MASK FAITHFULNESS DOES NOT GUARANTEE SELECTIVE CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

A parameter decomposition writes a weight matrix as a sum of rank-one components with learned gates. Its standard test is mask faithfulness. The network is expected to behave the same under every mask allowed by the gates. We explore a different question. Can a faithful decomposition erase one behavior while preserving others? We answer this exactly on a controlled task with d independent, always-active behaviors. For every $d \geq 2$ and any capacity $C \geq d$, we show that every faithful decomposition with the fewest active gates pairs coordinates. This is because a paired read cancels on half the inputs, so it appears sparser than a single coordinate. Erasing one behavior then raises another behavior's error by at least $1/8$ (MSE), for any real coefficients whereas the coordinate basis erases with no damage. The sparsity advantage that pairing creates depends on the input data. An input that activates only one behavior activates both the sum and difference reads in a pair but only one coordinate read. If single-behavior inputs become frequent enough, the coordinate basis is the optimal choice. A second theorem extends this bound to smooth targets, small input noise, continuous gates, and a finite penalty. It holds for every 5%-approximate optimum. In three small trained networks, the same change of training data cuts the provable cost of selective edits by more than $400\times$. Therefore, selective control should be judged separately from masked reconstruction.

1 INTRODUCTION

Parameter decomposition (PD) writes a weight matrix as a sum of rank-one components with learned, input-dependent causal-importance gates (Braun et al., 2025; Bushnaq et al., 2025; 2026). Its central promise is mask faithfulness: under the masks permitted by the gates, a masked network should behave like the unmasked one. We ask if a faithful decomposition enables us to erase one behavior while preserving another solely through editing component coefficients? Faithful masked reconstruction does not guarantee this independent control, a sparse faithful representation can merge independently implemented behaviors into shared components, resulting in downstream effects when one is erased.

We characterize when this obstruction arises and how it disappears. On a controlled identity-weight task with independent GELU and SiLU behaviors, every globally minimal exact faithful decomposition pairs read coordinates, for every $d \geq 2$ and $C \geq d$, so exact erasure of one behavior erases a partner (Theorem 1). A data intervention that preserves the target, mean, and covariance selects native coordinate reads and restores control (Proposition 1). A separate theorem bounds average-over-target collateral in a small-noise finite-penalty regime (Theorem 2, Appendix C). For saved dictionaries we prove bounds from their matrices alone (Section 5); approximate global optima here refer to the constructed population objective, whereas those bounds are for saved dictionaries. We train three small two-factor GELU MLPs ($2 \to 32 \to 32 \to 2$, shared hidden layers, $C = 2$ decompositions of the first weight matrices) and fit their dictionaries, obtaining feasible edits at matched budget with a provable control-cost improvement above $400\times$ under the fixed-moment intervention on finite held-out inputs (Section 6, Fig. 4).

Scope. In the controlled task every behavior is active on every input, because x_j is always $+1$ or -1. This density drives the obstruction since the native coordinate basis has expected gate count d,

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.

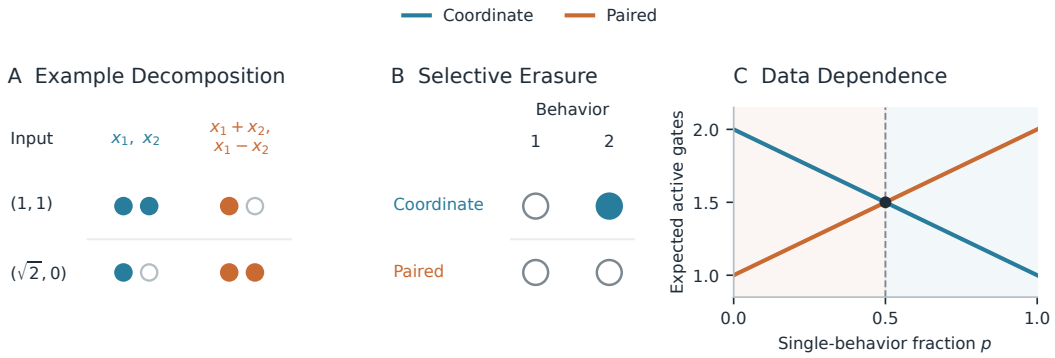

Figure 1: **Example Decomposition.** Analytic $d = 2$ illustration; both bases reconstruct the unedited target exactly. **(A)** Filled circles indicate active gates. On the displayed sign input, paired reads need one active gate while coordinate reads need two; a single-behavior input reverses the count. **(B)** Filled circles indicate preserved behaviors and hollow circles erased behaviors. In the orthogonal sum/difference decomposition, any coefficient edit that perfectly erases behavior 1 also erases behavior 2. The coordinate basis however, preserves behavior 2. **(C)** Under the mixture in Proposition 1, expected gate counts are $2 - p$ for coordinates and $1 + p$ for pairs. The preferred basis changes at $p = 1/2$, with target, mean and covariance fixed.

whereas a paired read $e_j + s\, e_k$ cancels on half the inputs and costs $1/2$ (Remark 1). The obstruction is therefore an indicator about densely co-active behaviors. Faithfulness alone does not force any basis, since any dictionary that has all gates equal to one is faithful; what selects the mixed basis is sparsity combined with faithfulness.

Low masking loss and sparse, faithful dictionaries don't necessarily permit independently erasable behaviors.

2 Setup: Independent Behaviors, Gates, and Control Costs

Task. Let X have i.i.d. entries uniform on $\{-1, +1\}$, and let each behavior read its own coordinate,

$$f_j(x) = \phi_j(x_j), \qquad j = 1, \dots, d, \qquad d \geq 2.$$

The target weight is I_d. Each head exposes binary logits $[0, \kappa\, \phi_j(z_j)]$, where z_j is the j-th preactivation ($z_j = x_j$ for the unedited target), with fixed gain $\kappa = 8$, and ϕ_j is GELU or SiLU, possibly differing across heads. Both satisfy

$$\phi_j(t) - \phi_j(-t) = t, \qquad \phi_j(0) = 0, \tag{1}$$

as does the identity. The support $\{\pm 1\}^d$ spans $\mathbb{R}^d$.

Metrics. An *erasure* edit is scored by the target behavior's loss, and *preservation collateral* by another behavior's loss under the same edit. We use output MSE, logit MSE, or binary cross-entropy (CE), in the unit selected in the experiment, and never combine across units. For an edit with $W' = I_d - E$, $E = \sum_i w_i A_i$, and row W'_j of W',

$$R_j = \mathbb{E}\big[\phi_j(W'_j X)^2\big], \qquad C_j = \sum_{k \neq j} \mathbb{E}\big[\big(\phi_k(W'_k X) - \phi_k(X_k)\big)^2\big]. \tag{2}$$

Here R_j is the output MSE distance to the erased (zero) head and C_j the preservation MSE to the clean output. The logit-erasure budget is $\kappa^2 R_j$; binary CE is the ordinary logistic loss for $Y_k = X_k$.

Atoms, gates, and all-mask faithfulness. A decomposition writes $I_d = \sum_{i=1}^{C} A_i$ with atoms of rank at most one, $A_i = u_i v_i^\top$, and zero slots allowed; an atom with $u_i = 0$ contributes nothing

and is omitted. Each atom carries a learned gate $g_i(x) \in [0,1]$, and a mask may take any value $m_i(x) \in [g_i(x), 1]$. Faithfulness is the nonlinear output identity

$$\phi_j\Big(\big(\textstyle\sum_{i=1}^{C} m_i(x) A_i x\big)_j\Big) = \phi_j(x_j) \qquad \text{for all } x \in \{\pm 1\}^d \text{ and all allowed } m_i. \tag{3}$$

We minimize the expected number of positive gates,

$$\min \;\; \mathbb{E}_x \#\{i : g_i(x) > 0\} \qquad \text{subject to (3).} \tag{4}$$

Equation (3) is the strong reading of mask faithfulness, requiring it for every allowed mask. Objective (4) is the $\gamma \to 0$ limit of the minimality loss (gate power γ) of APD, SPD and VPD (Braun et al., 2025; Bushnaq et al., 2025; 2026): it counts how many components an input needs, and it does not constrain how many input coordinates a component reads. We assume $\sum_i A_i = I_d$, so the all-on map is exactly the target.

Edits. A coefficient edit forms

$$W' = I_d - \sum_{i=1}^{C} w_i A_i. \tag{5}$$

It subtracts learned components from I_d, with $w \in \mathbb{R}^C$ in the exact theorem; the finite empirical bounds restrict to the box $[-1,1]^C$. For an inexact dictionary with residual $\Delta = I_d - \sum_i A_i$, a complete erasure $w_i = 1$ preserves Δ automatically. The native basis $A_j = e_j e_j^\top$ erases behavior j with $w = e_j$ and zero collateral; the paired optima of Theorem 1 can be orthogonal, so the obstruction is not ill-conditioning.

3 Every optimal faithful decomposition mixes behaviors

Theorem 1 (All optima and their erasure cost)**.** *Let $d \geq 2$, $C \geq d$, let X be uniform on $\{\pm 1\}^d$, let the target be I_d, and let each ϕ_j be differentiable with $\phi_j'(+1) \neq 0$ and $\phi_j'(-1) \neq 0$. Under all-mask faithfulness (3):*

(i) the minimum of (4) is $d/2$;

(ii) every optimum has exactly d nonzero atoms, and the read v_i of every nonzero atom is proportional to $e_j + s\, e_k$ with $k \neq j$ and $s \in \{\pm 1\}$;

(iii) if in addition every ϕ_j satisfies (1), then at every optimum, for every coefficient edit $W' = I - \sum_i w_i A_i$ with $w \in \mathbb{R}^C$ arbitrary and every behavior j whose output MSE is at most ϵ, some $k \neq j$ has preservation MSE at least $\frac{1}{4}\big[\frac{1}{\sqrt{2}} - 2\sqrt{\epsilon}\big]_+^2$; at perfect erasure ($\epsilon = 0$), some $k \neq j$ is itself perfectly erased: $W'_k = 0$, so its preservation MSE equals $\mathbb{E}[\phi_k(X_k)^2]$.

The bound in (iii) is independent of d, C and the coefficient magnitude.

The proof of Theorem 1 appears in Appendix A. GELU and SiLU satisfy both hypotheses. ReLU satisfies the odd identity (1) but has $\phi'(-1) = 0$, so it fails only the derivative hypothesis used in Step 1 of the proof of Theorem 1 (Appendix A) and is therefore not covered.

3.1 Finite prediction loss

Theorem 1(iii) bounds an abstract MSE. On this task it converts into a prediction-loss floor. Take binary labels $Y_j = X_j$, logits $\ell_j = \kappa \phi_j(x_j)$, and independent uniform signs. Both GELU and SiLU are globally L-Lipschitz with $L = 1.13$, and the clean per-head binary CE is at most $b = 0.124435$ nats (GELU is the larger).

Corollary 1 (Cross-entropy floor at an optimum)**.** *At every optimum of Theorem 1, if a coefficient edit makes the target head's mean squared logit at most ϵ_{logit}, then some other head has binary cross-entropy at least*

$$\frac{\log 2}{2} - L\sqrt{\epsilon_{\text{logit}}/2}. \tag{6}$$

At $\epsilon_{\text{logit}} = 0$ the partner head is fully erased by Theorem 1(iii), so its cross-entropy is exactly $\log 2$; the bound (6) keeps only the half of the inputs on the event H of the proof. At $\epsilon_{\text{logit}} = 0.01$ this is at least 0.266670 nats, an increase relative to the clean loss, bounded conservatively using b, of at least $0.266670 - b \geq 0.142235$ nats.

Corollary 1 is the case $\delta = 0$ of the prediction-loss bound (10), derived in Appendix E; the gain κ cancels and is set to 8 only so that the clean loss is small enough for the increase to be visible.

3.2 A DATA INTERVENTION RESTORES CONTROL

The obstruction used corner data. Change only the input distribution, holding the target, mean and covariance fixed. For $p \in [0, 1]$ let

$$D_p = (1 - p)\,\mathrm{Unif}\{\pm 1\}^d \;+\; p\,\mathrm{Unif}\{\pm\sqrt{d}\,e_j\}_{j=1}^{d}, \tag{7}$$

the mixture of uniform cube corners and uniform single-coordinate axis probes, scaled so both have the same norm. A corner activates all d behaviors, while an axis probe activates exactly one, so D_p interpolates between always-co-active data ($p = 0$) and one-behavior-at-a-time data ($p = 1$). Every D_p has mean zero and covariance I_d, and all support coordinates have nonzero activation derivative: the support coordinates are 0, ± 1 and $\pm\sqrt{d}$, $\phi'(0) = \frac{1}{2}$ for GELU and SiLU, $\phi'(\pm 1) \neq 0$, and the only zeros of GELU' and SiLU' lie at $t \approx -0.75$ and $t \approx -1.28$ respectively, which are not of the form $-\sqrt{d}$ for any integer $d \geq 2$.

Proposition 1 (Exact mixture threshold)**.** *Under D_p, every global optimum of (4) uses paired reads for $p < d/(d+2)$ and native coordinate reads for $p > d/(d+2)$; at $p = d/(d+2)$ mixed invertible native/paired bases tie. In particular the thresholds are $1/2, 2/3, 4/5$ for $d = 2, 4, 8$.*

The proof (Appendix B) compares activation densities: a k-coordinate read is active on axis probes with probability k/d and on corners with probability at least $1/2$, with equality only for paired reads.

Below p^* the erasure bound of Theorem 1(iii) still holds; above it the native basis erases with zero collateral. The threshold $p^* = d/(d + 2)$ tends to 1 as d grows, so an increasing fraction of the data must be single-behavior probes before the native basis becomes optimal. The threshold is thus a property of the representation and the input distribution together. The optimum set and the threshold are unchanged for any positive gate power and for the increasing population-frequency penalty of Appendix B.

4 FINITE PENALTIES AND NOISY INPUTS

Theorem 1 concerns exact optima of an idealized objective. Appendix C extends it to the finite-penalty objective $J = \mathbb{E}\sum_i g_i(X)^\gamma + \lambda\mathbb{E}e(X)^2$ with Gaussian-perturbed inputs $X = S + \sigma Z$, bounded coefficient edits, and any dictionary whose objective is within a factor $1+\xi$ of the Hadamard competitor. Theorem 2 there bounds the average-over-targets collateral by a rank–trace argument, and Corollary 2 instantiates it at $d = 16$, $\sigma = 10^{-3}$, $\xi = 0.05$ to give collateral at least 0.07448 while a coordinate dictionary erases with zero collateral. The bound is informative only for σ below about 3×10^{-3}, so the experiments of Sections 5 and 6 rely on the finite bounds of Section 5, not on it.

5 PROVABLE BOUNDS FOR TRAINED DICTIONARIES

Theorem 1 concerns optimal solutions of an idealized problem. Real dictionaries are trained with finite penalties and need not be optimal. This section therefore gives bounds that hold for any saved dictionary: from its matrices alone, a lower bound on the collateral of every coefficient edit in a declared box that erases a given behavior. No search over edits is involved, so a small bound cannot be blamed on a weak optimizer. We then train dictionaries with the reference implementation and evaluate the bounds on them.

MSE bound. Let $\{A_i\}_{i=1}^{C}$ be a saved dictionary, exact or not, and use the edit (5) acting on the original target, so the residual $\Delta = I_d - \sum_i A_i$ is preserved. Suppose behavior j is erased to output

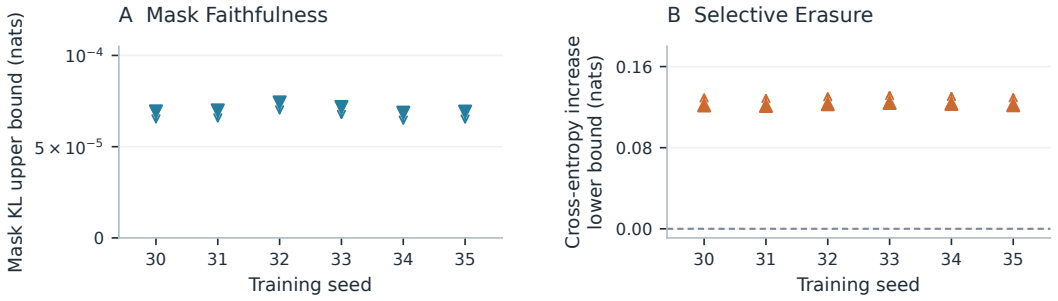

Figure 2: **Mask Faithfulness and Selective Erasure.** Six unselected training seeds on the controlled GELU/SiLU task. **(A)** Original learned-gate KL upper bounds over all allowed input-dependent masks, averaged over inputs and heads; all six are below 7.5×10^{-5} nats. **(B)** Cross-entropy increase lower bounds relative to the clean baseline, minimum across either erased head, at target-logit MSE .01, gain 8, and coefficient box $[-1, 1]^2$; every lower bound exceeds .12 nats. The dashed line marks zero collateral for the coordinate basis. Arrows point toward the guaranteed side of each bound; they are not error bars.

MSE at most ϵ. With $E(w) = \sum_i w_i A_i$ and row j denoted $E_j(w) = e_j^\top E(w)$, the erased row obeys $\|E_j(w) - e_j^\top\|^2 \le 4\epsilon$, and every preserved behavior contributes MSE at least $\frac{1}{4}\|E_k(w)\|^2$ by the odd-part fact (11) (with $q = 1$). Hence the nonlinear collateral is at least

$$\frac{1}{4} \min_w \sum_{k \neq j} \|E_k(w)\|^2 \quad \text{s.t.} \quad \|E_j(w) - e_j^\top\|^2 \le 4\epsilon, \quad |w_i| \le B. \tag{8}$$

The displayed expression supplies a global lower bound on the nonlinear collateral; the unscaled minimum is a relaxation, not the nonlinear optimum. Any $\lambda \ge 0$ and any $w_0 \in [-B, B]^C$ give a valid global bound on the primal value via the tangent of $q_\lambda(w) = \sum_{k \neq j} \|E_k(w)\|^2 + \lambda(\|E_j(w) - e_j^\top\|^2 - 4\epsilon)$; the explicit minimizing corner is in Appendix E. The empirical box is $[-1, 1]^C$. This applies to actual saved matrices and makes no optimality claim about the learned dictionary.

Prediction-loss bound. Let $r = e_j + s\, e_k$ (a column) and let $F_i = r^\top A_i - [(r^\top A_i r)/(r^\top r)]\, r^\top$ be the residual row of $r^\top A_i$ orthogonal to r, and let $H = \{r^\top X = 0\}$. For $\|w\|_\infty \le B$,

$$\left\|\sum_i w_i F_i\right\| \le B\sqrt{C}\, \left\|\mathrm{stack}_i F_i\right\|_F =: \delta. \tag{9}$$

Here δ is the *pairing defect* of the dictionary with respect to r; it vanishes when every $r^\top A_i$ is a multiple of $r^\top$, as at every optimum of Theorem 1, and it is not the dictionary residual Δ. On the event $H = \{r^\top X = 0\}$, which has probability $1/2$, the row evaluations satisfy $(W'_j + sW'_k)X = -(\sum_i w_i F_i)X$. Together with the target odd-part budget, the conditional covariance of X on H (operator norm 2), the Lipschitz bound $L = 1.13$, and the global tangent $\mathrm{softplus}(-Y\ell) \ge \log 2 - \frac{1}{2}|\ell|$, this gives

$$\mathrm{CE}_k \;\ge\; \frac{\log 2}{2} - L\sqrt{\frac{\epsilon_{\mathrm{logit}}}{2}} - \frac{\kappa L\, \delta}{2\sqrt{2}}, \tag{10}$$

where $\epsilon_{\mathrm{logit}}$ is the target logit-erasure MSE. For a saved dictionary the bound is evaluated for every partner $k \neq j$ and sign $s \in \{\pm 1\}$, and the largest value is reported. At an exact optimum $\delta = 0$ and this reduces to Corollary 1. The derivation is in Appendix E.

5.1 Native KL study with distinct nonlinearities

We train on the controlled task with the native implementation as a controlled native reference (not the unchanged complete released training setup) with $C = d = 2$, two independent uniform signs, GELU on the first head and SiLU on the second, binary logits $[0, 8\phi_j(z_j)]$, and the native KL reconstruction loss with native PGD. Six fresh starts (seeds 30–35) are reported without selection (Fig. 2). The pre-specified pass criterion required at least five of six models to satisfy both a shared-mask KL at most .002 and a provable CE increase at least .05 nats for both target erasures; all

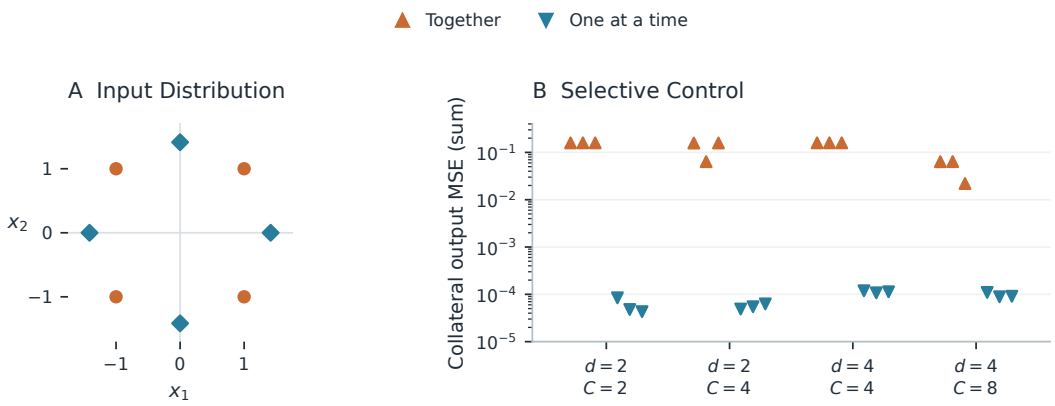

Figure 3: **Data Intervention.** **(A)** Clean $d = 2$ supports for behaviors active together ($p = 0$) or one at a time ($p = 1$). Both laws have covariance I; training adds $\sigma = .1$ noise and standardizes, with the target fixed. **(B)** Four dimension/capacity configurations (d, C), at output-erasure MSE .01, coefficient box $[-1, 1]^C$, and a common evaluation law. Up-triangles are minimum collateral lower bounds across erased heads for corner training; down-triangles are maximum feasible upper bounds across heads for axis training. Each point is one selected three-start replicate, not an error bar; selection uses only the validation objective and all 72 starts are retained.

six pass. A post-training analytic audit bounds the *original* learned gates over all input-dependent permitted masks (Appendix E). The learned gates have exactly one positive gate per input (strict gate $L_0 = 1$); the positive value need not equal 1, and the editing coefficients remain arbitrary real values in $[-1, 1]^2$. The audited guarantee is

$$\mathbb{E}_x \sup_{m \in [g(x),1]} \mathrm{KL}_{\mathrm{mean_heads}}(x, m) \;\le\; 7.43 \times 10^{-5}$$

for every seed, averaged over the four inputs. The bounds are nonvacuous: the coefficient vector $(1, 1)$ erases either target to logit MSE below 9.5×10^{-7} in all six seeds. The minimum provable CE increase is 0.121228 nats.

5.2 Noise, Capacity, and the Data Intervention

We add independent Gaussian noise of standard deviation $\sigma = 0.1$ to the training law and standardize, so each population has mean zero and covariance I_d; the target $\mathrm{GELU}(x)$ is unchanged. We run the four configurations $(d, C) \in \{(2, 2), (2, 4), (4, 4), (4, 8)\}$ at the extremes $p = 0$ (corners) and $p = 1$ (axis probes), three replicate groups of three starts each—72 starts, all retained—and select within each replicate by the validation objective only. The primary endpoint is output-erasure MSE at most .01 with coefficient box $[-1, 1]^C$; the secondary endpoint is the paired-read CE bound (10). The primary output-MSE bounds are universal over centrally symmetric evaluation laws with covariance I, whereas the secondary CE bound is not.

Outcomes. All 12 matched pairs qualify (the pre-specified criterion required at least 10 of 12, with at least two of three replicates per configuration); the per-configuration dense, probe, mask, and CE numbers are in Table 1. Every matched head's provable gap in optimal cost exceeds 300; because the lower bound is universal and the upper bound feasible, this bounds a ratio of optimal costs, not of found local optima. Every one of the 36 dense erasures is feasible in the population via $w = \mathrm{ones}$, with maximum erasure upper bound 3.781×10^{-4} and axis erasure upper bound 1.297×10^{-4}. The masking audit is a finite grid (a 31×31 grid for $C = 2$, 5^C for $C = 4$, fixed random shared masks plus endpoints for $C = 8$), not a continuous supremum. The secondary CE bound is positive on 24 of 36 targets. At $(d, C) = (4, 8)$ the dense lower bound and the audited masking MSE are the same order of magnitude, and the CE bound is inconclusive on all 12 targets there because its pairing defect δ scales with $\sqrt{C}$. The bounds thus weaken with spare capacity $C - d$ and do not show how the obstruction behaves at large capacity in trained dictionaries, even though Theorem 1 covers every $C \ge d$ at exact optima. An independent reconstruction of all 72 raw

Table 1: Native 72-start MSE confirmation by configuration. "Dense" is the minimum dense-trained collateral lower bound across all erased heads and the three replicates; "Probe" the maximum axis-trained population collateral upper bound across all heads and replicates; "Mask" the maximum finite-grid audited MSE; "CE" the positive secondary cross-entropy targets. Dense and probe units are summed output MSE; mask is the mean audited MSE over heads.

(d, C)	Pairs	Dense $\geq$	Probe $\leq$	Mask $\leq$	CE
$(2, 2)$	$3/3$	0.1591	8.44×10^{-5}	0.00825	$6/6$
$(2, 4)$	$3/3$	0.0637	6.27×10^{-5}	0.00672	$6/6$
$(4, 4)$	$3/3$	0.1593	1.19×10^{-4}	0.00395	$12/12$
$(4, 8)$	$3/3$	0.02198	1.10×10^{-4}	0.01307	$0/12$

U/V checkpoints verified the stored witnesses in exact rational arithmetic (maximum atom replay difference 5.89×10^{-8}; all 36 lower and 36 upper bounds pass).

6 Learned Shared Networks

We extend the evidence to trained networks. The target is a dense GELU MLP with architecture $2 \to 32 \to 32 \to 2$ and shared hidden layers. It is trained on two independent noisy binary-label tasks defined by rotated two-latent-sign targets. The first weight matrix is decomposed with $C = 2$. We use one pilot network and two preregistered confirmation networks. Architecture, gate training, noise, selection, and full details are in Appendix G, and these small two-factor networks show the mechanism in a trained network, not at scale.

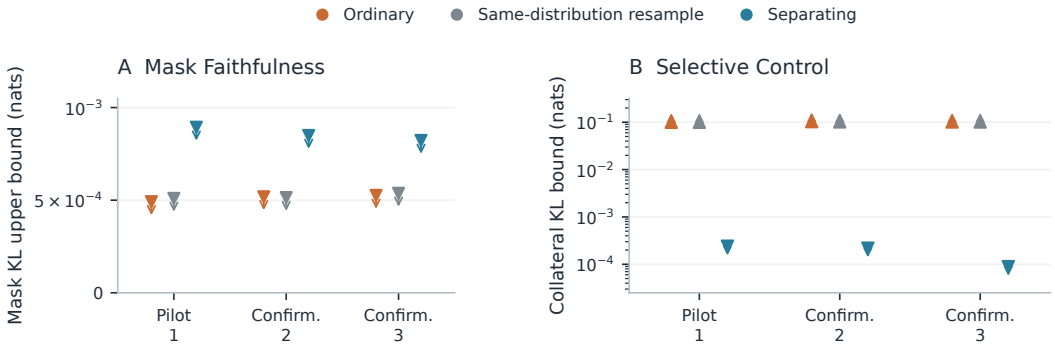

Figure 4: **Learned Networks.** One pilot and two preregistered confirmations; 2048 held-out inputs per network and coefficient box $[-1, 1]^2$. **(A)** For each data condition, the mean over inputs and heads of an upper bound on KL over all allowed masks. **(B)** Minimum over heads of ordinary and same-distribution-resample collateral lower bounds (up-triangles), and maximum over heads of separating feasible upper bounds (down-triangles), at the common erasure budget $\mathrm{ER}_j \leq 0.065$. Separating data reduce optimal control cost by more than $400\times$ in every network relative to either control condition. This is a certified ratio of optimal costs, not point estimates. Arrows show bound direction, not uncertainty.

We compare three data arms. In the *ordinary* arm, both factors are active together or both are inactive together. In the *separating* arm, exactly one factor is active and the two marginals are balanced. The *resample* arm is an independent fresh sample from the ordinary law and serves as a same-distribution control. All three arms have the same population mean 0 and covariance I. The target network is identical across arms; only the data used to fit the decomposition differs.

The fit has two stages: we train the components, freeze them, then re-initialize and train the gate networks identically in every arm on the same common data. Differences between arms therefore come from the components, not from gate training.

Let ℓ_j and ℓ'_j denote clean and edited logits, and define the erasure ratio on target-active inputs,

$$\mathrm{ER}_j \;=\; \frac{\mathrm{mean}_{\mathrm{active}}(\ell_j'^{\,2})}{\mathrm{mean}_{\mathrm{active}}(\ell_j^2)} \;\le\; 0.065.$$

Collateral cost is the clean-to-edited Bernoulli KL on active other-head samples, with coefficient box $[-1,1]^2$. Each selected dictionary satisfies a mean over 2048 inputs and heads of the per-input supremum of the KL over all allowed masks below 9×10^{-4}. Every feasible ordinary or resample edit costs more than 0.10 nats, whereas separating admits feasible edits below 2.5×10^{-4}, so the ratio of optimal ordinary or resample cost to optimal separating cost exceeds 400 per matched head, target, and control. Each lower bound holds for every coefficient vector in the box; it is proved by interval arithmetic over a complete cover of the box (details in Appendix G). Results concern 2048 finite held-out inputs per network.

7 Related work

Parameter decomposition. APD (Braun et al., 2025), SPD (Bushnaq et al., 2025) and VPD (Bushnaq et al., 2026) factor a weight matrix into rank-one components with learned gates, and the Goodfire handbook (Goodfire Research, 2026) already recommends measuring both target change and off-target divergence. We do not claim off-target damage or its measurement as new. Our contribution is an all-optima, arbitrary-capacity characterization of when one behavior cannot be independently erased. Faithfulness alone is trivially satisfiable by any dictionary with all gates equal to one; the faithfulness-plus-minimality objective is not merely blind to this obstruction, since on this task it selects the mixed basis at every optimum.

Composition, absorption and co-activity. Feature composition, data-dependent absorption, and orthogonality blindness are established (Bussmann et al., 2025; Korznikov et al., 2025; Vigouroux & Sharkey, 2026; Chanin et al., 2024); we claim no priority there. Our distinct result is that a whole-map criterion can be blind in the strong sense that *all* optima are inseparable: in the constructed task, every optimum faithful decomposition prevents independent behavioral erasure, and a causal data change removes that obstruction.

Erasure, editing and identifiability. Concept-erasure methods (INLP, LEACE) remove a linearly encoded concept from representations while minimizing damage, and model-editing methods (ROME) measure collateral change to unrelated behavior (Ravfogel et al., 2020; Belrose et al., 2023; Meng et al., 2022). Our erasure-with-preservation metric is in that tradition but is applied to edits expressed in a learned parameter decomposition. Classical dictionary learning and independent component analysis give conditions under which a basis is identifiable—sparsity of use (Spielman et al., 2012) and non-Gaussian independence (Comon, 1994)—and our always-active sign task violates the sparse-use condition, consistent with the scoping remark (Remark 1). Superposition toy models motivate sparse activity as the regime of interest (Elhage et al., 2022).

Non-identifiability and data dependence. Seed dependence and non-identifiability are well documented (Paulo & Belrose, 2025; Gerasimov et al., 2026; Nelson et al., 2026; Venkatesh & Kurapath, 2026; Korznikov et al., 2026). A public, non-peer-reviewed technical note (reuellee, 2026) is closest to our crossover and orthogonality-penalty blindness; our setting differs by addressing all-mask weight-space faithfulness, arbitrary component capacity, and erasure cost. We further add a dimension-independent collateral floor, finite bounds on actually learned dictionaries, and a causal data intervention removing the obstruction. Throughout, components are intervention-relative: we report task-relative erasure and preservation cost, not reconstruction quality.

8 Limitations and conclusion

We established an exact separable masking theorem and a smooth Gaussian-mixture finite-penalty result. The latter is informative only in a small-noise regime, is averaged over targets, and uses bounded fixed-coefficient edits, exact-sum atoms, and an ideal population all-mask loss and gate power that omits the frequency term. Neither result shows that the full VPD objective fails, nor that

finite PGD or minibatch SGD attains optimality. The learned two-factor result ($C = 2$) covers one site at finite held-out inputs; its mean all-mask KL is neither pointwise nor population-level and is not task accuracy, and its ratios verified in Arb interval arithmetic are conservative lower bounds, not the actual ratios. The 72-run mask-grid audit is finite, and the phase interior is unconfirmed. Masking and selective control therefore require separate evaluation.

REPRODUCIBILITY STATEMENT

Proofs appear in Appendices A–C; the derivation of the bounds, the training protocols, and the learned-network protocol are in Appendices E–G. The accompanying anonymous supplement includes the theorem verifiers, the raw saved checkpoints of the controlled and learned-network studies bundled per study with the training scripts as run, a pinned reference (repository URL and commit hash) to the public native implementation, which the entry point fetches and verifies, the computed bounds and results, the exact rational and Arb audits, the scripts that regenerate every figure, and a README with a single entry point that replays all checks on CPU.

AI USE STATEMENT

AI was used for data visualization, figure generation, and code assistance. The authors take responsibility for all content in accordance with ICLR policy.

REFERENCES

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems*, 2023.

Dan Braun, Lucius Bushnaq, Stefan Heimersheim, Jake Mendel, and Lee Sharkey. Interpretability in parameter space: Minimizing mechanistic description length with attribution-based parameter decomposition. *arXiv preprint arXiv:2501.14926*, 2025.

Lucius Bushnaq, Dan Braun, and Lee Sharkey. Stochastic parameter decomposition. *arXiv preprint arXiv:2506.20790*, 2025.

Lucius Bushnaq, Dan Braun, Oliver Clive-Griffin, Bart Bussmann, Nathan Hu, Michael Ivanitskiy, Linda Linsefors, and Lee Sharkey. Interpreting language model parameters. `https://www.goodfire.com/research/interpreting-lm-parameters`, 2026. Goodfire technical report.

Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning multi-level features with Matryoshka sparse autoencoders. *arXiv preprint arXiv:2503.17547*, 2025.

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, Satvik Golechha, and Joseph Bloom. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. *arXiv preprint arXiv:2409.14507*, 2024.

Pierre Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, 1994.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022.

Gleb Gerasimov, Timofei Rusalev, Nikita Balagansky, Daniil Laptev, Vadim Kurochkin, and Daniil Gavrilov. Unstable features, reproducible subspaces: Understanding seed dependence in sparse autoencoders. *arXiv preprint arXiv:2606.12138*, 2026.

Goodfire Research. Parameter decomposition — handbook. `https://github.com/goodfire-ai/param-decomp/blob/964d8baf49cb8b32a1a45081e39d47a3c11eef42/docs/handbook.md`, 2026. Commit 964d8baf49cb8b32a1a45081e39d47a3c11eef42.

Anton Korznikov, Andrey Galichin, Alexey Dontsov, Oleg Rogov, Elena Tutubalina, and Ivan Oseledets. OrtSAE: Orthogonal sparse autoencoders uncover atomic features. *arXiv preprint arXiv:2509.22033*, 2025.

Anton Korznikov, Andrey Galichin, Alexey Dontsov, Oleg Rogov, Ivan Oseledets, and Elena Tutubalina. Sanity checks for sparse autoencoders: Do SAEs beat random baselines? *arXiv preprint arXiv:2602.14111*, 2026.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, 2022.

Walter Nelson, Theofanis Karaletsos, and Francesco Locatello. Toward identifiable sparse autoencoders. *arXiv preprint arXiv:2605.31245*, 2026.

Gonçalo Paulo and Nora Belrose. Sparse autoencoders trained on the same data learn different features. *arXiv preprint arXiv:2501.16615*, 2025.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

reuellee. A solvable model of feature absorption in L1 sparse autoencoders: an exact pure-strategy crossover, a non-identifiability wall, and GPU-verified predictions. Public technical note, `https://github.com/reuellee/sae-identifiability/blob/master/report.md`, 2026. Not peer reviewed.

Daniel A. Spielman, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries. In *Proceedings of the 25th Annual Conference on Learning Theory*, 2012.

Sohan Venkatesh and Ashish Mahendran Kurapath. On the non-identifiability of steering vectors in large language models. *arXiv preprint arXiv:2602.06801*, 2026.

Antoine Vigouroux and Lee Sharkey. Targeted recovery of weight-space mechanisms from neural networks. *arXiv preprint arXiv:2607.13047*, 2026.

A Proof of the exact optimum theorem

The four steps below prove Theorem 1 and cover arbitrary component capacity.

Proof idea. Faithfulness under every allowed mask forces a gate on exactly when its component acts, so the objective counts nonzero reads $v_i^\top x$ per input. On ± 1 inputs a read is nonzero at least half the time, with equality only for equal-magnitude coordinate pairs, and rank forces d nonzero reads, giving $d/2$. In a paired basis the rows of any coefficient edit are tied, so erasing behavior j moves its partner's row by a constant amount, and exact erasure of j forces some atom's coefficient to one and erases the partner outright.

Lemma 1 (Odd-part bound)**.** *Let X be centrally symmetric, let ϕ satisfy $\phi(t)-\phi(-t) = q\,t$ for some $q > 0$, and for columns c, c' let $D(x) = \phi(c^\top x) - \phi(c'^\top x)$. Then $\mathbb{E}D(X)^2 \geq \frac{q^2}{4}\mathbb{E}\big[((c-c')^\top X)^2\big]$. If moreover $\mathrm{Cov}(X) = I_d$ then*

$$\mathbb{E}D(X)^2 \;\geq\; \tfrac{q^2}{4}\,\|c-c'\|^2. \tag{11}$$

Under (1), $q = 1$.

Proof. The odd part gives $D(x) - D(-x) = q(c-c')^\top x$. Since $(D(x) - D(-x))^2 \leq 2D(x)^2 + 2D(-x)^2$ and $X \overset{d}{=} -X$, averaging gives $q^2\mathbb{E}[((c-c')^\top X)^2] \leq 4\mathbb{E}D(X)^2$. If $\mathrm{Cov}(X) = I_d$ and the mean is zero, $\mathbb{E}[((c-c')^\top X)^2] = \|c-c'\|^2$. Hence (11). □

Proof of Theorem 1. We argue in four steps.

Step 1: gates count nonzero contributions. By the exact-sum assumption the all-on preactivation is x. If $g_i(x) < 1$ then the free mask m_i ranges over the nondegenerate interval $[g_i(x), 1]$, and (3) says $\phi_j\big(x_j + (m_i - 1)(A_i x)_j\big) = \phi_j(x_j)$ for every such m_i. If $(A_i x)_j \neq 0$ this makes ϕ_j constant on an interval with endpoint x_j, contradicting $\phi_j'(x_j) \neq 0$. Hence $g_i(x) < 1$ forces $A_i x = 0$. Conversely $A_i x = 0$ makes atom i contribute nothing, so $g_i(x) = 0$ is optimal: a gate is on exactly when its component acts on the input. Therefore, for a nonzero atom,

$$g_i(x) > 0 \iff A_i x \neq 0 \iff v_i^\top x \neq 0, \tag{12}$$

and the objective equals the expected number of atoms with nonzero read activation.

Step 2: every nonzero atom is active at least half the time. Fix a nonzero read v and a coordinate ℓ with $v_\ell =: a \neq 0$. Condition on all signs except X_ℓ and write the rest of the activation as $S = \sum_{j \neq \ell} v_j X_j$. Then $v^\top X = aX_\ell + S$ vanishes for at most one sign, so $\Pr[v^\top X \neq 0] \geq 1/2$. Equality forces $S \in \{a, -a\}$ almost surely. A sum of $m \geq 2$ independent nonzero signed terms takes at least three distinct values (its maximum B, its minimum $-B$, and $B - 2b$ with b the smallest magnitude), so $m = 1$ and the other coefficient has magnitude $|a|$. Finally $d = \mathrm{rank}(I_d) = \mathrm{rank}(\sum_i A_i) \leq \#\{i : A_i \neq 0\}$, so at least d atoms are nonzero and the expected count is at least $d/2$, with equality only for paired reads.

Step 3: attainment for every dimension. For even d, use disjoint 2×2 Hadamard blocks $\left(\begin{smallmatrix}1 & 1\\ 1 & -1\end{smallmatrix}\right)$. For odd $d \geq 3$, add the invertible 3×3 block with rows $(1,1,0)$, $(0,1,1)$, $(1,0,1)$, whose determinant is 2. Let V have these rows $v_i^\top$, let U have columns u_i with $UV = I_d$, and set $A_i = u_i v_i^\top$. Then $\sum_i A_i = UV = I_d$, every read is paired, and every atom is active with probability exactly $1/2$, attaining $d/2$.

Step 4: paired reads force collateral. At any optimum Steps 2–3 give exactly d nonzero atoms: fewer cannot reach rank d, and any additional atom would strictly increase the expected count. Arrange them as U, V with $UV = I_d$, so V is invertible and $VU = I_d$. For $E = \sum_i w_i A_i = U\,\mathrm{diag}(w)V$ we get $VE = \mathrm{diag}(w)V$, i.e. $v_i^\top E = w_i v_i^\top$ for every i. Every coordinate j appears in some read, which by Step 2 is proportional to $r = e_j + s\,e_k$ with $k \neq j$, so $r^\top E = w_i r^\top$. Equivalently $r^\top W' = (1 - w_i)\,r^\top$ for $W' = I - E$, so on every input with $r^\top x = 0$ the two rows satisfy $W'_j x = -s\,W'_k x$; Corollary 1 uses this. From $r^\top E = w_i r^\top$ we get $E_j + sE_k = w_i r^\top$, so $E_k = s(w_i r^\top - E_j)$. If $\|E_j - e_j^\top\| \leq \eta$ then

$$\|E_k\| \;\geq\; \|w_i r^\top - e_j^\top\| - \eta \;=\; \sqrt{(w_i - 1)^2 + w_i^2} - \eta \;\geq\; \frac{1}{\sqrt{2}} - \eta, \tag{13}$$

with the last minimum at $w_i = \frac{1}{2}$. Here $E_j = e_j^\top E$ is row j of E and $W'_j = e_j^\top - E_j$. Erasure with $c = (W'_j)^\top$, $c' = 0$ gives $\|W'_j\| \leq 2\sqrt{\epsilon}$ by Lemma 1 with $q = 1$, i.e. $\eta = \|E_j - e_j^\top\| \leq 2\sqrt{\epsilon}$. Preservation with $c = (W'_k)^\top$, $c' = e_k$ gives MSE $\geq \frac{1}{4}\|E_k\|^2 \geq \frac{1}{4}[1/\sqrt{2} - 2\sqrt{\epsilon}]_+^2$, where the last step uses (13).

If $E_j = e_j^\top$ exactly, then $e_j^\top U\,\mathrm{diag}(w)V = e_j^\top UV$, and since V is invertible, $U_{ji}(w_i - 1) = 0$ for every i, where $U_{ji} = (u_i)_j$ and $V_{ij} = (v_i)_j$. Because $(UV)_{jj} = \sum_i U_{ji}V_{ij} = 1$, some i has $U_{ji} \neq 0$ and $V_{ij} \neq 0$; for that atom $w_i = 1$ and its read is proportional to $r = e_j + s\,e_k$, so $r^\top E = r^\top$, i.e. $E_j + sE_k = e_j^\top + se_k^\top$, hence $E_k = e_k^\top$ and $W'_k = 0$. The partner head then outputs $\phi_k(0) = 0$ on every input, and its preservation MSE is $\mathbb{E}[\phi_k(X_k)^2]$. □

B MIXTURE THRESHOLD, GATE POWERS, AND SPARSE ACTIVITY

Proof of Proposition 1. A read with k nonzero entries is active on axis points with probability k/d, and on corners with probability at least $1/2$, with equality only for the paired form by Step 2 of the proof of Theorem 1 (Appendix A). Faithfulness is imposed on the support of D_p, and the densities below are expectations under D_p. A one-coordinate read is always active on corners ($k = 1$), so its density is $a_{\mathrm{native}} = 1 - p + p/d$. A two-coordinate read has density at least $a_{\mathrm{pair}} = (1-p)/2 + 2p/d$; unequal-magnitude reads are active on every corner, and $k \geq 3$ gives at least $(1-p)/2 + kp/d \geq a_{\mathrm{pair}}$. Rank needs d nonzero atoms, so the minimum is $d\min(a_{\mathrm{native}}, a_{\mathrm{pair}})$, attained by the native basis when $a_{\mathrm{native}} \leq a_{\mathrm{pair}}$ and by the invertible paired bases of Step 3 of that proof otherwise. Solving $a_{\mathrm{native}} = a_{\mathrm{pair}}$ gives $d/(d+2)$. At $p = 0$ the equality characterization of Step 2 of that proof excludes every $k \geq 3$ read, and at $p = 1$ the one-coordinate read is the unique minimizer. □

Gate powers and the frequency penalty. Under exact faithfulness Step 1 of the proof of Theorem 1 (Appendix A) forces relevant gates to one and irrelevant gates to zero, so every positive gate power takes the same value as the gate count and the optimum set is unchanged. For a complete population batch of size N over which activation frequency is counted, the increasing per-atom frequency penalty

$$h(a) = a\big[1 + \beta \log_2(1 + Na)\big], \qquad \beta \geq 0, \tag{14}$$

is strictly increasing in the activation density a; since every nonzero atom has density at least $1/2$ and rank needs d atoms, the optimum is again $d\,h(1/2)$, attained only by the paired bases. Since h is strictly increasing, under D_p the minimum of $\sum_i h(a_i)$ over dictionaries is $d\,h\big(\min(a_{\mathrm{native}}, a_{\mathrm{pair}})\big)$ by the same density comparison as in the proof of Proposition 1, so the crossover is still $a_{\mathrm{native}} = a_{\mathrm{pair}}$. This covers the increasing population-frequency penalty, not minibatch expected loss or finite-penalty trainers.

Remark 1 (Sparse activity)**.** *Suppose instead each behavior is active independently with probability π, with x_j uniform on $\{+1, -1\}$ when active and $x_j = 0$ when inactive. A native one-coordinate*

read is active with probability π. An equal-magnitude paired read $e_j + s\, e_k$ is active with probability $1 - [(1-\pi)^2 + \pi^2/2] = 2\pi - (3/2)\pi^2$. The paired read is sparser than the native read only when $\pi > 2/3$. This compares these two read types only; we do not characterize all optima under this activity model.

C CONTROL BUDGET AND FINITE-ERROR PROOF

Setup. Let d be even, and let $X = S + \sigma Z \in \mathbb{R}^d$, where the entries of S are independent uniform signs, the entries of Z are independent standard Gaussians, $\sigma > 0$, and S and Z are independent. Fix atoms $A_1, \dots, A_C$ of rank at most one with $\sum_{i=1}^{C} A_i = I_d$ and $C \geq d$. The link ϕ acts coordinatewise and satisfies $\phi(0) = 0$, the lower and upper Lipschitz bounds

$$\alpha|t - t'| \;\leq\; |\phi(t) - \phi(t')| \;\leq\; |t - t'|, \qquad 0 < \alpha \leq 1,$$

and the odd part $\phi(z) - \phi(-z) = qz$ with $q > 0$. Each gate g_i is a measurable function into $[0, 1]$. The all-mask error is

$$e(x) = \sup_{g_i(x) \leq m_i \leq 1} \Big\| \phi\Big(\sum_i m_i A_i x\Big) - \phi(x) \Big\|_2,$$

and the objective is

$$J = \mathbb{E} \sum_i g_i(X)^\gamma + \lambda\, \mathbb{E}\, e(X)^2, \qquad \gamma > 0,\ \lambda > 0.$$

For each target coordinate j separately, fix coefficients w_{ij} with $|w_{ij}| \leq B$ and form the edit matrix $E_j = \sum_i w_{ij} A_i$. Let ϵ be any upper bound on the average erasure error, and let the average collateral be

$$\frac{1}{d} \sum_j \mathbb{E}\, \phi\big(((I - E_j)X)_j\big)^2 \;\leq\; \epsilon, \qquad c = \frac{1}{d} \sum_j \sum_{k \neq j} \mathbb{E}\Big[\Big(\phi\big(((I - E_j)X)_k\big) - \phi(X_k)\Big)^2\Big].$$

Theorem 2. *Suppose $J \leq K$. For every $\tau \in (0, 1)$, set*

$$\nu = \frac{K}{d\, \tau^\gamma}, \qquad \eta = \sqrt{K/\lambda}, \qquad \zeta = \frac{2B\eta}{\alpha(1 - \tau)}, \qquad a = \frac{2\sqrt{\epsilon}}{q} + \zeta + \sigma.$$

Whenever $0 < \nu < 1$ and $a \leq 1 - \nu$, the average collateral obeys

$$c \;\geq\; \frac{q^2}{4} \Big[\sqrt{\tfrac{1}{\nu} - 1}\,(1 - a) - \zeta\Big]_+^2.$$

Hadamard competitor. For the objective J above, the Hadamard construction below gives

$$\inf J \;\leq\; \frac{d}{2} + \lambda\, \frac{d\sigma^2}{2} \;=\; J_H,$$

with J_H as in (16). Hence every $(1 + \xi)$-approximate global optimum satisfies $J \leq (1 + \xi)\big(d/2 + \lambda d\sigma^2/2\big)$.

Corollary 2. *Take*

$$\phi(z) = \tfrac{1}{2} z + \tfrac{1}{2} \log(1 + e^z) - \tfrac{1}{2} \log 2,$$

so $\alpha = \frac{1}{2}$ and $q = 1.5$, and take $B = 1$, $d = 16$, $\gamma = 0.4$, $\tau = 0.5$, $\sigma = 0.001$, $\lambda = 10^5$, $\epsilon = 0.01$, and $\xi = 0.05$. Then $J_H = 8.8$ and $K = (1 + \xi) J_H = 9.24$, so the bound of Theorem 2 gives $c \geq 0.07448$. A coordinate dictionary with all gates equal to one has $J = 16$ and admits erasures with zero collateral damage.

Noise regime and scope. The bound requires λ large, so that ζ is small, and $\lambda\sigma^2$ small, so that the competitor's value and hence ν stays below 1. With $d = 16$, $\gamma = 0.4$, $\epsilon = 0.01$, $\xi = 0.05$, and λ and τ optimized numerically over a grid, the bound is positive at $\sigma = 10^{-3}$ and is vacuous (equal to zero) at $\sigma = 3 \times 10^{-3}$ and above. It is an averaged-over-targets collateral guarantee with bounded fixed edits against a constructed Gaussian-mixture all-mask adversary, not VPD frequency or training dynamics.

General form. Under the above setup $X \overset{d}{=} -X$, and since $(|X_j| - 1)^2 \le \sigma^2 Z_j^2$,

$$\big\| |X_j| - 1 \big\|_{L^2} \le \sigma.$$

For gates $g_i(x) \in [0,1]$ and $\tau \in [0,1)$, let $S_x = \{i : g_i(x) > \tau\}$ and assume $\mathbb{E}|S_X| \le \nu d$ with $0 < \nu < 1$. A mask m is admissible if $g_i(x) \le m_i \le 1$; the allowed error satisfies

$$\sup_{g_i(x) \le m_i \le 1} \Big\| \phi\Big(\sum_i m_i A_i x \Big) - \phi(x) \Big\|_2 \le e(x), \qquad \mathbb{E}e(X)^2 \le \eta^2,$$

where $\eta \ge 0$. Set

$$\zeta = \frac{2B\eta}{\alpha(1-\tau)}, \qquad a = \frac{2\sqrt{\epsilon}}{q} + \zeta + \sigma, \qquad [t]_+ = \max(t, 0).$$

The proposition below is the general statement from which Theorem 2 follows by the substitutions of the "From a budget bound to collateral" paragraph of the penalty proof.

Proposition. *If $a \le 1 - \nu$, then*

$$c \;\ge\; \frac{q^2}{4} \Big[\sqrt{\tfrac{1}{\nu} - 1}(1 - a) - \zeta \Big]_+^2. \tag{15}$$

Proof idea. To erase target j while preserving the others, the edit E_j must act on inputs like $e_j e_j^\top$: its j-th diagonal entry is about one and its other entries in row and column j are small. Split $E_j x$ into the part carried by components active on x and the part carried by inactive ones. The inactive part is small because faithfulness bounds what any allowed mask can change (Lemma 2). The active part has rank at most the number of active components, which is at most νd on average, and a matrix with near-unit diagonal and rank well below d must have large off-diagonal entries (Lemma 3). Those off-diagonal entries are the collateral, and Lemma 1 converts them into output error.

For erasure take $b = e_j^\top (I - E_j)$, $b' = 0$; for preserved head k take $b = e_k^\top (I - E_j)$, $b' = e_k^\top$, so $b - b' = -e_k^\top E_j$. Applying Lemma 1 (first inequality, since $\mathrm{Cov}(X) = (1 + \sigma^2) I$ here, not I) with rows transposed to columns, this yields

$$\Big(\frac{1}{d} \sum_j \mathbb{E}\big[((I - E_j)X)_j^2 \big] \Big)^{1/2} \le \frac{2\sqrt{\epsilon}}{q},$$

$$c \ge \frac{q^2}{4} \rho^2, \qquad \rho^2 = \frac{1}{d} \sum_j \sum_{k \ne j} \mathbb{E}\big[(E_j X)_k^2 \big].$$

Lemma 2 (Leakage through inactive components)**.** *For every x and j, $Q_j(x) = \sum_{i \notin S_x} w_{ij} A_i x$ satisfies $\|Q_j(x)\|_2 \le 2Be(x)/(\alpha(1-\tau))$; hence $\big(\frac{1}{d} \sum_j \mathbb{E}\|Q_j(X)\|_2^2 \big)^{1/2} \le \zeta$.*

Proof. The two masks below are both allowed, so both outputs lie within $e(x)$ of $\phi(x)$, yet they differ by $(1-\tau)/B$ times $Q_j(x)$ before ϕ. Fix j and inactive i ($g_i(x) \le \tau$), and set

$$a_i^\pm = \Big[\pm \frac{(1-\tau) w_{ij}}{B(1 - g_i(x))} \Big]_+, \qquad m_i^\pm = g_i(x) + (1 - g_i(x)) a_i^\pm,$$

with $m_i^\pm = 1$ on S_x. Since $a_i^\pm \in [0,1]$, both masks are admissible, and

$$m_i^+ - m_i^- = \frac{1-\tau}{B} w_{ij} \quad \text{for } i \notin S_x,$$

$$\sum_i (m_i^+ - m_i^-) A_i x = \frac{1-\tau}{B} Q_j(x),$$

$$Q_j(x) = \sum_{i \notin S_x} w_{ij} A_i x.$$

Admissibility and lower Lipschitz give $\alpha \frac{1-\tau}{B} \|Q_j(x)\|_2 \le 2e(x)$, hence

$$\Big(\frac{1}{d} \sum_j \mathbb{E}\|Q_j(X)\|_2^2 \Big)^{1/2} \le \zeta.$$

□

Lemma 3 (Rank–trace). *Let M be a random $d\times d$ matrix with $\mathbb{E}\,\mathrm{rank}\,M \le \nu d$, $\vartheta = \mathbb{E}\,\mathrm{tr}(M-I) \ge -da$, and $T = \mathbb{E}\sum_j (M_{jj}-1)^2 \le da^2$, where $0<\nu<1$ and $a \le 1-\nu$. Then*

$$\frac{1}{d}\mathbb{E}\sum_j \sum_{k\neq j} M_{kj}^2 \ge \Big(\frac{1}{\nu}-1\Big)(1-a)^2.$$

Proof. For the orthogonal projector P onto the range of M, $\mathrm{tr}\,M = \mathrm{tr}(PM)$ and $|\,\mathrm{tr}(PM)| \le \|P\|_F\|M\|_F = \sqrt{\mathrm{rank}\,M}\,\|M\|_F$. Taking expectations and applying Cauchy–Schwarz, $(\mathbb{E}\,\mathrm{tr}\,M)^2 \le \mathbb{E}\,\mathrm{rank}\,M \cdot \mathbb{E}\|M\|_F^2 \le \nu d\,\mathbb{E}\|M\|_F^2$; with $\mathbb{E}\,\mathrm{tr}\,M = d+\vartheta$,

$$(d+\vartheta)^2 \le \nu d\,\mathbb{E}\|M\|_F^2 = \nu d\,(d + 2\vartheta + T + d\,c_{\mathrm{act}}), \qquad c_{\mathrm{act}} = \frac{1}{d}\mathbb{E}\sum_j\sum_{k\neq j} M_{kj}^2,$$

so

$$d\,c_{\mathrm{act}} \ge \frac{(d+\vartheta)^2}{\nu d} - d - 2\vartheta - T.$$

As $a \le 1-\nu$, for $\vartheta \ge -da$ the right side increases in ϑ and decreases in T; its minimum is at $\vartheta = -da$, $T = da^2$, giving $c_{\mathrm{act}} \ge (\frac{1}{\nu}-1)(1-a)^2$. □

Proof of the proposition. Define $\Lambda_j = E_j x - Q_j(x)$ and $M_{kj} = \mathrm{sgn}(x_j)(\Lambda_j)_k$. Its columns lie in $\mathrm{span}\{A_i x : i \in S_x\}$, so $\mathrm{rank}\,M \le |S_x|$ and $\mathbb{E}\,\mathrm{rank}\,M \le \nu d$. Its diagonal is near one:

$$M_{jj} - 1 = -\,\mathrm{sgn}(x_j)\big((I-E_j)x\big)_j + (|x_j|-1) - \mathrm{sgn}(x_j)(Q_j)_j,$$

three terms whose averaged L^2 norms are at most $2\sqrt{\epsilon}/q$ (Lemma 1), σ, and ζ (Lemma 2), so by Minkowski $T \le da^2$, and by Cauchy–Schwarz $\vartheta \ge -da$. Lemma 3 gives $c_{\mathrm{act}} \ge (\frac{1}{\nu}-1)(1-a)^2$ where $c_{\mathrm{act}} = \frac{1}{d}\mathbb{E}\sum_j\sum_{k\neq j} M_{kj}^2$. Since $E_j x = \Lambda_j + Q_j$, the reverse triangle inequality gives

$$\rho \ge \big[\sqrt{c_{\mathrm{act}}} - \zeta\big]_+ \ge \Big[\sqrt{\tfrac{1}{\nu}-1}(1-a) - \zeta\Big]_+.$$

Lemma 1 turns ρ into (15). □ We now instantiate (15) to prove the instance, Corollary 2.

Hadamard competitor. Partition the d coordinates of the input into $d/2$ disjoint pairs. For each pair with coordinates j,k, define the two orthonormal input directions

$$u_+ = \frac{e_j+e_k}{\sqrt{2}}, \qquad u_- = \frac{e_j-e_k}{\sqrt{2}},$$

and the corresponding rank-one atoms $A_\pm = u_\pm u_\pm^T$; the remaining $C-d$ slots are zero. In each pair the gate is binary: it assigns 1 to whichever of the two atoms has the larger absolute read $|u_+^T x|$ or $|u_-^T x|$ and 0 to the other, so exactly one atom per pair is active and $\mathbb{E}\sum_i g_i^\gamma = d/2$ for any $\gamma > 0$. Because the inactive atoms act on orthogonal directions, the reconstruction error admits the pointwise identity

$$\sup_{\text{allowed } m}\Big\|\sum_i m_i A_i x - x\Big\|^2 = \sum_{\text{pairs}} \frac{(|x_j|-|x_k|)^2}{2},$$

and since ϕ is 1-Lipschitz the realized error satisfies e^2 bounded by the right-hand side. Taking expectations under $X = S + \sigma Z$ with independent coordinates and using $\mathbb{E}|X_1| = \mathbb{E}|1+\sigma Z_1| \ge 1$ (conditional Jensen) together with $\mathbb{E}X_1^2 = 1+\sigma^2$,

$$\mathbb{E}e^2 \le \frac{d}{2}\,\mathrm{Var}\,|X_1| \le \frac{d}{2}\big(\mathbb{E}X_1^2 - 1\big) = \frac{d\sigma^2}{2}.$$

Hence the competitor objective is upper-bounded by

$$J_H \;=\; \frac{d}{2} + \lambda\,\frac{d\sigma^2}{2} \tag{16}$$

for $J = \mathbb{E}\sum_i g_i^\gamma + \lambda\,\mathbb{E}e^2$.

From a budget bound to collateral. Suppose a feasible dictionary and gating policy satisfy $J \le K$. Since $g \in [0,1]$, each gate exceeding $\tau > 0$ contributes at least τ^γ, so $\mathbb{E}\,\mathrm{count}\{g > \tau\} \le K/\tau^\gamma$; likewise $\mathbb{E}e^2 \le K/\lambda$, so the mask RMS obeys $\eta = \sqrt{K/\lambda}$. Setting $\nu = K/(d\,\tau^\gamma)$, together with $\zeta = 2B\eta/[\alpha(1-\tau)]$ and $a = 2\sqrt{\epsilon}/q + \zeta + \sigma$ as in the statement of (15), the bound (15) applies verbatim whenever $0 < \nu < 1$ and $a \le 1 - \nu$.

Main smooth activation. For $\phi'(z) = \frac{1}{2} + \frac{1}{2}\,\mathrm{sigmoid}(z) \in (\frac{1}{2}, 1)$ with $\phi(0) = 0$, the lower-Lipschitz constant is $\alpha = \frac{1}{2}$, $B = 1$, and the odd part has slope $q = 1.5$. Fix $d = 16$, $\gamma = 0.4$, $\tau = 0.5$, $\sigma = 0.001$, $\lambda = 10^5$, $\epsilon = 0.01$, $K = 9.24$. The competitor upper bound (16) is $J_H = 8 + 0.8 = 8.8$, and every 5%-approximate optimum satisfies $J \le 1.05 \inf J \le 1.05\, J_H = 9.24 = K$; no attainment assumption is needed. Interval arithmetic gives

$$\nu \in [0.762015,\ 0.762016], \qquad \zeta \in [0.0768999,\ 0.0769000],$$

$$a \in [0.211233,\ 0.211234],$$

so $0 < \nu < 1$ and $a < 1 - \nu$. Evaluating the bracket in (15) on an Arb interval at 100-bit precision yields a collateral lower bound in $[0.0744876,\ 0.0744877]$, hence $c \ge 0.07448$; all constants above are rigorous and the 100-bit interval evaluation is verified in Arb. This completes the proof of Corollary 2.

D Scalar constants and elementary identities

Odd-part identity. For GELU, $\phi(t) = t\Phi(t)$ and $\phi(-t) = -t(1 - \Phi(t))$, so $\phi(t) - \phi(-t) = t\Phi(t) + t(1 - \Phi(t)) = t$. For SiLU, $\phi(t) = t\sigma(t)$ and $\phi(-t) = -t(1 - \sigma(t))$, so $\phi(t) - \phi(-t) = t\sigma(t) + t(1 - \sigma(t)) = t$. The identity case is immediate.

Lipschitz constants. GELU satisfies $\mathrm{GELU}'(t) = \Phi(t) + t\,\mathrm{pdf}(t)$, whose extrema occur at $t = \pm\sqrt{2}$ and whose maximum is below 1.13. The SiLU derivative maximum occurs where $z\tanh(z/2) = 2$ and equals $\frac{1}{2} + \frac{z}{4}$; since $e^{2.5} > 9$ gives $z < 2.5$, the maximum is below $9/8 < 1.13$, and the negative extremum is $1 - \max$. The code proves these bounds with outward rational arithmetic.

Clean cross-entropy. Let $\theta = \phi(1)$ be the clean correct-logit value, i.e. $\Phi(1)$ for GELU and $\sigma(1)$ for SiLU; it is *not* a correct-choice probability. The actual binary correct-choice probabilities are $\sigma(8\theta)$ and $\sigma(8(1-\theta))$ on the two signs. For a binary label, gain 8 and correct-logit value θ, the clean per-head CE is $\mathrm{CE}(\theta) = \frac{1}{2}[\mathrm{softplus}(-8\theta) + \mathrm{softplus}(-8(1-\theta))]$, increasing for $\theta > 1/2$. GELU gives $\theta = \Phi(1) < .841345$ and hence CE $< .124435$; SiLU gives $\theta = \sigma(1) < \Phi(1)$, so the GELU value is the conservative baseline.

Logistic tangent. $\mathrm{softplus}(z) = \log 2 + \frac{1}{2}z + \log\cosh(\frac{1}{2}z) \ge \log 2 + \frac{1}{2}z$, so $\mathrm{softplus}(-Y\ell) \ge \log 2 - \frac{1}{2}|\ell|$ for all real ℓ.

Conditional covariance. On $H = \{X_j = -sX_k\}$ the (j,k) block of $\mathrm{Cov}(X \mid H)$ is $\left(\begin{smallmatrix} 1 & -s \\ -s & 1 \end{smallmatrix}\right)$, with eigenvalues 2 and 0; other diagonal entries are one and other cross entries zero. Its operator norm is therefore 2.

E Derivation of the bounds

Convex dual and exact box tangent. With $q_\lambda(w) = \sum_{k \ne j} \|E_k(w)\|^2 + \lambda(\|E_j(w) - e_j^\top\|^2 - 4\epsilon)$, convexity gives, for any w_0 and $g = \nabla q_\lambda(w_0)$,

$$q_\lambda(w) \ \ge\ q_\lambda(w_0) + g^\top(w - w_0) \qquad \text{for all } w.$$

The right side is affine, so its minimum over the box $[-B, B]^C$ is attained at the corner $w^\star = -B\,\mathrm{sign}(g)$, with value

$$\mathcal{L}_\lambda(w_0) = q_\lambda(w_0) - g^\top w_0 - B\|g\|_1.$$

For every $\lambda \geq 0$ and every w_0, on every target-feasible edit w with $\|E_j(w) - e_j^\top\|^2 \leq 4\epsilon$ the chain

$$\mathcal{L}_\lambda(w_0) \;\leq\; q_\lambda(w) \;\leq\; \sum_{k \neq j} \|E_k(w)\|^2$$

holds: convexity gives the first inequality, and feasibility makes the λ penalty nonpositive for the second. Thus $\mathcal{L}_\lambda(w_0)$ lower-bounds the constrained primal row energy $\sum_{k\neq j}\|E_k(w)\|^2$; dividing by 4 yields the nonlinear bound of (8). The reported witness is $\max(0, \mathcal{L}_\lambda(w_0)/4)$ for one chosen finite λ, w_0; searching over more λ, w_0 can only improve it, and no computed supremum over all λ, w_0 is claimed. Its arithmetic is exactly checked. This tangent bound is the one used for the empirical box bounds.

Prediction-loss bound. Write $r = e_j + s\, e_k$ (a column) and let $F_i = r^\top A_i - [(r^\top A_i r)/(r^\top r)]\, r^\top$, a row, be the component of $r^\top A_i$ orthogonal to r, so $r^\top A_i = \beta_i r^\top + F_i$ for some scalar β_i. Then $r^\top W' = r^\top - r^\top E = (1 - \sum_i w_i\beta_i) r^\top - \sum_i w_i F_i$, and on $H = \{r^\top X = 0\}$ this becomes $(W'_j + sW'_k)X = -(\sum_i w_i F_i)X$, not an unconditional row equality. Bounding $\|\sum_i w_i F_i\| \leq \delta$ as in (9), the conditional RMS of the residual is at most $\sqrt{2}\,\delta$. Here δ is the *pairing defect* of the dictionary with respect to r, vanishing when every $r^\top A_i$ is a multiple of $r^\top$ (as at every optimum of Theorem 1) and distinct from the dictionary residual Δ. For a saved dictionary the bound is evaluated for every partner $k \neq j$ and sign $s \in \{\pm 1\}$, and the largest value is reported. The erased row obeys $\|W'_j\| \leq 2\sqrt{\epsilon_{\text{logit}}}/\kappa$; since $W'_k X = -sW'_j X - s(\sum_i w_i F_i)X$ on H, the edited preserved logit has conditional RMS at most $2L\sqrt{2\epsilon_{\text{logit}}} + \kappa L\sqrt{2}\,\delta$. Averaging $\text{softplus}(-Y\ell) \geq \log 2 - \frac{1}{2}|\ell|$ over H and using nonnegativity off H gives (10). The exact-optimum case $\delta = 0$ is Corollary 1.

All-mask KL bound. For the six frozen native-KL models we bound the *original learned gates*. Write $z_o(m, x) = (\sum_i m_i A_i x)_o$ for the o-th preactivation and let d be the number of heads. For the two logits $(0, \ell)$, the Bernoulli KL is the softplus Bregman divergence, so the mean-head KL obeys the uniform bound

$$\overline{\text{KL}}(x, m) \;\leq\; \frac{\kappa^2}{2d} \sum_o M_o\, L_o^2 \left(z_o(m, x) - x_o\right)^2, \tag{17}$$

where the per-input sign $x_o \in \{+1, -1\}$ selects (L_o, M_o). Exact rational vertex evaluation verifies that every allowed mask $m \in [g(x), 1]$ keeps z_o in $[.95, 1.05]$ when $x_o = +1$ and in $[-1.05, -.95]$ when $x_o = -1$, and the clean preactivation $z_o = x_o$ lies in the same interval. Both activations satisfy $|\phi'| \leq 1.13$ globally and $\leq .15$ on the negative interval, so $L_o = 1.13$ on the positive side and $L_o = .15$ on the negative side; on the negative interval rational scalar bounds give GELU derivatives in $[-.133, -.029]$ and SiLU derivatives in $[.0295, .121875]$, both within $[-.15, .15]$. On the positive interval both logits satisfy $8\phi(z) > 5.3$, so the Bernoulli softplus curvature is at most $e^{-5.3} < 1/200$, giving $M_o = 1/200$; on the negative interval we use the global bound $M_o = 1/4$. Since $|\phi(z_o) - \phi(x_o)| \leq L_o|z_o - x_o|$ on the verified interval, it suffices to bound the right-hand side, a convex quadratic in m whose maximum over the box $[g(x), 1]$ is attained at a vertex; we maximize at the vertices for each of the four sign inputs and average the four per-input maxima, covering arbitrary input-dependent masks. All constants are verified by rational series, and the bound is computed exactly as a rational from the unchanged checkpoint U, V and recorded gates; native CI replay matches with zero difference and native forward replay differs by at most 9.54×10^{-7}, with printed bounds rounded outward. The provable all-mask bound is $\leq 7.43 \times 10^{-5}$, with prediction-loss increase $\geq .12122$ nats and strict $\text{L0} = 1$.

Population upper bounds. For a fixed edit E with $W' = I - E$, any zero-mean law with $\text{Cov}(X) = I$ gives $R_j \leq L^2\|W'_j\|^2$ and $C_j \leq L^2 \sum_{k\neq j}\|E_k\|^2$. Fitting E to $e_j e_j^\top$ by constrained least squares yields a feasible $w \in [0, 1]^C$; these formulas prove feasible erasure and preservation without sample-generalization assumptions.

F NATIVE TRAINING PROTOCOLS

The native reference implementation is used unmodified at commit `74146b555f3eb3cd2bc639c3789fd3b7a099094a`.

Table 2: Frozen native protocols. Both studies share the native shared-MLP CI (two hidden layers of width 16), leaky-hard gates, $\beta = .5$, shared PGD of step size .2 and weight 1, a gate-power anneal over 20%–80% of training, and random orthogonal exact reconstruction initialization.

	KL study	MSE confirmation
Task	$d = C = 2$, GELU+SiLU	four (d, C) in Table 1
Steps	8,000	6,000
Optimizer	Adam .003 $\to$ .0003	Adam .003 $\to$ .0003
Weight penalty	10	10
Minimality	.03 $\to$ .001	.03 $\to$ .001
Gate power	2 $\to$ .4	2 $\to$ .4
Masks per step	4 stochastic	4 stochastic
Shared PGD	5 steps, weight 1, from 20%	5 steps, size .2, from 80%
Seeds	30–35	100–108

The selection paragraph below applies only to the 72-run MSE study; the KL study reports all six starts with no selection. In the MSE study each replicate selects its lowest validation objective among its three starts; the objective is 10·weight MSE + mean-mask MSE + max audited MSE + .001·minimality, using common validation noise seed 960000 and independent test noise seed 960001. The main pre-specified pass criterion requires every dense bound $\geq .02$, axis actual erasure $\leq .01$ with collateral $\leq .005$, and both finite audited mask MSE $\leq .02$; success needs $\geq 10/12$ overall and $\geq 2/3$ per configuration. No edit, bound or behavioral outcome enters selection. Mask audits use a 31×31 grid for $C = 2$, a 5^C grid for $C = 4$, and 1,024 fixed random shared masks plus all-zero/all-one endpoints for $C = 8$.

G Learned-Bridge Experiment

Setup. We train fully learned dense GELU MLPs with shared hidden layers, architecture $2 \to 32 \to 32 \to 2$, on independent noisy binary-label tasks defined by rotated two-latent-sign targets. Activity laws: ordinary data with both factors active or inactive together; separating data with exactly one factor active and balanced marginals. Independent signs, $\sqrt{2}$ scaling, and normalized Gaussian $\sigma = 0.04$ noise give equal population moments. The label is correct with probability 0.95 when a factor is active and is an independent fair coin when it is inactive. Target training uses all activity combinations for 4000 steps. The first weight matrix is decomposed with $C = 2$; two starts per arm.

Training and selection. Native component losses train for 2000 steps; components are then fixed, and identically reset causal importance (CI) MLPs with two hidden layers of width 32 train for 2500 steps on the same common data, with 512 validation inputs per target. Selection is by validation objective among masking-qualified, edit-feasible candidates; collateral metrics do not rank dictionaries. One exploratory pilot failed the initial gate qualification, which checks the learned gates; a gate-only change preceded any test evaluation, and two fresh preregistered confirmations repeat the full procedure. The native code is unmodified; the finite training schedule is ours.

Rigorous bounds. Edits preserve the residual; masking omits it. We use Arb 80-bit ball arithmetic through saved binary weights with exact-erf GELU, enclose the critical minimum, and bound output intervals via convex KL extrema. Exact rational dyadic volume and non-overlap prove full coefficient cover for the ordinary and resampled dictionaries, for which every box is infeasible or incurs collateral exceeding 0.10, our conservative stopping target; the separating dictionary provides feasible low-cost repair witnesses, checked in Arb. The all-mask guarantee averages per-input mask suprema over finite test inputs and both heads.

Results. Gate counts are $\sim$ 1.24–1.26 for ordinary and resampled dictionaries versus $\sim$ 0.99–1.01 for separating; the target network is unchanged across arms. NumPy, direct, and native replays agree to $< 1.4 \times 10^{-6}$ in logits. All claimed ratios are conservative provable lower bounds over all coefficient edits in the declared box, over finite held-out inputs and controlled two-factor networks; they do not prove global optimality of decomposition training. Starts and artifacts are included in the supplementary material.