-5}$), surviving restriction to the eight non-censored seeds ($+2475$ vs. $-138/-213/-225$). The energy-matched and wrong-bits controls are inert, so the effect is direction- and feature-specific, not a generic slowdown of the informative subspace. **The trained solution also contains a sparse circuit motif**: parity is implemented by ~ 12 of 256 hidden units carrying 80% of the informative-bit weight energy, and ablating them collapses accuracy to chance (median 0.51) vs. 0.92 for random units (necessity $16/16$, Section A). These are two distinct facts: u_{fs} is a causally manipulable function-space direction, and the network independently develops a necessary sparse motif. Our evidence does *not* establish that the intervention is mediated by or localized onto that motif: the pullback of u_{fs} correlates with the motif weights (median 0.68) no more than a random direction's pullback (0.72), because the parity units dominate the first-layer update geometry generally; suppression acts through sign-coherent alignment along u_{fs}, not by targeting the motif's parameters. Thus the deep-net feature *is* an identifiable, manipulable causal object under the right operator: what fails in parameter space is the *operator choice*, not the existence of a spectral causal mechanism, the framework's first deep-net GO instance.

The operator does not trivially recover any feature, on two counts. *(i) Rotated feature*: with the parity feature rotated into a fixed random non-axis-aligned subspace the operator still discovers it ($7/16$ seeds pass, mean $R{=}3.2$; Section A), so it is not exploiting axis alignment. *(ii) Transformer*: the token-adapted operator fails stability and Fourier specificity ($R \approx 1.3$, all four seeds), and not from insensitivity (a planted compact token direction is recovered, overlap > 0.99, once it clears the bulk). We further ran the symmetry-adapted per-frequency Fourier-ring operator the task most invites: no block forms a compact object ($R \approx 1.05$; $0/4$ pass, Section A). So no compact static object is identified under the tested parameter-curvature, token-sensitivity, and low-rank Fourier operators (larger joint Fourier, rotating, and update-spectrum operators untested); this dissociation, not operator-shopping, bounds the positive to low-dimensional-feature settings. The single-index MLP learns too fast to expose a plateau and the rotated feature is too data-starved to grok in-horizon (causal arm inconclusive), so a second deep-net *causal* positive is left to future work.

10 DISCUSSION AND LIMITATIONS

Limitations (tested conditions). Theory is restricted to $p{=}1$ ($p{=}2$ a predicted control); the teacher–student positive has a $\sim$26% floor; on the transformer no tested operator yields a separated stable object (update spectrum untested); we suppress the top-$|\lambda|$ eigenspace without exhaustively characterizing curvature sign. **Future work** (preregistered): the function-space operator at larger scale.

USE OF AI

Generative AI tools were used for writing/editing and code support. All claims, proofs, results, and conclusions were developed and verified by the authors, who take full responsibility.

REPRODUCIBILITY STATEMENT

All numerical claims are reproducible from committed artifacts in the anonymous supplement (code, configurations, all seeds, raw/processed results, figure scripts, an evidence ledger mapping each claim to its artifact, and mechanical validation tests including an exact-eigendecomposition cross-check). Appendices detail theory assumptions, the full rank 2–50 audit, the calibration grid, statistics, and compute.

REFERENCES

Emmanuel Abbe, Enric Boix-Adsera, and Theodor Misiakiewicz. The merged-staircase property: A necessary and nearly sufficient condition for sgd learning of sparse functions on two-layer neural networks. *arXiv preprint arXiv:2202.08658*, 2022.

Gérard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *Journal of Machine Learning Research*, 22, 2021.

Alex Damian, Jason D. Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. *arXiv preprint arXiv:2206.15144*, 2022.

Chandler Davis and William M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via hessian eigenvalue density. *arXiv preprint arXiv:1901.10159*, 2019.

Guy Gur-Ari, Daniel A. Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace. *arXiv preprint arXiv:1812.04754*, 2018.

Jaerin Lee, Bong Gyun Kang, Kihoon Kim, and Kyoung Mu Lee. Grokfast: Accelerated grokking by amplifying slow gradients. *arXiv preprint arXiv:2405.20233*, 2024.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. *arXiv preprint arXiv:1804.08838*, 2018.

Ziming Liu, Ouail Kitouni, Niklas Nolte, Eric J. Michaud, Max Tegmark, and Mike Williams. Towards understanding grokking: An effective theory of representation learning. *arXiv preprint arXiv:2205.10343*, 2022a.

Ziming Liu, Eric J. Michaud, and Max Tegmark. Omnigrok: Grokking beyond algorithmic data. *arXiv preprint arXiv:2210.01117*, 2022b.

Neil Mallinar, Daniel Beaglehole, Libin Zhu, Adityanarayanan Radhakrishnan, Parthe Pandit, and Mikhail Belkin. Emergence in non-neural models: grokking modular arithmetic via average gradient outer product. *arXiv preprint arXiv:2407.20199*, 2024.

Charles H. Martin and Michael W. Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *arXiv preprint arXiv:1810.01075*, 2018.

Andrea Montanari and Zihao Wang. Phase transitions for feature learning in neural networks. *arXiv preprint arXiv:2602.01434*, 2026.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations (ICLR)*, 2023.

Vardan Papyan. The full spectrum of deepnet hessians at scale: Dynamics with sgd training and sample size. *arXiv preprint arXiv:1901.08244*, 2019.

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.

Adityanarayanan Radhakrishnan, Daniel Beaglehole, Parthe Pandit, and Mikhail Belkin. Mechanism for feature learning in neural networks and backpropagation-free machine learning models. *Science*, 383(6690):1461–1467, 2024.

Vikrant Varma, Rohin Shah, Zachary Kenton, János Kramár, and Ramana Kumar. Explaining grokking through circuit efficiency. *arXiv preprint arXiv:2309.02390*, 2023.

Y. Xu. Feature repulsion and spectral lock-in: An empirical study of two-layer network grokking. *arXiv preprint arXiv:2605.08119*, 2026.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. Galore: Memory-efficient llm training by gradient low-rank projection. *arXiv preprint arXiv:2403.03507*, 2024.

A Appendix: protocol, proofs, and full results

The complete frozen protocol and amendment ledger, theorem assumptions and proofs, implementation-validation tests, the full matched-control definitions, statistical-analysis plan, seed-level tables, Lanczos convergence and eigenpair residuals, the complete rank 2–50 identification audit, the detector calibration grid, per-seed cross-seed results, compute/hardware details, and the claims/evidence ledgers are provided in the supplement. All main-text claims are supported within the nine main pages; appendices provide detail only.

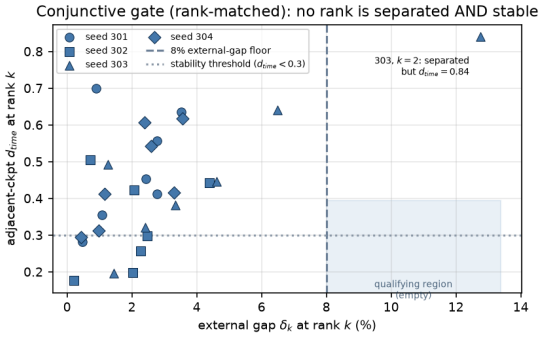

Figure 7: Conjunctive identification across seeds (supporting Table 1). Each point is a $(\delta_k, d_{\text{time}}(k))$ pair evaluated *jointly at the same rank* k for a given seed (coordinates are not decoupled across ranks); the qualifying region (gap $\geq$ floor *and* $d_{\text{time}} < 0.3$) is empty. Seed 303 clears the gap floor only at k=2, where it is temporally unstable (d_{time}=0.84), so no single rank is separated and stable.

Preregistration and amendment ledger. The identification gates (G1–G4), estimands, and stop-rules were frozen before confirmatory outcomes; the deep-net extension protocol was frozen separately before its outcomes. A sequence of post-review amendments, each content-hashed and committed *before* its results were observed, authorized one or two further runs apiece: a true Gauss–Newton operator sweep and a deep-net causal contrast at the preregistered n=16; a curvature-sign and dose-response battery and an energy-matched control; a frozen quantitative specificity gate ($R \geq 2$); the function-space operator hunt (with its disconfirming verdict tokens pre-committed); a second-positive hunt on a rotated (non-axis-aligned) feature (Section A); and a symmetry-adapted per-frequency operator on the transformer (Section A). Each prohibited threshold re-tuning, outcome-selected checkpoints, and seed backfilling. The supplement contains each protocol file, its freeze timestamp and hash, and the verdict tokens, so confirmatory and exploratory analyses

are separable by construction: the teacher–student positive and the transformer identification audit are confirmatory against frozen gates; the deep-net breadth settings and operator comparisons are exploratory extensions reported under the amendment.

Rotated-feature identification (amendment `03ed54c7`). To test whether the function-space positive is an artifact of axis alignment, we rotate the k=3 parity feature by a fixed random orthogonal Q so the informative subspace is the (non-axis-aligned) span of the first three rows of Q; the specificity basis B is that rotated subspace. Under the identical G1–G4 detector the top function-space mode passes the full gate on $7/16$ seeds with mean R=3.2 (a random subspace scores $R\approx1$), so the operator discovers a rotated feature it cannot read off input coordinates. The causal arm is inconclusive: the detector window occurs only at a training-set size where the base run does not grok within the 8000-step horizon, so all arms censor at the horizon and the delay contrast is 0 (rank p=1.0). Precommitted token: `IDENTIFIED_NOT_MANIPULABLE`. Evidential interpretation after observing failure of the baseline event: `INCONCLUSIVE_DUE_TO_NO_BASELINE_TRANSITION` (the baseline never groks in this window, so the run neither supports nor refutes manipulability; contrast the parity MLP of Section 8, where grokking occurs and targeted suppression is no more disruptive than controls). A follow-up (amendment `f54173aa`) re-selected the config by a *grokking-window* rule (base must generalize within a raised 12000-step horizon and leave a pre-generalization plateau) over a $\{800, 1000, 1200\} \times \{0.5, 1.0\}$ (ntr, wd) grid; *no* config groks within budget ($0/8$ base seeds reach test accuracy 0.9, though the detector still fires with $R \approx 2.2$–2.4). We do not conclude the rotated task is un-grokkable: no tested rotated-feature configuration produced an observed grokking transition within budget, so a causal contrast could not be formed. We report this honestly; the deep-net *causal* positive remains the axis-aligned parity MLP, and a second deep-net causal positive is left to future work.

Symmetry-adapted transformer operator (amendment `ee2e3fb`). We ran the operator a $\mathbb{Z}_p$ task most invites: per-frequency Fourier-ring blocks (the basis of the known modular-addition circuit). For each frequency we score its rank-2 block with the same G1–G4 conjunction. No block forms a compact object on any of the four seeds: each captures $\approx 4\%$ of the function-space sensitivity ($R \approx 1.05$, mean 1.05, max 1.07, indistinguishable from a random rank-2 subspace), with relative separation $\ll 0.08$ and unstable across bootstrap/adjacent checkpoints ($0/4$ pass). Multi- frequency blocks (rank 4, 6) do no better ($R \approx 1.05$). Verdict token: `ADAPTED_OPERATOR_NO_COMPACT_OBJECT`. The representation is not concentrated in any tested rank-2, rank-4, or rank-6 Fourier block; larger joint Fourier subspaces, rotating subspaces, activation-space and update-spectrum structure remain untested, so this is scoped to the tested low-rank Fourier-adapted operators.

Specificity-ratio extraction map T (per setting). Teacher–student / single-index MLP: $T = I$, $D_T = d$, $B = w^*$ (b=1). Parity MLP: $T = I$ on the D=40 input coordinates, B the k=3 informative bits. Transformer: T slices the token-embedding block (D_T = embedding dimension), B the closed-form Fourier ring basis; $\widetilde{U} = \mathrm{orth}(TU)$ is renormalized to unit Frobenius norm before projection, and the per-mode top-three-frequency selection is an empirical upper bound on Fourier concentration.

Mechanistic circuit identity (amendment `f0279bd`). Confirmatory run on the 16 parity seeds (401–416; all grok). At the post-grok checkpoint we rank layer-1 hidden units by weight energy on the informative bits and take the smallest set capturing 80% of that energy (median k=12.5 of 256; median participation ratio ≈ 10; $14/16$ seeds sparse). *Necessity*: ablating this set drops test accuracy to chance (median 0.51) on $16/16$ seeds. *Specificity*: ablating an equal number of random units preserves accuracy (median 0.92) on $13/16$. So the learned parity solution contains a sparse, necessary motif. Precommitted token: `MECHANISM_IDENTIFIED`. Evidential interpretation: `NECESSARY_SPARSE_MOTIF_IDENTIFIED`; `MEDIATION_OF_PULLBACK_EFFECT_NOT_ESTABLISHED` – the present evidence does not establish that the function-space intervention is selectively mediated by that motif. *Honest scope*: the parameter pullback of u_{fs} correlates with the motif (median energy-correlation 0.68), but a random direction's pullback correlates comparably (0.72), so we do *not* claim the pullback targets the motif more specifically than a generic direction; on a 256-wide MLP the parity units dominate the W_1 update geometry.

On the 8% separation floor. This threshold is *detector-relative*, not a universal causal boundary: it is the empirically measured sensitivity of our detector on planted objects ($\approx 1.08\times$ the bulk edge

with zero false positives), and the transformer negative is robust to sweeping it over $\{5\%, 8\%, 12\%\}$. It bounds when a separated object is recoverable by this assay; it makes no claim about causal relevance below the floor. It bounds detector *sensitivity*, not a formal type-I error rate.

On multiplicity. The confirmatory inferences are pre-specified single contrasts (targeted-vs-random; the $p{=}1/p{=}2$ interaction; targeted-vs-energy-matched), each at $p < 10^{-5}$, orders of magnitude below any Bonferroni correction over the $\sim$15–20 tests in the arm/setting family; the placebo arms are declared as TOST equivalence tests, not screened for significance. The function-space positive is tested against three matched controls (random, energy-matched, wrong-bits), each at $p = 5 \times 10^{-5}$ ($16/16$ seeds), which survives Bonferroni correction over the three contrasts ($3 \times 5 \times 10^{-5} < 10^{-3}$). The transformer identification sweep (ranks 2–50$\times$ seeds $\times$ thresholds) is a *negative* under a conjunction, where multiplicity only makes a spurious pass more likely, so failing anyway is conservative.

Scope diagram. The core identification protocol targets compact, static, parameter-space spectral objects under the full-Hessian and Gauss–Newton operators. Mechanism classes it does *not* adjudicate in parameter space: distributed (non-compact) directions, trajectory-tracked rotating subspaces, and symmetry-adapted (e.g. per-frequency) bases. The function-space (input-sensitivity) operator is a demonstrated *extension* of the protocol (Section 9) that recovers a compact object when the feature is low-dimensional; activation-space and update-spectrum operators remain future work. A parameter-space negative is a statement about the covered class, not about these other operators.