CALIBRATING CAUSAL-LOCALIZATION SCORES: FEATURE-SPACE COMPARISONS DEPEND ON INTERVENTION BUDGET AND WRITE-BACK SEMANTICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Causal-localization benchmarks compare representations by intervening on selected features and measuring the resulting counterfactual behaviour, and the scores are often read as properties of the representations themselves. We argue that this reading is incomplete: the measured effect is jointly determined by the representation and by the intervention used to interrogate it. We therefore treat causal localization as an *intervention-relative measurement*. Instead of assigning each feature space one intrinsic score, we characterise an *intervention-response profile*: causal response as a function of the residual-space intervention actually realized, together with its selection, write-back and evaluation conditions. Across RAVEL, MIB's IOI setting and a synthetic model with known causal structure, this distinction is consequential. Method orderings change across effective residual-space budgets, and rank-matched random subspaces acquire substantial causal effect. Nominally comparable learned interventions can occupy residual subspaces whose ranks differ by up to $9.9\times$ in our replay of MIB's learned-mask procedure, and re-optimising at a common residual-support rank changes the ordering in both successfully optimised large-mismatch tests. Write-back semantics add a separate asymmetry: changing only the operator substantially changes an SAE's causal score while leaving orthonormal bases invariant, and swapping an entire dictionary recovers 7% to 99% of the corresponding full-state causal effect across the settings we measure. These results motivate a calibration principle: compare interventions in the residual space where they act, report the intervention alongside the score, and distinguish localization from the capacity and transmission of the measuring intervention.

1 INTRODUCTION

Interchange-intervention benchmarks ask a seemingly simple comparative question: *which representation best localizes a causal variable?* A representation is selected, some of its coordinates are replaced using a source example, the modified representation is mapped back into the model, and the resulting counterfactual behaviour is scored. Causal abstraction gives the framework (Geiger et al., 2024); RAVEL (Huang et al., 2024), the Mechanistic Interpretability Benchmark (MIB) (Mueller et al., 2025) and SAEBench (Karvonen et al., 2025) use it to compare sparse autoencoder (SAE) dictionaries (Bricken et al., 2023; Huben et al., 2024), fitted rotations, PCA and neurons.

The difficulty is that the benchmark never observes the representation in isolation. It observes the outcome of an *intervention constructed through that representation*: which coordinates change, how many, how the edit is written back into the residual stream and on which counterfactuals it is graded together determine the perturbation the model actually receives.

This creates a measurement problem. Consider comparing "k SAE features with k dimensions of a rotation". For an orthogonal basis, selecting k coordinates defines a k-dimensional residual subspace; for an overcomplete dictionary, it specifies an edit whose realized residual perturbation depends on the selected decoder directions and on the write-back operation. Equal nominal feature counts need not give equal intervention capacity, so a benchmark that equates them can compare not only *where* a causal variable is represented but also *how much of the model state each method may manipulate*.

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.

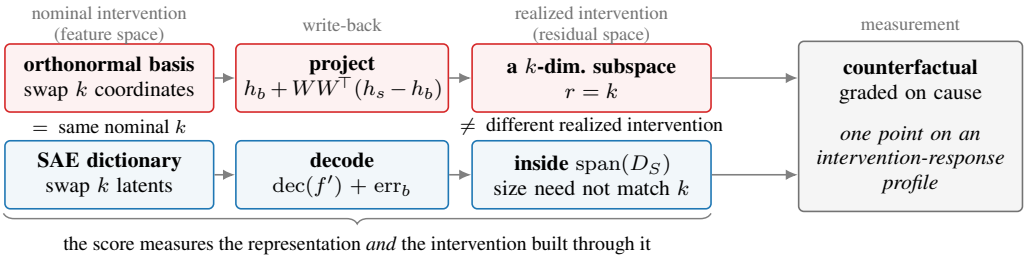

Figure 1: A causal-localization score is produced by a representation–intervention pipeline. A benchmark specifies an intervention in each representation's coordinates, but the model receives the resulting edit in residual space. *Top:* for an orthonormal basis the two coincide, $r = k$. *Bottom:* for an overcomplete dictionary they need not — the decoded edit lies in $\mathrm{span}(D_S)$, and only latents whose codes differ between base and source move it.

We therefore distinguish the *nominal intervention*, expressed in a representation's own coordinates ("swap k features"), from the *realized intervention*, the perturbation that reaches the residual stream (Figure 1). Different families turn nominally identical interventions into residual-space edits of different dimension, geometry and transmission, which makes residual space the natural common domain for comparing them. A leaderboard entry is then not an intrinsic property of a feature space but one point on an **intervention-response profile**: causal response as a function of the effective residual-space intervention, under a specified selector, write-back operator and evaluation distribution. A scalar score remains useful once its operating point is specified, but asking whether one representation "localizes better" first requires commensurate interventions.

Three distinctions are necessary for such a comparison. First, **intervention capacity must be calibrated in a common space** (§3). Method orderings are not stable across effective residual rank (Figure 2): a fitted rotation can lead at low rank and a dictionary at a larger one, even in a synthetic setting where the causal localization is known by construction. Capacity alone also produces causal effect — a rank-matched random subspace recovers a non-negligible fraction of the full-state intervention — so a score at one arbitrary budget cannot characterise a representation.

Second, **nominally equal feature-space interventions need not be equal realized interventions** (§4). Coordinate selection matters with the dictionary, nominal size and operator fixed, and changing only the write-back operator substantially moves an SAE's causal score while leaving an orthonormal basis unchanged, as the algebra requires. More importantly, learned masks over different families settle on very different residual-space capacities: in our replay of MIB's learned-mask procedure, the residual-support ranks selected for SAE and neuron interventions differ by factors of 1.1–9.9, and re-optimising both at a common residual-support rank changes the ordering in both successfully optimised $\sim$70% rank-reduction tests, while an 8% reduction leaves it unchanged. Residual rank need not determine the ordering, but an unmatched residual budget can be sufficient to change it.

Third, **localization must be separated from intervention transmission** (§5). A dictionary's edited values must still be mapped back into the model, so we measure *full-swap recovery*: the fraction of a full-state patch's causal effect recovered when every dictionary coordinate is swapped. Across the dictionary–site settings we measure it ranges from about 7% to 99%, so a low score through a dictionary can reflect poor transmission through its encode–swap–decode pathway as well as poor localization; none of seven cheaper proxies tracks it consistently across dictionary and site changes.

Contributions. **First**, we formulate causal localization as an intervention-relative measurement problem and introduce the intervention-response profile as the object being estimated (§2), separating the representation from the procedure that interrogates it. **Second**, we develop a residual-space calibration for comparing heterogeneous representation families, distinguishing nominal feature count from effective and realized residual-support rank and pairing the latter with rank-matched random controls (§3, §4.3–4.5). **Third**, we separate representational localization from intervention transmission by making selectors and write-back operators explicit (§4.1–4.2) and introducing full-swap recovery as a dictionary-specific diagnostic (§5). Together they yield a calibration protocol (§6).

We evaluate on RAVEL `city` attributes in gemma-2-2b, where we also replay MIB's learned-mask procedure; on MIB's IOI setting in GPT-2-small, under both write-back operators; and on a synthetic

transformer whose bottleneck makes the causal localization known by construction, so that budget sensitivity cannot be blamed on an unknown ground truth. The evaluation distribution matters as well: RAVEL's own fitting set exposes 99.2% of the graded `city/Country` answers, and a fitted rotation's advantage at large fitting sets holds only under that exposure (Appendix D).

None of this establishes a universally preferred feature family, operator or rank: some cells keep stable orderings across the ranks we test, and the common-rank comparison rests on three successfully optimised informative cells at two budget levels. The claim is methodological: without calibrating the realized intervention, a score cannot separate the representation from the intervention measuring it.

This relates to work showing that activation-patching conclusions depend on corruption procedures and output metrics (Zhang & Nanda, 2024), and to work normalising intervention magnitude across methods by residual edit norm, the closest existing attempt at a common scale (Bhalla et al., 2024). Our problem arises one level higher, when interventions built through *different representation families* are compared, and edit magnitude alone does not identify capacity, selection or write-back semantics. The closest precedent is probing controls (Hewitt & Liang, 2019; Voita & Titov, 2020): as probe accuracy is read relative to what capacity alone achieves, causal-localization scores should be read relative to what a matched residual-space intervention achieves.

2 Measurement setup

We now define the intervention that underlies the intervention-response profile. The key distinction is between the intervention specified in a representation's coordinates and the intervention ultimately realized in the model's residual space.

Benchmarks and models. **RAVEL** (Huang et al., 2024) asks whether an intervention can change one attribute of an entity while preserving another. We denote a task as `entity/attribute`; `city/Country` asks for a city's country. We run RAVEL on gemma-2-2b ($d = 2304$) at layer 14 with 65 536-atom Gemma Scope SAEs (Lieberum et al., 2024). On **MIB's IOI task** (Mueller et al., 2025; Wang et al., 2023) we localize two variables: `position`, which of two mentioned names is copied (two values; MIB's subject-position variable), at the final token, and `token`, which name is emitted (ranging over held-out names; an output variable of ours in place of MIB's subject token), at the indirect-object token. We run GPT-2-small ($d = 768$) at layers 9 and 8 respectively, with 24 576-atom jbloom SAEs. Finally, a 25M-parameter **synthetic transformer** ($d = 512$) has a bottleneck that makes "variable B is causally localized here" true by construction; an unrestricted twin differing only in its attention mask has cause 0.000 (defined below) at every layer.

From a nominal edit to a residual-space intervention. Let $h_b, h_s \in \mathbb{R}^d$ be the residual states of a base and a source example at the intervention site (a layer and token position). For an orthonormal basis W of the selected directions, our **projected-donor** intervention is $h' = h_b + WW^\top(h_s - h_b)$: it replaces the component of the base state in $\mathrm{span}(W)$ with the source's. The special case $W = I_d$ is the **full-state patch**, $h' = h_s$; its cause in a given cell is that cell's **full-state cause**, the denominator whenever we report a fraction of the full-state effect — a reference, not an upper bound.

Two notions of intervention size are central: k, the number of selected representation coordinates, and r, the dimension of the residual subspace the intervention can occupy, which we call the **effective residual rank** and normalise as r/d. For an orthogonal k-dimensional basis $r = k$; for other representations the two need not coincide. For a dictionary with decoder D and selected latent support S, the intervention lies in $\mathrm{span}(D_S)$, and we call $\mathrm{rank}(D_S)$ the **residual-support rank**: the same notion as r, named explicitly because it is not determined by the number of selected latents alone. Matching it matches the *dimension* of the space two interventions act in — not their orientation, per-example norm or learned contents — and our rank-matching claims are limited to it.

Representation families and selectors. For methods with an explicit collection of candidate directions we separate the **basis** from the **selector** that chooses coordinates from it, and orthonormalise the selected directions before applying the projected donor; `supervised` and `DAS` instead fit a rank-k subspace directly (Table 1). Basis controls — `pca`; `randrot`, a random rotation; `randdict`, a random dictionary matched to the SAE in width, norm and sparsity; `unembed`, the model's vocabulary

Table 1: Representation families and coordinate selectors. The two SAE rows use exactly the same dictionary and differ only in how coordinates are selected, a controlled test of selector dependence.

method	candidate directions	selector / fitted subspace
`sae`	SAE decoder columns	labelled between/within-class score
`saeunsup`	same SAE decoder columns	label-free mean activation
`neuron`	residual basis I_d	labelled between/within-class score
`supervised`	linear-discriminant (LDA) subspace fitted directly at rank k	
`DAS`	rank-k rotation fitted by single-task Distributed Alignment Search (Geiger et al., 2024)	

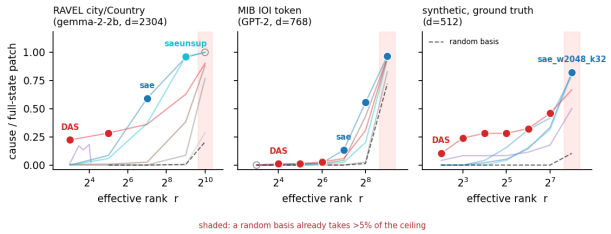

Figure 2: Method ordering changes across effective residual rank. Held-out cause over the same cell's full-state cause, against effective rank r (log scale). Filled markers name the leading family; an open grey marker means the apparent leader fails to beat the rank-matched random baseline, fails to separate from the runner-up, or both. Dashed: strongest random basis; shaded: ranks where a random basis recovers over 5% of the full-state cause. A fitted rotation leads at lower rank and a dictionary at larger rank in all three settings.

rows; and `random`, a random k-dimensional subspace — separate what learned geometry contributes from what an intervention of similar capacity obtains anyway.

Causal metrics. A **graded pair** is a base–source pair for which the unpatched model answers both prompts correctly and the demanded counterfactual token differs from the base answer. **Cause** is the fraction of graded pairs on which the patched model emits the source's value of the intervened attribute; **isolation** is the fraction on which a different attribute survives unchanged. We report them separately: RAVEL's aggregate is their mean, which awards 0.500 to an intervention that changes nothing (cause 0, isolation 1) and can rank methods very differently from causal accuracy.

Evaluation distribution and answer transport. **Coverage** is how much of the graded answer distribution the fitting set already contains: under *full* coverage all answer values may occur during fitting, under *zero* coverage every evaluated value is held out, and an **answer partition** is one such held-out split, across which comparisons are paired. A patch should transfer the causal variable, not copy the token the source prompt happened to produce, so we use **cross-attribute sourcing**: the source prompt asks about a different attribute from the base, and the measured source-answer copy rate is exactly 0 over all 4 685 graded cells, including for the full-state patch (Appendix A). A **cell** is one (method, rank, run) configuration under a fixed evaluation condition.

3 Intervention Capacity

Nominal feature count gives no common answer to how much of the model each method may manipulate, so we measure effective residual rank r, and find that **the comparative conclusion can change with intervention capacity**.

3.1 Method ordering is conditional on residual rank

Figure 2 exposes the intervention-response profile directly, and Table 2 gives its clearest case, RAVEL `city/Country` on gemma-2-2b at layer 14 under full coverage. At rank 8 `DAS` leads the runner-up by 0.176 cause; at rank 128 an SAE dictionary leads by 0.190. Both gaps are positive in all four answer partitions, and paired cluster bootstraps over graded pairs exclude zero in each ($+0.097$ to $+0.259$ at rank 8, $+0.141$ to $+0.266$ at rank 128). *Both conclusions are valid readings of the same benchmark at different intervention capacities*, so a benchmark that reports one operating point folds its budget choice into the apparent property of the representation.

Table 2: The RAVEL leader changes with effective rank. Held-out cause, `city/Country`, full coverage.

effective rank r	leader	runner-up	cause gap	partitions > 0
8	`DAS`	`supervised`	0.176	**4/4**
32	`DAS`	`supervised`	0.133	**4/4**
128	`sae`	`saeunsup`	0.190	**4/4**
512	`saeunsup`	`sae`	0.006	2/4
1024	`sae`	`saeunsup`	0.000	identical

The effect is not specific to one benchmark. On IOI `token`, GPT-2-small has three distinct leaders over the rank ladder; between the method orderings at any two ranks, mean Kendall τ_b is $+0.581$ with a minimum of 0.000. In the synthetic transformer, where the causal localization is fixed by construction, `DAS` leads at ranks 4–128 and a dictionary at 256 (mean τ_b $+0.245$, minimum -0.600, falling to -0.837 with a smaller fitting set), so rank sensitivity is not merely a consequence of not knowing where a variable resides. Nor must rankings always change: on the two-valued IOI `position` variable, `DAS` leads at every rank from 8 to 512 (minimum rank-to-rank τ_b of $+0.400$). A benchmark cannot know which regime it is in without sweeping.

3.2 Useful comparisons occupy an intermediate capacity regime

A single operating point can be uninformative even when the ordering does not change. At small budgets most methods sit at the causal floor (under zero coverage, 2 of 11 exceed cause 0.02 at $k \leq 32$); at large ones the leaders differ by as little as 0.006, distinct selectors give identical per-example outcomes, and random directions recover a substantial fraction of the full-state effect. We call the band in between the **discriminating window**; its location is benchmark- and variable-dependent, so it must be measured rather than fixed.

3.3 Intervention capacity has its own causal baseline

Capacity produces causal effect even when the subspace is random, so for each cell we measure a **random-basis floor**: the cause of a random residual subspace at the same effective rank. Below $r/d = 0.22$ four random constructions — a rotation, a subspace, a norm- and sparsity-matched 65 536-atom dictionary and its dense variant — agree to within 0.004, so shared capacity sets the baseline; it reaches 21% of the full-state cause at $r/d = 0.44$ on gemma-2-2b and 72% at 0.667 on IOI `token` (Figure 4, Appendix C). Our registered rule — a floor below 0.009 of the full-state cause for $r/d < 0.2$, drawn from high-cardinality cells — is falsified by IOI `position` (0.276 at $r/d = 0.167$), and across five RAVEL attributes with 6 to 137 answer values the floor falls as cardinality rises (Spearman -0.900, exact two-sided $p = 0.083$). It must be measured at the same rank and on the same cell, and a method judged by what it achieves *beyond* it.

Requested rank is not necessarily effective rank. LDA caps `supervised` at classes $- 1$, so `supervised_k512` on `city/Country` has effective rank 127 and one score at requested ranks 128–1024 (Appendix B); we report budgets where they are realized.

4 Intervention construction

Matching capacity does not determine the intervention: methods in subspaces of equal dimension can choose different directions (*selection*), and one edit can be written back differently (*write-back*). We isolate each, then turn to learned interventions, where selection and capacity interact.

4.1 Coordinate selection changes the intervention at fixed capacity

`sae` and `saeunsup` (Table 1) use the same dictionary and the same encode $\rightarrow$ score $\rightarrow$ top-k $\rightarrow$ orthonormalise $\rightarrow$ projected-donor pipeline, and differ only in the scoring rule: labelled between/within-class for `sae`, label-free mean activation for `saeunsup`. Any difference is therefore coordinate selection alone, and it is largest where the benchmark discriminates: held-out cause of `sae` minus `saeunsup` reaches $+0.208$ on `city/Language` at zero coverage, $+0.336$ on IOI `token`, and $+0.512$ on IOI `position` at $k = 8$ (Table 6, Appendix E). At the top of the RAVEL ladder the two become

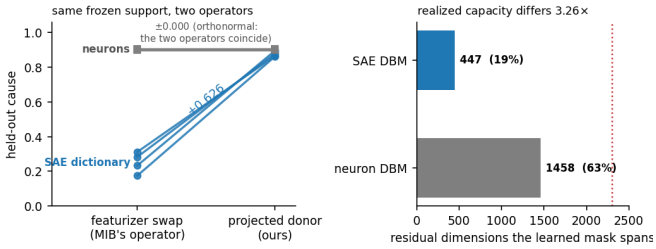

Figure 3: Write-back semantics affect an overcomplete dictionary but not an orthonormal basis. Masks learned over SAE latents and over raw residual coordinates by MIB's described procedure are frozen and scored under both operators. *Left:* averaged over the four runs shown, SAE cause rises from 0.250 (featurizer swap) to 0.876 (projected donor) while the neuron basis moves by under 10^{-6}; the ordering does *not* reverse. *Right:* the learned masks span very different residual budgets, 447 against 1458 dimensions on average.

identical on every graded example, while on IOI `position` the effect peaks at the opposite, low-rank end. A representation therefore does not uniquely determine a causal-localization intervention; the benchmark must also specify how directions are chosen from it.

4.2 Write-back semantics are part of the intervention

For a non-orthogonal or overcomplete representation a second choice remains: how the feature-space edit returns to the residual stream. MIB's causal-variable track uses the `CausalAbstraction` library (causalab contributors, 2025), whose SAE stage maps $h \mapsto (\mathrm{enc}(h),\ h - \mathrm{dec}(\mathrm{enc}(h)))$ and back by $\mathrm{dec}(f) + \mathrm{err}$. With decoder $D \in \mathbb{R}^{d \times M}$ (M atoms), codes $f_b = \mathrm{enc}(h_b)$ and $f_s = \mathrm{enc}(h_s)$, $\Delta f = f_s - f_b$, and f' the base code with a selected support S taken from the source, MIB's **featurizer swap** returns $\mathrm{dec}(f') + (h_b - \mathrm{dec}(f_b))$, keeping the base's reconstruction residual, while the **projected donor** replaces the residual component in the orthonormalised selected span. **For an orthonormal basis the two are algebraically identical** — encoding is exact, the error vanishes, and swapping *is* the projection — a direct correctness control.

Changing only write-back changes the SAE score. `neuron` scored both ways agrees to the last reported digit in 110 of 110 comparisons (96 on RAVEL, 14 on IOI, cause and isolation, including values as large as 0.813), as does an independently sampled orthonormal rotation, localising the asymmetry to the overcomplete representation. For the SAE, with the same basis, selector, k and graded pairs, `city/Country` cause at rank 128 falls from 0.481 under the projected donor to 0.060 under the featurizer swap, and IOI `token` at rank 512 from 0.924 to 0.000 (Appendix F); over the 145 cells with nonzero projected-donor cause the median ratio is $8\times$ and 34 are infinite (cells nest within runs, so this is descriptive), while isolation moves the other way (median $0.410 \to 0.886$). The same holds on supports learned by MIB's procedure (Figure 3): **write-back semantics are part of the intervention being measured**, moving one family while leaving another exactly invariant.

4.3 Why matching feature count is not enough

Selecting k orthogonal coordinates gives a k-dimensional residual intervention; for the featurizer swap, k says only that k code entries were replaced, and the perturbation lies in $\mathrm{span}(D_S)$, depending on the selected columns and on which entries of Δf_S are nonzero per example. So **matched k does not imply a matched residual-space intervention**. Matching SAE latents and neurons by nominal k, our preliminary comparison had the operator alone reverse their ordering in 12 of 12 live RAVEL cells at $k = 1024$ and at every k in the second setting (Figure 5, Appendix G); we deliberately do *not* use this as evidence that write-back explains MIB's ordering.

4.4 Learned selectors can choose different residual budgets

MIB does not set equal numbers of SAE and neuron coordinates by hand; it learns a differentiable binary mask (DBM) for each. We replay MIB's *described* procedure under our protocol — the interchange objective with no sparsity term, gates hard-thresholded at evaluation — with one DBM over the 65 536 SAE latents and a corresponding one over the d raw residual coordinates, freeze each learned support, and score it under both operators: a specification-level replay, not MIB's pipeline.

Table 3: Learned supports: representation-specific write-back, unequal budgets. Ranges over seven cells; Δ cause is projected donor minus featurizer swap. In every cell cardinality equals residual-support rank.

	residual-support rank	featurizer swap	projected donor	Δ cause
`sae`, 7 cells	150–1367	0.000–0.593	0.148–0.896	$+0.148$ **to** $+0.689$
`neuron`, 7 cells	1348–1489	0.793–0.907	*identical*	$+0.000$

Table 4: Re-optimising at a common residual-support rank can change the ordering when the mismatch is large. Neuron rank reduced to the SAE's m; projected donor throughout.

cell	neuron rank, free $\to m$	neuron	Δ neuron	SAE	ordering
`Country/:284`	1489 $\to$ 1367 (8%)	0.919	-0.015	0.889	neuron, unchanged
`Country/:73`	1489 $\to$ 445 (70%)	0.115	$\mathbf{-0.789}$	0.859	**SAE**
`Language/:73`	1348 $\to$ 402 (70%)	0.024	$\mathbf{-0.769}$	0.709	**SAE**
`Country/:21`	1489 $\to$ 150 (90%)	both arms at floor			no inference

Two observations matter (Table 3). First, the operator effect stays representation-specific on supports we did not choose: the projected donor raises SAE cause by $+0.148$ to $+0.689$ while the neuron basis stays bit-identical (our registered prediction of > 0.30 in every cell failed; the effect is bounded by what the dictionary transmits). Second, and more important for the cross-family comparison, the two learned selectors choose very different capacities: the neuron DBM settles at 1348–1489 of 2304 residual dimensions and the SAE DBM at 150–1367, a ratio of $1.1\times$ to $9.9\times$ across releases. With each family free to size its own mask, **changing the operator reverses the SAE–neuron ordering in none of the seven cells**. This falsified our hypothesis that write-back alone explained the ordering, and exposed the larger issue: the families also choose their own residual budgets.

4.5 Re-optimising at a common residual budget can change the comparison

To separate representation family from realized capacity we re-optimise both masks at a fixed size m, with a hard top-m gate in the forward pass and straight-through gradients, so each is trained *at* the target size rather than truncated afterwards. All arms use the projected donor, $|S| = \mathrm{rank}(D_S) = m$ throughout, and results (Table 4) are held-out cause at zero coverage.

With a small mismatch, removing 8% of the neuron arm's rank changes its cause by -0.015 and leaves the ordering unchanged. In the two $\sim$70% reductions the unconstrained neuron advantage disappears (0.115 and 0.024 against SAE scores of 0.859 and 0.709); the cells are different attributes whose drops agree within 0.020, and both constrained neuron masks stay above the rank-matched random floor, so they are optimised interventions rather than arbitrary subsets — the check that post-hoc truncation fails (Appendix G). The 90% cell fails the optimisation diagnostic (final objective 8.523 against 0.08–2.67 in the runs that trained; the criterion predates the runs) and supports no inference. The conclusion is limited: rank need not determine which family wins, but **when learned methods receive substantially different residual-space capacities, their ordering can confound representation quality with intervention budget**. None of this shows that MIB's numbers are wrong, that either operator is correct, or that either choice produced MIB's ordering.

5 Dictionary transmission

For dictionaries a separate question remains: how much of a specified intervention's causal effect survives the mapping back into the model? This **intervention transmission** is distinct from localization — a dictionary can hold coordinates for the variable yet transmit their values poorly through its encode–swap–decode pathway — so we separate the two by removing selection.

5.1 Full swap isolates dictionary transmission

With reconstruction residuals $\mathrm{err}_b = h_b - \mathrm{dec}(f_b)$ and $\mathrm{err}_s = h_s - \mathrm{dec}(f_s)$, a featurizer swap of *every* coordinate ($k = M$) sets the code to f_s and keeps the base residual, giving $\mathrm{dec}(f_s) + \mathrm{err}_b = h_s - \mathrm{err}_s + \mathrm{err}_b$ — the full-state patch modulo a difference of reconstruction errors, with no selector or budget left to vary. A dictionary's **full-swap recovery** at a site is the cause of this full swap over

the full-state cause: if selection were no limitation, how much of the source state's effect would survive write-back? It is a transmission diagnostic, not a general score of dictionary quality.

It is not an upper bound, and we do not claim one. A partial mask μ induces $D(\mu \odot \Delta f)$ and the full swap $D\Delta f$; these are not ordered, since decoder columns can cancel ($d_1 = u$, $d_2 = -u$, $\Delta f = (1, 0.9)$ gives $0.1u$ for the full swap and u for the first coordinate alone). **It does not happen in our data**: across **352** partial-mask cells — every measured $k < M$, in every release, site and coverage arm — none exceeds its own run's full swap, the closest reaching 95.9%, and recovery is identical across paired coverage conditions in all 16 repeated runs. We report this as an empirical regularity of the interventions evaluated, not a theorem.

5.2 Transmission varies dramatically across dictionaries and sites

A complete dictionary swap transmits between about **7% and 99%** of the corresponding full-state causal effect across the settings we measure (Figure 6, Appendix H). *Across four Gemma Scope releases at one site*, differing in width and trained sparsity, recovery is descriptively ordered by measured code density L_0, the average number of latents a prompt activates (10.9 / 40.6 / 52.7 / 213.6 giving 26.5% to 81.8%; spread 0.553 against a registered 0.10), though the site axis breaks any simple density account and measured L_0 is 53–135% of the advertised value (Appendix I). *Across sites in one model and family* recovery ranges from 7.2% to 99.1% over layers 18 / 14 / 10: layers 10 and 14 have similar raw full-state causes (0.863, 0.796) yet recover 99.1% and 47.0%, and layers 14 and 18 have code densities within 1% yet differ $6.5\times$ — though layer 18's own full-state cause is only 0.256, so its 7.2% is a small share of a small number. *Across the two benchmark/model/SAE settings* it differs 47.0% vs 96.5%, but those factors change together and are not separately identified.

5.3 Common representation proxies do not identify transmission

Seven cheaper candidate proxies, spanning reconstruction, decodability, geometry, sparsity and structural controls, fail to track full-swap recovery across both dictionary and site changes (Table 9, Appendix J). Reconstruction explained variance (EV) ranks synthetic SAEs *inverse* to causal recovery, and a multinomial logistic probe decodes a random basis and `saeunsup` equally well (0.776) while their causes are 0.003 and 0.825. The two strongest, registered in advance with declared falsifiers, compare the reconstructed and true state difference — geometric delta fidelity, $\cos(\mathrm{dec}(f_s) - \mathrm{dec}(f_b),\, h_s - h_b)$, and a version weighted by the output gradient — and both *reverse* the site axis: layer 18 reconstructs the state difference better than layer 14 (0.694 vs 0.663) while transmitting roughly $7\times$ less. Full-swap recovery costs one extra intervention; we recommend reporting it with the raw full-state and full-swap causes, so that 7% of a weak effect is not read like 47% of a strong one.

6 A calibration protocol for causal-localization comparisons

A causal-localization score mixes the representation with the measurement: capacity sets how much of the residual state a method may manipulate, construction which edit is realized, and evaluation and write-back what effect is observed. Calibration is therefore part of defining what a leaderboard entry means, not a correction applied afterwards. Table 5 lists the quantities our experiments show are needed; Appendix K expands each into a minimal reporting unit. It is not exhaustive: patching estimands also carry interaction effects (Vaidyanathan et al., 2026), which it does not address.

7 Related work

Intervention measurements depend on intervention design. Activation-patching conclusions change with corruption procedure and outcome metric (Zhang & Nanda, 2024); ours is the complementary point that comparisons *across representation families* change because the families realize different interventions. Bhalla et al. (2024) normalise interventions by residual edit magnitude, but rank, selection and write-back also affect the measurement.

Cross-representation comparisons need a common intervention scale. RAVEL (Huang et al., 2024) sweeps percentages of each featurizer's own dimensionality, so on Llama2-7B "1%" is ≈ 41

Table 5: Calibration protocol for cross-representation causal localization. Each row changed a comparative conclusion in our experiments or is needed to identify the intervention behind a reported score.

stage	report	guards against
Capacity	requested k; effective rank r and r/d; a rank sweep; a rank-matched random control	arbitrary operating points; requested–effective rank mismatch; effect available from capacity alone
Construction	basis; selector; write-back operator; realized residual-support rank of learned or dictionary supports	crediting selector, write-back or learned-budget effects to the representation
Evaluation	cause and isolation separately; fitting/test value coverage; held-out partition; fitting- vs held-out performance	metric aggregation; answer coverage; conclusions tied to one held-out distribution
Transmission	raw full-state cause; full-swap cause; full-swap recovery	reading poor transmission as poor localization; recovery without its denominator

directions for a rotation but ≈ 164 latents for its 16 384-feature SAE; MIB (Mueller et al., 2025) integrates over a size curve in its circuit track, but its causal-variable comparison still needs a capacity notion common to heterogeneous representations, which learning subspace extent within one family (Wu et al., 2023) does not supply. Rank-matched random controls, which we found in none of RAVEL, MIB's causal-variable track or SAEBench (Karvonen et al., 2025), act as probing controls (Hewitt & Liang, 2019; Voita & Titov, 2020), measuring what capacity alone achieves.

Representation quality does not imply causal intervention quality. Makelov et al. (2024) show that successful subspace interventions can activate dormant pathways. SAEBench finds that proxy-metric gains need not translate to practical performance, and later audits find that its metrics are noisier than assumed (Chanin, 2026), that features passing geometric recovery can be causally inert (Bal, 2026), and that controlling an SAE feature need not control the behaviour it tracks (Cui et al., 2026). Our transmission results approach the issue from the other side: even when a representation reconstructs or decodes the relevant information, its write-back pathway need not preserve the causal effect of the corresponding state change, which motivates measuring the causal intervention directly.

8 Limitations

Several of these are themselves findings. **The site effect is unexplained**: recovery ranges 7.2%–99.1% across three layers, including 6.5× between L14 and L18 at matched measured code density. **Registered predictions that failed** include three mechanisms for the operator gap, α-interpolation (the basis-free partial patch $h_b + \alpha(h_s - h_b)$ at $\alpha = r/d$ as a model of the random-basis floor), a uniform operator cost and a uniform capacity gap; most consequentially, the learned-mask replay withdrew our claim that the operator alone reverses MIB's ordering. Two rank-matching procedures were discarded as invalid before the third worked. **Small samples**: the cardinality, dictionary and rank-matching sweeps rest on 3–5 cells, criteria were registered as effect sizes for that reason, and exact permutation p-values are used throughout (smallest attainable 0.333, 0.083, 0.017, 0.003 at $n = 3, 4, 5, 6$), and a within-cell budget sweep and a seed replication of the matched-rank result are the obvious next tests. **Scope**: two pretrained models and one synthetic model, two benchmark families, single-layer interventions, and a replay, not a run, of MIB's pipeline.

9 Conclusion

Causal-localization benchmarks aim to compare representations, but their scores come from *representation–intervention pairs*: orderings change with residual-space capacity, comparable-looking learned interventions realize very different residual budgets, write-back moves an overcomplete dictionary but not an orthonormal basis, and dictionary transmission ranges from 7% to 99% of the full-state effect. Causal localization is therefore an *intervention-relative measurement*, and a score is interpretable only relative to the intervention that produced it. The object to report is an intervention-response profile on a common residual-space scale, with its selector, write-back operator, evaluation conditions and, for dictionaries, a direct transmission measurement.

REPRODUCIBILITY STATEMENT

Every run reported in this paper is logged with its configuration and outputs, and every reported result can be recovered from these logs.

AI USE STATEMENT

Large language models were used for drafting and editing prose, for literature search, for writing analysis and plotting scripts, and for critical review of the manuscript — including suggestions for experimental controls and methodological checks, several of which we adopted. The authors independently evaluated those suggestions, made all final design decisions, pre-registered and executed the experiments, verified all AI-assisted code and every claim against the run files, and take responsibility for all results and interpretations reported here.

REFERENCES

Mohamed Abdessalem Bal. From geometric recovery to causal validation: A reproducible audit of sparse autoencoder features, from superposition geometry to causal inertness. *arXiv preprint arXiv:2607.12166*, 2026.

Usha Bhalla, Suraj Srinivas, Asma Ghandeharioun, and Himabindu Lakkaraju. Towards unifying interpretability and control: Evaluation via intervention. *arXiv preprint arXiv:2411.04430*, 2024.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

causalab contributors. causalab (formerly CausalAbstraction): Causal abstraction for mechanistic interpretability. https://github.com/goodfire-ai/causalab, 2025. Originally https://github.com/atticusg/CausalAbstraction, the causal-variable track repository linked by MIB; accessed 2026-09-15.

David Chanin. Are sparse autoencoder benchmarks reliable? *arXiv preprint arXiv:2605.18229*, 2026.

Mingyue Cui, Linghui Shen, and Xingyi Yang. SAE interventions are unreliable: Post-intervention recovery of suppressed behavior. *arXiv preprint arXiv:2606.18322*, 2026.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Causal Learning and Reasoning (CLeaR)*, 2024.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *EMNLP*, 2019.

Jing Huang, Zhengxuan Wu, Christopher Potts, Mor Geva, and Atticus Geiger. RAVEL: Evaluating interpretability methods on disentangling language model representations. In *ACL*, 2024.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *ICLR*, 2024.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum McDougall, Kola Ayonrinde, Demian Till, Matthew Wearden, Arthur Conmy, Samuel Marks, and Neel Nanda. SAEBench: A comprehensive benchmark for sparse autoencoders in language model interpretability. In *ICML*, 2025.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma Scope: Open sparse autoencoders everywhere all at once on Gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, 2024.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *ICLR*, 2024.

Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fiotto-Kaufman, Tal Haklay, Michael Hanna, Jing Huang, Rohan Gupta, Yaniv Nikankin, Hadas Orgad, Nikhil Prakash, Anja Reusch, Aruna Sankaranarayanan, Shun Shao, Alessandro Stolfo, Martin Tutek, Amir Zur, David Bau, and Yonatan Belinkov. MIB: A mechanistic interpretability benchmark. In *ICML*, 2025.

Sankaran Vaidyanathan, David Arbour, Aaron Mueller, Scott Niekum, and David Jensen. The curse of multiple mediators: Hidden interaction effects in activation patching. *arXiv preprint arXiv:2606.27510*, 2026.

Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. In *EMNLP*, 2020.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *ICLR*, 2023.

Zhengxuan Wu, Atticus Geiger, Thomas Icard, Christopher Potts, and Noah D. Goodman. Interpretability at scale: Identifying causal mechanisms in Alpaca. In *NeurIPS*, 2023.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *ICLR*, 2024.

A THE ANSWER-TRANSPORT CONTROL

Across **4685** graded (method, rank, run) cells in the **103** cross-attribute runs behind this paper, the rate at which a patch causes the model to emit the token the *source prompt* would itself have produced is exactly **0**, including for the full-state patch. The control is not vacuous: under the matched-source protocol that copy target *is* the graded token, so a subspace can score there by transporting an answer rather than by localising the variable.

B EFFECTIVE RANK, NOT REQUESTED RANK

LDA's subspace is capped at classes -1, so `supervised_k512` on `city/Country` is rank **127**, and its score is identical at requested ranks 128, 512 and 1024. Plotted against requested rank it looks like a method that does not benefit from more dimensions; it was never given them. This trap appeared three times in this work, each time in a different disguise.

C THE RANDOM-BASIS FLOOR

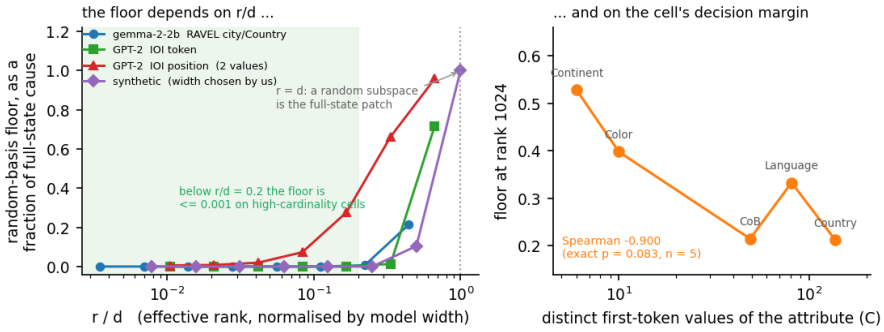

Figure 4: The causal effect available from intervention capacity alone. *Left:* random-basis cause, normalized by full-state cause, as a function of r/d across three model widths and two benchmark families. On high-cardinality variables the floor is small below $r/d \approx 0.2$ and large above ≈ 0.4; the synthetic setting reaches $r = d$, where a "random subspace" is the full residual space. IOI `position` is an important exception: despite $r/d = 0.167$, its random-basis floor is 0.276, for a variable with only two possible values (IOI series average two name splits). *Right:* at fixed rank 1024 across five RAVEL attributes, the random-basis floor decreases as answer cardinality increases. Together the panels show that the appropriate control is cell- and rank-specific rather than a universal function of r/d. The two panels agree at `city/Country` (0.21).

D VALUE COVERAGE AND THE FITTING-SIZE REVERSAL

Cross-attribute sources, gemma-2-2b L14, `city/Country`, DAS under RAVEL's protocol. The fitting-size reversal reproduces — DAS 0.893 against the SAE's 0.770 at $k = 128$, $n_{\text{fit}} = 1024$ — but *only* at full value coverage. At zero coverage the SAE leads at every fitting-set size we ran (40, 256, 1024) and at both k, while DAS climbs from 0.015 to 0.367 and never overtakes. Expected answer coverage for `city/Country` is 99.2% at 1787 fitting entities, 49.6% at 40 and 37.8% at 21 (400 draws; data and tokenizer only, no model run).

E COORDINATE SELECTION IN FULL

At the top of the RAVEL rank ladder, the two selectors become not merely statistically indistinguishable but *identical on every graded example*. The location of the selector-sensitive regime is itself task-dependent: on IOI `position` the largest effect occurs at the opposite, low-rank end of the sweep.

The implication is simple. A representation does not uniquely determine a causal-localization intervention. Even after the representation and residual-space capacity are fixed, the benchmark must specify how directions are chosen from that representation.

Table 6: Coordinate selection changes causal response even with the representation and intervention size fixed. Entries are held-out cause for `sae` minus `saeunsup`, averaged over answer partitions. The two methods use the same SAE dictionary and intervention pipeline and differ only in their selector. The effect is largest in the benchmark's discriminating regime and vanishes once both selectors saturate.

	`city/Country`		`city/Language`		
rank	full	zero	full	zero	IOI `token`
32	+0.030	+0.005	+0.009	+0.003	+0.005
64	+0.159	+0.019	+0.082	+0.125	+0.013
128	**+0.164**	**+0.169**	**+0.104**	**+0.208**	**+0.108**
256	−0.013	**+0.158**	−0.062	**+0.155**	**+0.336**
512	−0.001	−0.000	−0.001	+0.008	+0.026
1024	**0.000**	**0.000**	**0.000**	**0.000**	—

F Detailed write-back cost at matched nominal k

Table 7 gives the RAVEL values underlying the write-back comparison. Each cell uses the same SAE dictionary, selector, nominal k, graded pairs, and evaluation condition; only the write-back operator changes. Arrows denote projected-donor → featurizer-swap cause. Over all 145 live cells the median ratio is $8\times$, and $8.0\times$ on the 77 cells of the labelled selector alone.

Table 7: SAE causal response can change substantially under different write-back semantics at the same nominal k. Values are mean cause over the three answer partitions. Parentheses give the projected-donor / featurizer-swap ratio.

	$k = 128$	$k = 256$	$k = 1024$
`city/Country`, full coverage	$0.481 \to 0.060$ $(8.1\times)$	$0.621 \to 0.100$ $(6.2\times)$	$0.836 \to 0.144$ $(5.8\times)$
`city/Country`, zero coverage	$0.209 \to 0.004$ $(56\times)$	$0.587 \to 0.029$ $(20\times)$	$0.821 \to 0.078$ $(10.5\times)$
`city/Language`, zero coverage	$0.304 \to 0.015$ $(21\times)$	$0.604 \to 0.087$ $(7.0\times)$	$0.852 \to 0.135$ $(6.3\times)$

G The superseded matched-k diagnostic

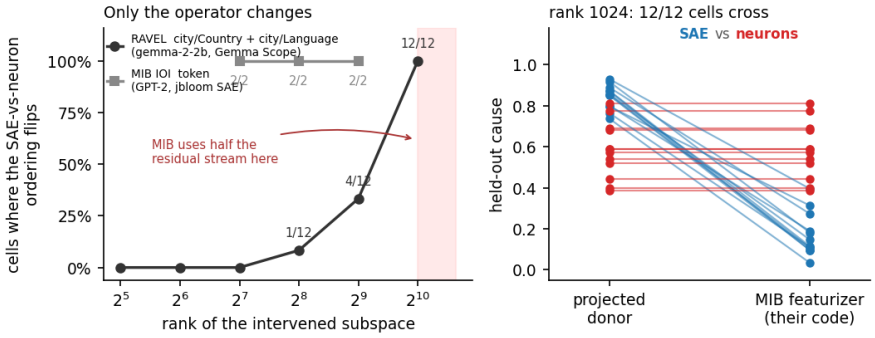

Figure 5: A superseded matched-k comparison illustrates why feature count is not a common intervention budget. *Left:* under a comparison that fixes the same nominal k for SAE latents and residual coordinates, the fraction of live RAVEL cells whose SAE–neuron ordering changes when only the operator is switched rises with k and reaches 12/12 at $k = 1024$. *Right:* at $k = 1024$, every live cell crosses. This is *not* the comparison on which our main conclusion rests: replaying MIB's described learned-mask procedure shows that the operator still changes SAE scores but no longer reverses the SAE–neuron ordering, because the learned masks occupy different residual-support ranks.

Table 8 gives the flip rate at each nominal rank.

The matched-k experiment in Figure 5 was an important diagnostic but is not our final cross-family comparison. Twelve RAVEL runs combine two attributes, two coverage conditions, and three answer partitions over eight ranks. A cell is *live* when the SAE has nonzero cause under the projected donor.

Across four Gemma Scope releases at the same site (`width_65k:21`, `:73`, `:284`, and `width_16k:83`), there are 102 live cells. The SAE leads under the projected donor in 99, while

Table 8: Ordering reversals under the superseded matched-k diagnostic. A cell is "live" when the SAE has nonzero cause under the projected donor. These results motivate, rather than establish, the need for a residual-space comparison.

nominal rank	ordering flips	live cells	flip rate
128	0	12	0%
256	1	12	8%
512	4	12	33%
1024	**12**	**12**	**100%**

neurons lead under the featurizer swap in 30. The operator-induced matched-k flip rate rises from 0% to 19%, 62%, and 88% from ranks 128 through 1024.

All three exceptions to the SAE lead under the projected donor occur for `width_65k:21`, the sparsest release and the release with the lowest full-swap recovery. We note the link to transmission (Section 5) but do not treat it as establishing a monotonic relationship.

Learned mask size varies with the dictionary. Across the three releases included in the learned-mask sweep, measured code density $L_0 = 11.3$, 42.1, and 217.9 corresponds to learned latent masks of 150, 445, and 1367 coordinates, respectively—from approximately 7% to 59% of the residual width. Thus under this learned-mask construction the realized budget is partly a consequence of the dictionary release rather than an externally fixed experimental constant.

The fourth matched-rank cell, `width_65k:21`, requires removing 90% of the neuron arm's unconstrained rank. Both constrained arms end near the floor, and the neuron optimization loss moves only from 10.528 to 8.523, compared with 0.08–2.67 in the matched-rank runs that trained. We therefore report this cell as an optimization failure rather than as a fourth ordering reversal.

Post-hoc truncation is not a valid rank-matching procedure. Truncating each family's freely learned mask to the other's cardinality sends the neuron arm to exactly 0.000 in the two deep cuts while leaving the shallow 94%-retention case at 0.856. Because the masks were optimized without a sparsity penalty, their gate values do not provide a reliable post-hoc importance ordering. The resulting deep cuts therefore behave like arbitrary subsets rather than interventions optimized at the target capacity.

Our final matched-rank experiment consequently trains each mask at the common size from the outset.

What the construction experiments establish. The section separates three quantities that a single causal-localization score otherwise mixes together.

Changing only the *selector* changes causal response even with the representation and intervention capacity fixed. Changing only the *write-back operator* can strongly change an overcomplete dictionary while leaving an orthonormal representation algebraically invariant. And allowing each representation to learn its own selector can implicitly allocate different *residual-space capacities*, large enough in our tests to change the resulting family comparison once the mismatch is removed.

These are properties of the measurement procedure, not interchangeable explanations of one another. In particular, our learned-mask replay falsifies our initial hypothesis that the operator alone explains the SAE–neuron ordering. The stronger conclusion is that a cross-representation comparison must specify both *what* intervention is selected and *what residual-space intervention that selection realizes*.

H Full-swap recovery across dictionaries and sites

I Measured versus advertised code density

Where we use code density L_0, we measure it directly on the prompts used in the experiment. Across six Gemma Scope releases, measured L_0 ranges from 53% to 135% of the nominal value; one release advertised as $L_0 = 66$ measures 89.1 on our RAVEL prompts. Since L_0 is used to characterize the

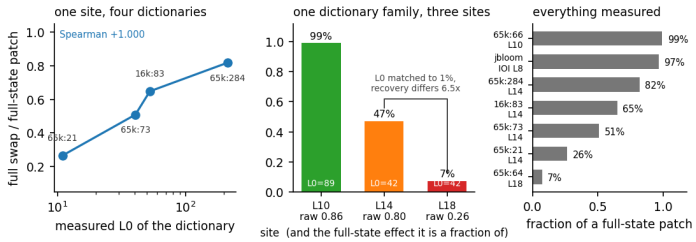

Figure 6: Full-swap recovery varies from 7% to 99% across the dictionary–site settings we measure. Recovery is the causal effect of swapping every dictionary coordinate divided by the causal effect of directly patching the complete residual state. Because every coordinate is swapped, coordinate selection is removed from the comparison. *Left:* four Gemma Scope releases evaluated at one site; recovery is descriptively ordered by measured code density L_0. *Middle:* one dictionary family evaluated across three layers. Layers 14 and 18 have measured code densities matched to within 1%, yet their recovery differs by $6.5\times$. Each bar also reports the raw full-state cause, because the denominator itself varies across sites: 7% of 0.26 is not comparable to 47% of 0.80. *Right:* all measured dictionary–site settings, spanning approximately 7%–99% recovery.

dictionaries and is associated descriptively with recovery along one axis, using measured rather than nominal density avoids introducing another hidden mismatch.

J TRANSMISSION PROXIES IN FULL

We test seven candidate proxies spanning reconstruction, decodability, geometry, sparsity, and structural controls (Table 9).

Table 9: Seven candidate proxies fail to track causal transmission consistently across the settings in which we test them. No proxy succeeds across both dictionary changes and intervention-site changes. In particular, the two preregistered delta-fidelity metrics reverse the site ordering: layer 18 reconstructs the source–base state difference better than layer 14 while transmitting roughly seven times less causal effect.

proxy	where measured	result
reconstruction EV	synthetic, ground truth	ranks the tested SAEs **inverse** to causal recovery
decodability	gemma-2-2b, matched rank	`random` and `saeunsup` both decode at **0.776**, while their causal effects are **0.003** and **0.825**
decodability	zero coverage	**undefined**: no target value is shared between fitting and test, leaving no common probe class
between-class variance concentration	2 cells	does not track causal performance; the basis concentrating the **least** between-class variance in its top coordinates performs best (SAE: 1.7% in top 128, causal rank 58; PCA: 27.4%, rank 362)
structural matching (`randdict`)	zero coverage	a 65 536-atom random dictionary matched in norm, L_0, and selector does not reach the registered threshold
$2L_0$ sparsity cap	*preregistered*	mean absolute error 0.104 for observed plateaus spanning 0.033–0.189
geometric delta fidelity	*preregistered*, 6 configurations	orders the tested dictionary axis correctly but **reverses the site axis**: L18 has higher fidelity (0.694 vs. 0.663) while transmitting approximately $7\times$ **less**
output-weighted delta fidelity	*preregistered*, 6 configurations	also reverses the site axis (-0.037) and performs worse overall ($+0.543$ vs. $+0.600$)

Several failures illustrate why transmission is not simply another name for representation quality.

Reconstruction quality is insufficient. On the synthetic model, reconstruction explained variance orders the tested SAEs in the opposite direction from causal recovery. Reconstructing residual activations well in aggregate therefore does not imply that the intervention pathway preserves the directions responsible for the causal behaviour being measured.

Information availability is insufficient. A multinomial logistic probe provides an even sharper example. At matched rank, a random basis and `saeunsup` both achieve decodability of 0.776, yet

their causal effects are 0.003 and 0.825. The relevant information can therefore be similarly decodable from two subspaces while interventions through those subspaces have radically different behavioral consequences.

Under zero coverage, the same decodability test is not even defined: the fitting and test sets share no target values, so there is no common output class on which the probe can be trained and evaluated.

Concentrating class variance is insufficient. The amount of between-class variance captured by the selected coordinates also fails to explain causal response. In the tested comparison, the SAE places only 1.7% of between-class variance in its top 128 coordinates while attaining causal rank 58; PCA places 27.4% there while attaining causal rank 362. The representation that concentrates less of this conventional signal measure performs better under the causal intervention.

Dictionary structure is insufficient. A random dictionary matched to the SAE in width, atom norms, measured sparsity, and selector does not reproduce the causal result under zero coverage. Matching these structural statistics therefore does not account for transmission.

Even direct state-difference fidelity fails across sites. Our strongest attempted proxies measure the actual source–base displacement.

The **geometric delta fidelity** is

$$\cos(\mathrm{dec}(f_s) - \mathrm{dec}(f_b),\ h_s - h_b)$$

over graded pairs. It asks whether the dictionary reconstructs the residual-state difference that an ideal source-to-base intervention would transfer.

The **output-weighted delta fidelity** additionally weights this reconstruction by

$$g = \nabla_h \left[\mathrm{logit}(\text{demanded token}) - \mathrm{logit}(\text{base answer})\right],$$

so that errors in directions to which the output is locally sensitive receive greater importance.

Both were preregistered with declared falsification criteria.

Geometric delta fidelity works along one axis: it correctly orders the tested dictionaries at a fixed site. But it fails on the site comparison. Layer 18 reconstructs the source–base state difference *better* than layer 14,

$$0.694 \quad \text{versus} \quad 0.663,$$

while transmitting roughly **seven times less** of its full-state causal effect.

Output weighting does not rescue the relationship. It also reverses the site ordering and performs worse overall in the tested configurations.

This failure clarifies what full-swap recovery measures. Aggregate geometric reconstruction—even reconstruction weighted toward directions locally associated with the output—does not establish that the model will respond to the reconstructed state difference as it responds to the original one.

Transmission should therefore be measured causally. None of the seven proxies we test tracks full-swap recovery consistently across dictionary and site changes. This does not establish that no useful proxy exists. It does show that the most natural inexpensive substitutes available to us—including two preregistered measures that explicitly compare reconstructed and true state differences—are insufficient over the settings tested here.

Full-swap recovery requires only one additional intervention endpoint: swap all dictionary coordinates and compare the resulting cause with the full-state patch. Because this operation removes feature selection from the comparison, it provides a direct diagnostic of the complete dictionary write-back pathway.

We therefore recommend reporting, for dictionary interventions,

$$\boxed{\text{raw full-state cause} \quad + \quad \text{full-swap cause} \quad + \quad \text{full-swap recovery.}}$$

The raw full-state cause establishes how much causal effect is available at the site; the full-swap cause measures how much of that effect survives dictionary transmission; and their ratio makes that

loss comparable across cells. Reporting all three prevents, for example, 7% of a weak full-state effect from being interpreted in the same way as 47% of a strong one.

Taken together with the previous sections, this completes the measurement decomposition. A causal-localization score can vary because the intervention has a different *capacity*, because it selects different *directions*, because those directions are written back with different *semantics*, or because the representation/write-back pathway transmits different fractions of the relevant causal state.

K THE CALIBRATION PROTOCOL IN FULL

This appendix expands each row of Table 5.

CALIBRATE INTERVENTION CAPACITY

A cross-representation comparison should first establish a common residual-space intervention scale.

We recommend reporting both the **requested size** k and the **effective residual rank** r. The distinction matters even for methods that appear to take an explicit rank argument: an LDA subspace, for example, is capped at classes $- 1$, so increasing requested k need not increase the intervention actually available to the method.

The primary comparison should be made against effective rank r, preferably over a sweep rather than at a single operating point. Our rank experiments show why: the leading representation family can change within the discriminating region of the same benchmark. A scalar measurement at one rank is therefore one sample from an intervention-response profile rather than a complete characterization of the representation.

Each operating point should also include a **rank-matched random-subspace control**. Intervention capacity alone can produce nontrivial causal effect: in our experiments a random subspace reaches 21% of the full-state cause at $r/d = 0.44$ on gemma-2-2b and substantially more at larger normalized ranks on GPT-2-small. Moreover, this baseline is not a universal function of r/d; it also varies with properties of the evaluation cell, including answer cardinality. The relevant baseline must therefore be measured at the same rank and on the same task as the method being evaluated.

A minimal capacity report is thus

$$\boxed{(k,\ r,\ r/d,\ \text{method cause},\ \text{rank-matched random cause}).}$$

Where space permits, the intervention-response curve should replace the point estimate.

SPECIFY INTERVENTION CONSTRUCTION

Matching effective rank does not determine which intervention is performed. A benchmark should therefore specify both **selection** and **write-back semantics**.

For methods represented as a basis or dictionary plus a coordinate selector, the selector should be reported explicitly. Our controlled SAE comparison shows that changing the selector alone can substantially change causal response while holding the dictionary, rank, operator, and evaluation examples fixed.

The write-back operator must also be part of the method specification. In our experiments the featurizer swap and projected-donor operators are identical for orthonormal bases but substantially different for an overcomplete SAE. Reporting only "we intervened on these features" therefore leaves unspecified an operation capable of changing the resulting causal score.

For learned selectors, nominal mask size is insufficient. The **realized residual-support rank** should be measured and reported for each representation family. In our MIB-style replay, freely learned SAE and neuron masks occupy residual subspaces differing by factors of 1.1–9.9 across tested settings. Re-optimizing at a common residual-support rank changes the ordering in both successfully optimized large-mismatch comparisons, while the small-mismatch control does not.

The appropriate intervention description is therefore not merely

"method X, k features,"

but

$$\boxed{\text{representation} + \text{selector} + \text{operator} + \text{realized residual rank.}}$$

This does not make two interventions identical: equal residual rank does not equalize subspace orientation or per-example edit magnitude. It does, however, prevent differences in available residual-space dimension from remaining hidden inside a representation-family comparison.

State the evaluation distribution

An intervention-response profile is conditional not only on the intervention but also on the examples on which it is fitted and evaluated.

We therefore recommend reporting **cause** and **isolation** separately. Their mean can obscure the causal quantity of interest: RAVEL's aggregate assigns 0.500 to an intervention with cause 0 and isolation 1, i.e. an intervention that changes nothing.

The relationship between fitting and evaluation values should also be explicit. In particular, an entity-level train/test split does not imply that the values being predicted are unseen during fitting. We use **value coverage** to distinguish these cases:

$$\text{coverage} = \text{fraction of the graded answer distribution represented during fitting.}$$

This distinction changes conclusions on RAVEL. Under RAVEL's fitting setup, the expected coverage of graded `city/Country` answers is 99.2% at 1787 fitting entities. The fitting-size reversal between DAS and the SAE reproduces under full value coverage, but disappears under zero coverage: when every graded test value is withheld during fitting, the SAE leads at every fitting-set size we test.

When a selector or subspace is fitted using task labels, it is also useful to report performance on the fitting entities beside performance on held-out entities. The distinction can be large. At rank 8 in our RAVEL experiment, `DAS` reaches a fitting-entity cause of 0.995 but a held-out cause of 0.190, and `supervised` 0.316 against 0.013. A high fitted causal response and a high held-out causal response therefore answer different questions.

Finally, the source construction should rule out trivial answer transport where applicable. Our cross-attribute sourcing makes the token naturally produced by the source prompt different from the token on which the counterfactual is graded; across 4 685 graded cells, the measured source-answer copy rate is exactly zero.

A minimal evaluation specification should therefore identify

$$\boxed{\text{cause}, \quad \text{isolation}, \quad \text{fitting/test coverage}, \quad \text{held-out partition},}$$

together with any source construction needed to distinguish causal transfer from answer copying.

Report dictionary transmission separately from localization

For dictionary representations, the score obtained from a selected subset should be accompanied by a direct measurement of the write-back pathway.

We recommend three quantities:

$$\boxed{C_{\text{state}}, \qquad C_{\text{swap}}, \qquad R_{\text{swap}} = \frac{C_{\text{swap}}}{C_{\text{state}}},}$$

where C_{state} is raw full-state cause, C_{swap} is cause after swapping every dictionary coordinate, and R_{swap} is full-swap recovery.

The three numbers answer different questions. C_{state} measures how much causal effect is available at the intervention site. C_{swap} measures how much of that effect survives the complete dictionary write-back pathway. Their ratio makes transmission losses easier to compare across settings.

Table 10: Minimal reporting unit for a causal-localization comparison. The first four rows apply generally; the final row applies to dictionary-based interventions.

component	report
experimental setting	benchmark/task, model, intervention layer/site, evaluation partition, value-coverage condition
intervention capacity	requested k, effective residual rank r, normalized rank r/d, rank-matched random baseline
intervention construction	representation/basis, selector, write-back operator, realized residual-support rank where applicable
outcome	cause and isolation separately, together with raw full-state cause
dictionary transmission	full-swap cause and full-swap recovery

All three should be shown together. A recovery of 7% when the underlying full-state cause is 0.256 should not be read in the same way as 47% recovery when the underlying full-state cause is approximately 0.8.

We prefer direct full-swap evaluation to substituting reconstruction or geometric proxies. Across the seven proxy families tested in Section 5.3, none tracks recovery consistently across both dictionary and site changes.

A MINIMAL REPORTING UNIT

The protocol can be summarized as a small report card attached to each headline comparison (Table 10):

For example, instead of reporting

> "SAE cause $= c_1$ and DAS cause $= c_2$,"

the comparison should identify the operating point that generated those values:

> "Held-out cause under zero coverage at effective residual rank r, with the selector and write-back operator named; the rank-matched random cause and the full-state cause C_{state} reported beside it; and, for the dictionary arm, C_{swap} and $R_{\text{swap}} = C_{\text{swap}}/C_{\text{state}}$."

The exact presentation can vary, but these quantities make explicit what the scalar scores are conditional on.

WHAT CALIBRATION DOES—AND DOES NOT—ESTABLISH

The goal of this protocol is not to define a single "correct" causal intervention. Different selectors, operators, ranks, and evaluation distributions can correspond to different legitimate questions.

Calibration instead makes those questions identifiable.

A comparison at one rank may be meaningful, but it should be understood as a comparison *at that rank*. A learned mask may legitimately choose different numbers of directions for different representations, but then its score jointly reflects representation quality and learned intervention capacity. A dictionary may yield a low selected-feature score, but that score should be interpreted differently when a full swap transmits nearly all of the full-state effect than when the same pathway transmits only a small fraction.

In this sense, the protocol changes the object being reported:

method score $\longrightarrow$ calibrated intervention-response measurement.

Our narrower claim is that the dimensions above are already necessary: each either changes a comparative conclusion in our experiments or determines what intervention a reported score actually represents.