# DE-ABSORPTION REPARAMETERIZATION: POST-HOC REPAIR OF FEATURE ABSORPTION IN SPARSE AUTOENCODERS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Sparse autoencoders (SAEs) are widely used to decompose neural-network activations into sparse, human-interpretable features. However, when the underlying features are hierarchically related, the same sparsity that promotes interpretability can distort their representation. In particular, a broad parent feature may stop activating when one of its more specific child features fires. This phenomenon, known as feature absorption, makes individual SAE features less reliable for interpretation and intervention. Existing approaches mainly aim to reduce absorption by modifying SAE training. In contrast, we ask whether feature absorption can be corrected after the SAE has already been trained. To address this question, we introduce De-Absorption Reparameterization (DAR), a post-hoc method that corrects feature absorption for a supplied relation between a parent feature and one or more child features in a trained SAE. DAR redistributes absorbed parent-related information from the child decoder back to the parent feature, while applying a compensating decoder update that preserves the SAE reconstruction exactly. Since reconstruction is preserved for any transfer coefficient, reconstruction alone cannot identify the correct repair. DAR therefore estimates the coefficients from parent–child activation patterns using sign-constrained least squares, ensuring that the parent activation only increases while all latent activations remain nonnegative. We evaluate DAR on JumpReLU and BatchTopK SAEs for Gemma-2-2B and CLIP ViT-L. Given the supplied parent–child relations, DAR restores parent activations on absorbed inputs across both language and vision. On the main language benchmark, DAR restores parent activation on up to $44.6\%$ of previously absorbed examples. The repaired activations are more usable for their intended interpretation: intended-latent AUROC on absorbed inputs increases from 0.535 to 0.941. Casually, ablating the repaired parent produces a 15× larger effect on the relevant model logit.

## 1 INTRODUCTION

Large Multimodal Models (LMMs) have achieved remarkable performance across a wide range of tasks, including visual question answering, image captioning, and visual reasoning (Alayrac et al., 2022; Laurençon et al., 2023). However, understanding their internal representations and decision-making processes remains a significant challenge, particularly in ensuring the reliability and trustworthiness of their predictions (Jacovi et al., 2021; Markus et al., 2021). As LMMs are increasingly adopted in high-stakes applications, the need for methods that provide meaningful insights into their internal representations and make their decision-making processes more interpretable has become increasingly important.

Sparse autoencoders (SAEs) have emerged as a powerful tool for interpreting the internal representations of neural networks (Templeton et al., 2024). SAEs decompose complex activation of neural network into sparse, human-interpretable features. Recently, SAEs have shown significant improvement in decomposing the dense, polysemantic activations of LMMs into more interpretable latent feature, through new sparsity mechanisms such as TopK (Gao et al., 2025), JumpReLU (Rajamanoharan et al., 2024), BatchTopK (Bussmann et al., 2024), and Matryoshka SAEs (Bussmann et al., 2025)



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.

Despite their effectiveness, SAEs still face important challenges that can limit the reliability of their learned features, particularly *feature splitting* and *feature absorption* (Chanin et al., 2025). As SAE width increases, broad features may split into increasingly specific features; for example, a general feature such as "math" may split into more specific features such as "calculus" and "geometry." Such splitting naturally induces hierarchical relations among SAE features, where a broad feature acts as a *parent* and more specific features act as its *children* (Bussmann et al., 2025). Ideally, when a child concept is present, the corresponding parent feature should remain active because the specific concept also implies the broader one. For instance, "elephant" belongs to the broader concept "animal". Feature absorption occurs, when a specific child latent captures part of the signal associated with its parent, causing the parent to remain inactive on inputs where the parent feature is still present(Chanin et al., 2025). For example, "a starts with L" latent may fail to activate on `lotus` or `long` because token-specific latent absorb the corresponding signal. As a result, the parent latent effectively represents starts with L, except for `lotus`/`long`," breaking the intended pattern. Such failures make individual SAE latents less reliable as units of interpretation and can affect downstream analyses that depend on stable feature identities, including causal interventions and circuit discovery.

Several recent methods attempt to mitigate feature absorption by modifying SAE training, including Matryoshka SAEs, Cross-sample Consistency Regularization ($C^2R$) (Jin et al., 2026), and Subspace-Aware Sparse Autoencoders (SASA) (Dalili & Mahdavi, 2026). These approaches intervene through changes to the SAE objective or architecture and therefore require training a new model. The distinction matters in practice: retraining does not repair the representation of an already-trained SAE, but learns a different sparse dictionary whose latent identities and decoder directions change, invalidating existing feature annotations, probes, dashboards, and circuit analyses.

This motivates a natural question: ***Can feature absorption be repaired directly in an already-trained SAE, without retraining its encoder or changing its representation?***

We answer this question with **De-Absorption Reparameterization (DAR)**, a post-hoc method for repairing supplied parent-child absorption relations in trained SAEs. DAR assumes a supplied hierarchy between broad parent latents and their more specific child latents, and redistributes representation within each parent-child group through a coupled code–decoder transformation. For a parent whose activation is suppressed when its children fire, DAR transfers a share of the representation carried by the supplied child latents to the parent coordinate: it increases the parent activation and correspondingly updates the child decoder directions. The paired updates preserve the decoded representation exactly, so DAR can change how representation is distributed among the supplied hierarchical latents without changing the SAE reconstruction.

Since reconstruction is preserved for any transfer coefficients under this reparameterization, reconstruction alone cannot determine how much should be transferred. DAR therefore estimates the transfer coefficients from parent-child activation patterns on a calibration set using sign-constrained least squares. The sign constraint restricts the coefficients to the direction required for parent restoration and guarantees that the repaired parent activation is non-decreasing while all repaired latent activations remain non-negative. DAR preserves the pretrained encoder and existing latent set while repairing the learned sparse representation to restore the supplied parent–child activation pattern. In this way, feature absorption can be addressed directly within an existing SAE while retaining its original encoder and latent indexing.

Our main contributions are summarized as follows:

- We introduce **De-Absorption Reparameterization (DAR)**, the first method to mitigate feature absorption in a trained SAE without retraining and without altering its reconstruction. Given a supplied parent-child relation, DAR reallocates credit for the parent's contribution within the group through a coupled code-decoder transformation.
- We show that reconstruction is invariant to the DAR transfer coefficients and therefore cannot determine the repair by itself. We estimate these coefficients using sign-constrained least squares, which preserves latent non-negativity and guarantees non-decreasing parent activations.
- We evaluate DAR across multiple model families and SAE architectures. DAR partially restores previously absorbed parent activations while preserving reconstruction to numeri-

cal precision, improves intended-latent discrimination on held-out absorbed examples, and produces substantially smaller changes on matched non-absorbing controls.

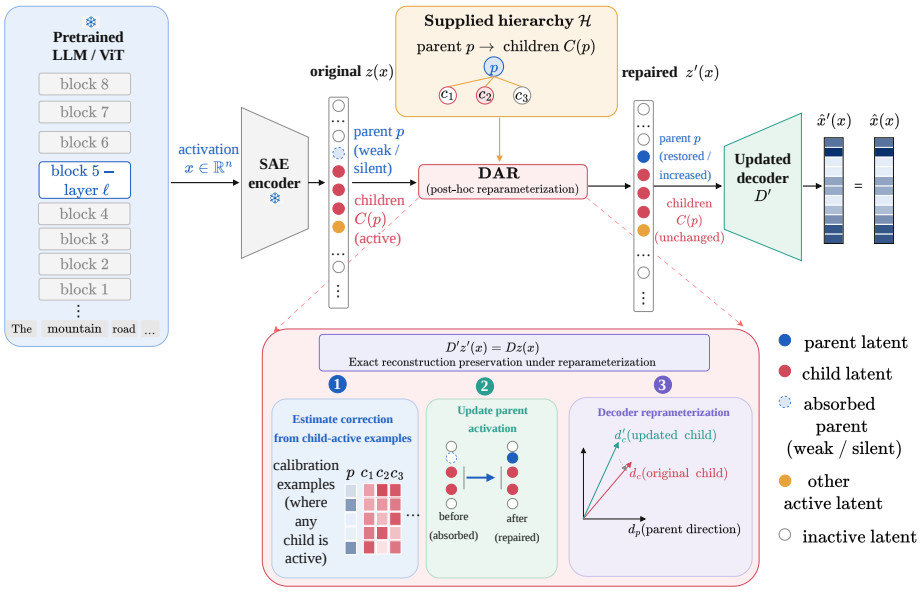


Figure 1: Overview of De-Absorption Reparameterization (DAR). Given a pretrained SAE and a supplied parent–child hierarchy $\mathcal{H}$, DAR estimates transfer coefficients from child-active calibration examples and reparameterizes the parent code and corresponding child decoder atoms. The resulting representation redistributes representation within the supplied parent–child group while preserving the decoded SAE representation exactly, $D'z'(x) = Dz(x)$.

## 2 Preliminaries

**Sparse autoencoders.** Given an activation $x \in \mathbb{R}^n$, an SAE is parameterized by an encoder weight matrix $W_{\text{enc}} \in \mathbb{R}^{n \times m}$, a decoder $D \in \mathbb{R}^{n \times m}$ whose columns $d_1, \ldots, d_m$ (with $d_j := D_{:,j} \in \mathbb{R}^n$) are decoder directions, and a decoder bias $b \in \mathbb{R}^n$, where $m = \varepsilon n$ with expansion factor $\varepsilon \geq 1$. The forward pass produces a sparse non-negative code $z(x)$ and reconstruction $\widehat{x}$:

$$z(x) = E(x) := \sigma\big(W_{\text{enc}}^\top (x - b)\big) \in \mathbb{R}^m_{\geq 0}, \qquad \widehat{x} = Dz(x) + b, \tag{1}$$

where $\sigma$ denotes the architecture-specific sparsification mechanism and $E$ the encoder map.

A common SAE objective is

$$\mathcal{L}_{\text{SAE}}(x) = \underbrace{\|x - \widehat{x}\|_2^2}_{\mathcal{L}_{\text{rec}}} + \lambda \underbrace{S(z(x))}_{\mathcal{L}_{\text{sparse}}}, \tag{2}$$

where $S(\cdot)$ denotes a sparsity penalty and $\lambda$ controls the reconstruction–sparsity trade-off; in hard-sparsity variants, sparsity is instead imposed by the encoder's selection rule.

For $p \in \{1, \ldots, m\}$, let $z_p(x)$ denote the $p^{\text{th}}$ latent activation. Latent $p$ *fires* when $z_p(x) > 0$. JumpReLU (Rajamanoharan et al., 2024) uses a learned latent-specific threshold, whereas TopK and BatchTopK (Makhzani & Frey, 2013; Gao et al., 2025) use hard selection rules.

**Hierarchical features.** Features $f_1$ and $f_2$ form a hierarchy, with $f_1$ as the *parent* and $f_2$ as the *child*, if $f_2 \Rightarrow f_1$. For example, "starts with S" is a parent of "`short`", while "animal" is a parent of "dog". Such relations may be nested.

We use the same terminology for supplied relations between SAE latents. A *parent latent* $p$ corresponds to a parent concept, while child latents $c_1, \ldots, c_k$ correspond to its more specific descendants. A group is $(p, C)$, where $C = \{c_1, \ldots, c_k\}$ and $p \notin C$, written $p \to C$; a collection of

groups forms a hierarchy $\mathcal{H}$. For such a relation, the parent latent is expected to remain active when the corresponding child concept is present. As SAE width increases, broader features may split into more specific features, a phenomenon known as *feature splitting* (Bricken et al., 2023).

**Feature absorption.** For a supplied parent–child relation, *feature absorption* occurs when, on an input $x$ for which a child concept and hence its parent concept are present, the child latent fires while the parent latent remains inactive.

## 3 METHODOLOGY

We propose **De-Absorption Reparameterization** (DAR), a post-hoc reparameterization method for repairing absorbed parent latents in a trained SAE. DAR redistributes representation within supplied parent–child groups through a coupled transformation of the latent code and decoder, while preserving the original SAE reconstruction exactly.

Consider a pretrained SAE $(E, D, b)$ with fixed encoder $E : \mathbb{R}^n \to \mathbb{R}^m_{\geq 0}$, decoder $D = [d_1, \ldots, d_m] \in \mathbb{R}^{n \times m}$, and decoder bias $b \in \mathbb{R}^n$. For an input activation $x \in \mathbb{R}^n$, the SAE produces

$$z = E(x) \in \mathbb{R}^m_{\geq 0}, \qquad \widehat{x} = Dz + b \in \mathbb{R}^n. \tag{3}$$

Let $\{1, \ldots, m\}$ index the SAE latents. DAR is given a supplied hierarchy

$$\mathcal{H} = \{(p, C_p)\}, \qquad p \in \{1, \ldots, m\}, \qquad C_p = \{c_1, \ldots, c_k\} \subseteq \{1, \ldots, m\} \setminus \{p\},$$

where $p$ denotes a parent and $C_p$ its set of $k$ children. For an absorbed input, one or more child latents fire while the corresponding parent latent remains inactive. Our goal is to construct a repaired code $z' \in \mathbb{R}^m_{\geq 0}$ and decoder $D' \in \mathbb{R}^{n \times m}$ such that the parent activation is recovered while the decoded representation remains unchanged. In particular, we seek a transformation satisfying

$$z'_p \geq z_p, \qquad D'z' = Dz, \tag{4}$$

while keeping the pretrained encoder fixed.

**Parent–child reparameterization.** Consider a single supplied parent latent $p$ and child latent $c$. Let $\beta \in \mathbb{R}_{\leq 0}$ be a scalar transfer coefficient. DAR modifies the parent activation and child decoder direction as

$$z'_p = z_p - \beta z_c, \qquad d'_c = d_c + \beta d_p, \tag{5}$$

while leaving $z_c$, $d_p$, and all other latent activations and decoder directions unchanged.

The parent-code and child-decoder updates are coupled so that their effects cancel exactly in the reconstruction. Thus, DAR transfers a share of the representation carried through the supplied child latent to the parent coordinate while preserving the decoded representation exactly.

**Extension to multiple children.** For a group $(p, C_p)$ with $C_p = \{c_1, \ldots, c_k\}$, collect the child activations into

$$y(x) = z_p(x) \in \mathbb{R}_{\geq 0}, \qquad u(x) = \big(z_{c_1}(x), \ldots, z_{c_k}(x)\big)^\top \in \mathbb{R}^k_{\geq 0}, \tag{6}$$

where $y(x)$ denotes the parent activation. The scalar transfer coefficient of the single-child case is generalized to a coefficient vector $\beta \in \mathbb{R}^k_{\leq 0}$. DAR updates the parent activation as

$$z'_p(x) = z_p(x) - u(x)^\top \beta, \qquad z'_j(x) = z_j(x), \quad j \neq p, \tag{7}$$

and modifies each child decoder direction according to

$$d'_{c_j} = d_{c_j} + \beta_j d_p, \qquad j = 1, \ldots, k. \tag{8}$$

The parent decoder direction, all decoder directions outside $C_p$, and the decoder bias remain unchanged.

**Estimating the transfer coefficients.** Let $\mathcal{D}_{\text{fit}} = \{x_i\}_{i=1}^{N_{\text{fit}}}$ denote a calibration set of $N_{\text{fit}}$ model activations used to estimate the DAR transfer coefficients. For each group $(p, C_p)$, we retain only those calibration examples on which at least one child latent fires. Their indices form the set

$$B_p = \{i \in \{1, \dots, N_{\text{fit}}\} : \exists\, c \in C_p \text{ such that } z_c(x_i) > 0\}\,. \tag{9}$$

Let $N = |B_p|$ denote the number of such child-active calibration examples.

Using these $N$ examples, we collect the child activations into $U \in \mathbb{R}_{\geq 0}^{N \times k}$ and the corresponding parent activations into $\mathbf{y} \in \mathbb{R}_{\geq 0}^{N}$. Specifically, each row of $U$ contains the $k$ child activations $u(x_i)^\top$, while the corresponding entry of $\mathbf{y}$ is $z_p(x_i)$, for $i \in B_p$. To remove the mean activation levels, we define the centering matrix

$$H_N = I_N - \frac{1}{N}\mathbf{1}\mathbf{1}^\top \in \mathbb{R}^{N \times N}, \tag{10}$$

where $I_N$ is the $N \times N$ identity matrix and $\mathbf{1} \in \mathbb{R}^N$ is the all-ones vector. We then form

$$\widetilde{U} = H_N U, \qquad \widetilde{\mathbf{y}} = H_N \mathbf{y},$$

and estimate the transfer coefficients by solving

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^k,\, \beta \leq 0}{\arg\min} \left\| \widetilde{\mathbf{y}} - \widetilde{U}\beta \right\|_2^2, \tag{11}$$

where the inequality $\beta \leq 0$ is applied elementwise. Let

$$\widehat{\beta} = (\widehat{\beta}_1, \dots, \widehat{\beta}_k)^\top \in \mathbb{R}_{\leq 0}^k,$$

where $\widehat{\beta}_j$ is the estimated transfer coefficient associated with child latent $c_j$, for $j \in \{1, \dots, k\}$.

A fitted coefficient $\widehat{\beta}_j < 0$ assigns child $c_j$ a positive parent-correction weight $-\widehat{\beta}_j$. Accordingly, $-\widehat{\beta}_j$ determines the parent-activation correction applied per unit activation of child $c_j$, while $\widehat{\beta}_j = 0$ assigns no transfer from that child.

The non-positivity constraint enforces a one-sided correction. Since $u(x) \geq 0$ and $\widehat{\beta} \leq 0$ elementwise, $-u(x)^\top \widehat{\beta} \geq 0$, and therefore $z'_p(x) \geq z_p(x)$ for every input. DAR can thus increase the parent activation but never suppress it.

**Correction strength.** A correction strength $\gamma \in [0, 1]$ is introduced by replacing $\widehat{\beta}$ with $\gamma\widehat{\beta}$. Thus,

$$z'_p(x; \gamma) = z_p(x) - \gamma\, u(x)^\top \widehat{\beta}, \tag{12}$$

and

$$d'_{c_j}(\gamma) = d_{c_j} + \gamma \widehat{\beta}_j d_p. \tag{13}$$

At $\gamma = 0$, DAR reduces to the original SAE, while $\gamma = 1$ applies the full estimated correction. The paired transformation preserves the decoded representation for every $\gamma \in [0, 1]$.

## 4 Theoretical Analysis

### 4.1 Exact reparameterization, admissibility, and locality

We analyze the structural properties of the DAR reparameterization and its coefficient estimator. In particular, we establish exact reconstruction preservation, admissibility and locality of the paired code–decoder transformation, and characterize the solution of the sign-constrained least-squares estimator.

**Proposition 1.** *For a supplied group $(p, C_p)$, let $(z', D')$ be obtained from Eqs. equation 7–equation 8. Then:*

1. *for every $\beta \in \mathbb{R}^k$: $b + D'z' = b + Dz$;*
2. *if $\beta \leq 0$ elementwise, then $z'_p \geq z_p \geq 0, z' \in \mathbb{R}_{\geq 0}^m$;*

3. *only the parent activation and the supplied child decoder directions are modified, and* $z' = z$ *whenever* $u = 0$.

*Proof.* Using Eqs. equation 7–equation 8, the contribution of the supplied group after the transformation is: $z'_p d_p + \sum_{j=1}^{k} z_{c_j} d'_{c_j} = (z_p - u^\top \beta) d_p + \sum_{j=1}^{k} z_{c_j} (d_{c_j} + \beta_j d_p) = z_p d_p - (\sum_{j=1}^{k} z_{c_j} \beta_j) d_p + \sum_{j=1}^{k} z_{c_j} d_{c_j} + (\sum_{j=1}^{k} z_{c_j} \beta_j) d_p = z_p d_p + \sum_{j=1}^{k} z_{c_j} d_{c_j}$, since $u^\top \beta = \sum_j z_{c_j} \beta_j$. All contributions outside the group and the decoder bias are unchanged.

If $\beta \leq 0$, then $u \geq 0$ implies $u^\top \beta \leq 0$, and therefore $z'_p = z_p - u^\top \beta \geq z_p \geq 0$.

All remaining code coordinates are unchanged and hence remain non-negative. The locality statement follows directly from Eqs. equation 7–equation 8; when $u = 0$, the parent activation is unchanged as well. □

Proposition 1 shows that DAR can reallocate representation within a supplied parent–child group without changing the decoded SAE representation. Under the sign constraint, this reallocation is admissible: the repaired code remains non-negative, the parent activation can only increase, and all non-parent code coordinates remain unchanged. Thus, changes in the repaired parent representation can be studied independently of changes in reconstruction. A detailed theoretical analysis of DAR is provided in the appendix A.1.

## 5 Experiments

**Models and metrics** Our main language experiments use the pretrained Gemma 2 2B model (Team et al., 2024) with Gemma Scope residual-stream SAEs (Lieberum et al., 2024). We use JumpReLU SAE with 16,384 latent features and an average of approximately 82 active features per input. Both the language model and SAE encoder remain frozen; DAR operates post-hoc by reparameterizing the SAE latent codes and decoder weights.

To evaluate generalization across model and SAE families, we also consider GPT-2 Small (Radford et al., 2019) with the OpenAI v5 SAE (Gao et al., 2025), using the layer-6 residual-stream SAE with 32,768 latent features and $k = 32$. For vision experiments, we use the vision tower of CLIP ViT-L/14-336 (Radford et al., 2021) with a locally trained layer-22 BatchTopK SAE (Bussmann et al., 2024), containing 65,536 latent features with $k = 20$. To evaluate DAR across different types of semantic hierarchies, we obtain these relations from three sources: spelling-based feature absorption (Chanin et al., 2025), visual object taxonomies (Russakovsky et al., 2015), and lexical semantic hierarchies (Miller, 1995). Across all three settings, the supplied parent-child relations are fixed before DAR evaluation. DAR uses these relations to estimate transfer coefficients ($\beta$) without adding, removing, or selecting relations based on the resulting evaluation performance. We evaluate DAR using multiple criteria: parent recovery, leakage, absorption reduction, Spearman's correlation ($s_{\mathrm{cc}}$) and Reconstruction. Recovery is reported using both macro-averaging across parent groups and example-weighted averaging to account for differences in the number of absorbed examples. We additionally measure the parent-given-child activation rate, denoted by $s_{\mathrm{ai}}$. Detailed evaluation setup details is provided in A.2.5.

### 5.1 De-absorption of features

To investigate DAR's effectiveness to restores parent activations affected by feature absorption while preserving the original SAE reconstruction. All experiments in this subsection use Gemma-2-2B at layer 12 with the Gemma Scope 16K JumpReLU SAE. Parent-child relations are obtained from the first-letter absorption benchmark of (Chanin et al., 2025), which identifies absorption through an ablation-based procedure independent of DAR. We first examine individually certified parent-child pairs from (Chanin et al., 2025), evaluating changes in parent activation and comparing them with matched non-absorber controls. We then evaluate a complete supplied hierarchy to assess parent recovery, absorption reduction, leakage, and representation cost.

**Certified absorption pairs.** We first examine a single parent–child pair on 17 absorption-affected tokens. After applying DAR, the mean parent activation increases from $0.206$ to $4.154$, an absolute

Table 1: Changes in parent activation for certified absorption pairs and matched non-absorber controls.

| Evaluation | Pairs | Median $\Delta z_p$ |
|---|---|---|
| Identified absorber | 25 | $+\mathbf{5.79}$ |
| Absorber, control subset | 17 | $+\mathbf{3.38}$ |
| Matched non-absorber | 17 | 0.00 |

Table 2: Evaluation of DAR on held-out examples using a supplied hierarchy of 24 parent latents and 245 parent–child relationships.

| Quantity | Before → After |
|---|---|
| *Repair* | |
| Recovery on absorbed tokens | $0.000 \rightarrow \mathbf{0.446}$ |
| Signature $s_{\mathrm{cc}}$ | $-0.204 \rightarrow +\mathbf{0.102}$ |
| Rate $s_{\mathrm{ai}}$ | $0.527 \rightarrow \mathbf{0.944}$ |
| Groups improved | $\mathbf{24/24}$ |
| *Cost* | |
| Negative-control firing | $+0.0116$ |
| Active latents per token | $+0.094$ |
| Reconstruction change | $3.72 \times 10^{-17}$ |

increase of $3.948$. The relative reconstruction difference is only $1.16 \times 10^{-17}$, confirming that the reconstruction is preserved to numerical precision.

We then extend the experiment to 25 certified parent–child pairs, repairing each pair independently. As shown in Table 1, parent activation increases for 21 of the 25 pairs, with a median increase of $5.79$ (95% CI $[3.05, 6.86]$). The remaining four pairs receive zero transfer coefficients and remain unchanged. The maximum relative reconstruction difference across all pairs is $3.05 \times 10^{-17}$. To examine whether DAR corrects feature absorption rather than simply increasing parent activation, we conduct a control experiment. For 17 absorbing parent–child pairs, we select non-absorbing children with similar activation frequencies and apply DAR separately to both sets. The absorbing pairs show a median increase of $3.38$ in parent activation, whereas the non-absorbing pairs show a median change of $0.00$. Among the 17 non-absorbing control pairs, only two show an increase in parent activation, while the remaining 15 remain unchanged. These results suggest that DAR primarily responds to the identified absorption relationships.

**On supplied hierarchy.** We further evaluate DAR on a hierarchy of 24 parent latents and their 245 associated child latents, as identified by (Chanin et al., 2025). We apply DAR to all parent–child groups and evaluate its effectiveness on held-out examples using the fixed firing criterion defined in Section A.2.5.

As shown in Table 2, DAR successfully restores parent activation on $44.6\%$ of previously absorbed examples, with improvements across all 24 parent groups. The absorption signature $s_{\mathrm{cc}}$, defined in Eq. equation 29, increases from $-0.204$ to $+0.102$, indicating a shift from a negative to a positive association between parent and child activations. Similarly, the parent-given-child activation rate $s_{\mathrm{ai}}$ increases from $0.527$ to $0.944$, showing that parent latents activate more consistently when their corresponding children are active. These improvements are accompanied by an increase of $0.0116$ in the negative-control firing rate and an average increase of $0.094$ active latents per token. Meanwhile, the relative reconstruction difference remains $3.72 \times 10^{-17}$, confirming that DAR preserves the original reconstruction to numerical precision.

### 5.2 Quantitative Comparison with Post-hoc Baselines

We compare DAR with three post-hoc baselines to investigate whether increasing parent activation alone is sufficient to address feature absorption. All methods are evaluated using the same supplied hierarchy of 24 parent latents and 245 parent–child relationships on the same held-out data.

The first baseline, *constant boost*, adds a fixed amount to the parent activation whenever a child latent fires. The second, *magnitude boost*, increases the parent activation proportionally to the maximum child activation. The third, *recovery-matched boost*, uses a constant boost whose strength is selected to match DAR's recovery on the calibration data. Unlike DAR, these baselines modify only the parent activation without updating the decoder. We compare the methods in terms of parent firing rate, absorption signature, leakage, and reconstruction preservation. The results are reported in Table 3. DAR increases the parent firing rate from $0.085$ to $0.431$ and changes the absorption signature $s_{\mathrm{cc}}$ from $-0.204$ to $+0.102$, while preserving reconstruction to numerical precision ($4.1 \times 10^{-18}$). The

constant and recovery-matched boosts achieve firing rates of $0.887$ and $0.488$, respectively. However, both leave the absorption signature unchanged at $-0.204$, despite increasing parent activation. The magnitude boost increases $s_{\mathrm{cc}}$ to $+0.397$, showing that scaling the correction with child activation can also change the parent–child activation relationship. Nevertheless, all three baselines alter the reconstruction, with relative changes ranging from $7.7 \times 10^{-3}$ to $3.7 \times 10^{-2}$.

The baselines also introduce additional unintended activations. For example, the recovery-matched boost achieves a held-out firing rate of $0.488$ with leakage of $0.0592$, compared with DAR's firing rate of $0.431$ and leakage of $0.0292$. The constant and magnitude boosts achieve higher firing rates of $0.887$ and $0.790$, respectively, with leakage rates of $0.0969$ and $0.0579$. Overall, these results show that increasing parent activation does not necessarily reduce the negative parent–child association or preserve the original reconstruction. Although some activation boosts also improve the absorption signature, they alter the decoded representation. In contrast, DAR improves parent activation and the measured absorption signature while preserving reconstruction through its paired code–decoder transformation. Additional quantitative results are provided in Appendix A.2.

Table 3: Comparison of DAR with post-hoc activation boosting methods on a supplied hierarchy of 24 parent latents and 245 parent–child relationships. Before correction, $s_{\mathrm{cc}} = -0.204$, the parent firing rate is $0.085$, and leakage is $0.0185$.

| Method | Groups repaired | $s_{\mathrm{cc}}$ | Recon. change | Firing rate | Leakage |
|---|---|---|---|---|---|
| **DAR (ours)** | $\mathbf{24/24}$ | $\mathbf{+0.102}$ | $\mathbf{4.1 \times 10^{-18}}$ | $0.431$ | $\mathbf{0.0292}$ |
| Constant boost | $22/24$ | $-0.204$ | $2.0 \times 10^{-2}$ | $0.887$ | $0.0969$ |
| Magnitude boost | $21/24$ | $+0.397$ | $3.7 \times 10^{-2}$ | $0.790$ | $0.0579$ |
| Recovery-matched boost | $11/24$ | $-0.204$ | $7.7 \times 10^{-3}$ | $0.488$ | $0.0592$ |

### 5.3 Qualitative Examples

We illustrate how DAR restores parent activations and changes their influence on model predictions. We use Gemma-2-2B at layer 12 with the Gemma Scope 16K JumpReLU SAE and parent–child relations from the first-letter absorption benchmark of Chanin et al. (2025).

**Restoring parent activation.** Figure 2 illustrates DAR on the token `Statistics`. Before repair, the parent latent #6508, representing "starts with S", has zero activation, while the child latent #16181 carries a component of $+0.2111$ along the parent decoder direction. After applying DAR, the parent activation increases from $0$ to $9.24$, while the corresponding component in the child decoder direction decreases to $0.0014$, leaving only $0.65\%$ of its original magnitude. Additional examples in Figure 4 show similar changes across different parent–child pairs. In each case, DAR increases parent activation while reducing the parent-direction component of the corresponding child decoder. Importantly, these changes preserve the original SAE reconstruction.

**Effect on model predictions.** Figure 3 examines the effect of intervening on three parent latents associated with the letters `F`, `U`, and `C`. Their activations increase from $0.00$, $0.00$, and $3.53$ to $75.30$, $18.56$, and $21.80$, respectively. We examine the effect of scaling each parent activation before decoding. A multiplier of $s = 0$ suppresses the parent, $s = 1$ leaves it unchanged, and $s = 2$ amplifies its contribution. For each intervention, we record the model's five most probable next tokens. At $s = 1$, predictions remain unchanged before and after DAR, consistent with exact reconstruction preservation. At $s = 0$, suppressing the repaired parent removes the intended first-letter continuation from the top predictions in all three examples. In contrast, suppressing the original parent leaves the intended continuation among the top predictions. At $s = 2$, the intended continuation remains present, although the other predictions may change. These examples illustrate that DAR not only redistributes parent-related information without changing the reconstruction but also makes the repaired parent a more effective target for causal interventions.

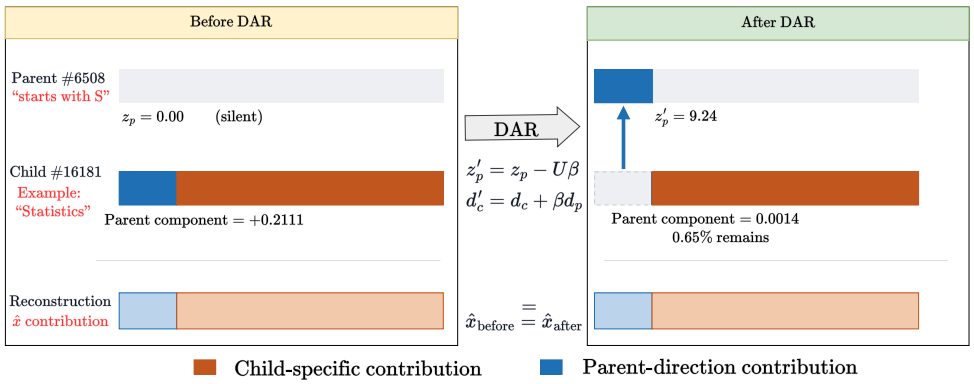


Figure 2: DAR restores parent activation for `Statistics` by transferring parent-related information from the child decoder while preserving the original reconstruction.

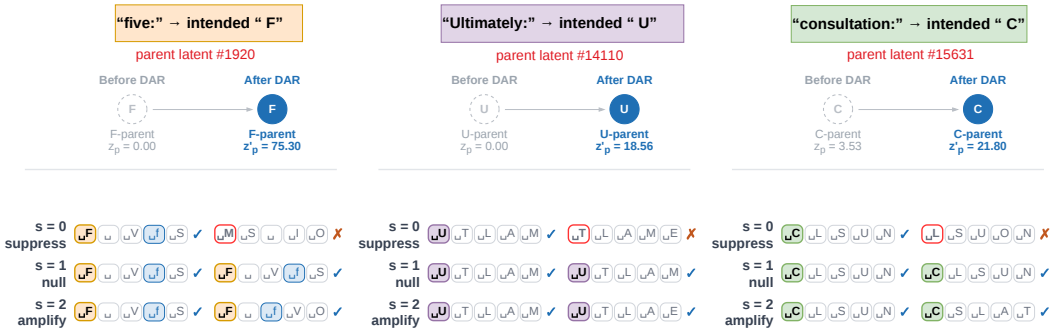


Figure 3: Effect of steering parent latents before and after DAR. Scaling the parent activation by $s \in \{0, 1, 2\}$ illustrates how repair changes the parent's influence on model predictions.

## 6 Conclusion

We introduced **De-Absorption Reparameterization (DAR)**, a post-hoc method for repairing supplied parent–child absorption relations in pretrained sparse autoencoders. DAR reparameterizes the latent code and decoder within a parent–child group so that parent activation can be increased while the decoded SAE representation is preserved exactly. Its sign-constrained coefficient estimation preserves code non-negativity and guarantees non-decreasing parent activations.

We empirically validated DAR across multiple SAE architectures and model families, showing that it can partially recover absorbed parent activations, improve the discrimination of the repaired parent coordinates, and remain selective relative to matched non-absorbing controls while preserving reconstruction to numerical precision. These results demonstrate that feature absorption can be addressed post hoc for supplied hierarchical relations without retraining the SAE encoder. More broadly, DAR shows that reconstruction-equivalent SAE representations can differ in the usefulness of their latent coordinates, motivating post-hoc methods for improving existing sparse representations without replacing the underlying SAE.

## AI Use Statement

Generative AI tools were used solely for language editing, grammatical correction, and improving the clarity and readability of the manuscript. All AI-assisted revisions were carefully reviewed by the authors, who take full responsibility for the final content.

## REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikołaj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a Visual Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2204.14198.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick 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. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Bart Bussmann, Patrick Leask, and Neel Nanda. BatchTopK Sparse Autoencoders. *arXiv preprint arXiv:2412.06410*, 2024. doi: 10.48550/arXiv.2412.06410. URL https://arxiv.org/abs/2412.06410.

Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning Multi-Level Features with Matryoshka Sparse Autoencoders. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 6077–6101. PMLR, 2025. URL https://proceedings.mlr.press/v267/bussmann25a.html.

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. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2755. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/764ff7477b8e24dbe01531f6791e8bdf-Abstract-Conference.html.

Seyed Arshan Dalili and Mehrdad Mahdavi. Subspace-Aware Sparse Autoencoders for Effective Mechanistic Interpretability. *arXiv preprint arXiv:2606.06333*, 2026. doi: 10.48550/arXiv.2606.06333. URL https://arxiv.org/abs/2606.06333.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255. IEEE, 2009. doi: 10.1109/CVPR.2009.5206848.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and Evaluating Sparse Autoencoders. In *The Thirteenth International Conference on Learning Representations (ICLR)*, volume 2025, pp. 26721–26754, 2025. URL https://arxiv.org/abs/2406.04093.

Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 624–635, 2021. doi: 10.1145/3442188.3445923.

Haoran Jin, Xiting Wang, Shijie Ren, Hong Xie, and Defu Lian. $C^2R$: Cross-Sample Consistency Regularization Mitigates Feature Splitting and Absorption in Sparse Autoencoders. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. doi: 10.48550/arXiv.2606.30609. URL https://arxiv.org/abs/2606.30609. arXiv:2606.30609.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36:71683–71702, 2023.

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. *arXiv preprint arXiv:2408.05147*, 2024. doi: 10.48550/arXiv.2408.05147. URL https://arxiv.org/abs/2408.05147.

Alireza Makhzani and Brendan Frey. k-Sparse Autoencoders. *arXiv preprint arXiv:1312.5663*, 2013.

Aniek F Markus, Jan A Kors, and Peter R Rijnbeek. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics*, 113:103655, 2021.

George A. Miller. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11): 39–41, 1995.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. A Semantic Concordance. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*, 1993.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139, pp. 8748–8763. PMLR, 2021. URL https://arxiv.org/abs/2103.00020.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping Ahead: Improving Reconstruction Fidelity with JumpReLU Sparse Autoencoders. *arXiv preprint arXiv:2407.14435*, 2024. doi: 10.48550/arXiv.2407.14435. URL https://arxiv.org/abs/2407.14435.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Charles Spearman. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 15(1):72–101, 1904. doi: 10.2307/1412159.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving Open Language Models at a Practical Size. arXiv preprint arXiv:2408.00118, 2024. URL https://arxiv.org/abs/2408.00118.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

# A Appendix

## A.1 Theoretical Analysis

This appendix develops the theoretical properties of DAR beyond the exact reparameterization established in Proposition 1. We first characterize the sign-constrained coefficient estimator and explain

why reconstruction alone cannot determine the transfer coefficients. We then examine the implications for sparsity, establish population-level properties under explicit assumptions, and show how the guarantees extend to multiple groups. Finally, we characterize the effect of correction strength and the associated representational cost.

**Optimality conditions for the transfer coefficients.** We begin with the coefficient estimator in Eq. equation 11. Its optimality conditions describe the remaining linear association between the repaired parent activation and the supplied child activations on the fitting set.

**Proposition 2** (KKT characterization of the DAR estimator)**.** *Let*

$$\widetilde{r} := \widetilde{y} - \widetilde{U}\widehat{\beta} \tag{14}$$

*be the residual at the solution of Eq. equation 11. Then, for every supplied child $j$, the following conditions hold:*

$$\widehat{\beta}_j < 0 \Longrightarrow \quad \widetilde{U}_{:,j}^{\top}\widetilde{r} = 0, \qquad \widehat{\beta}_j = 0 \Longrightarrow \quad \widetilde{U}_{:,j}^{\top}\widetilde{r} \geq 0.$$

*Equivalently, on the fitting rows, the repaired parent activation is empirically uncorrelated with every child receiving a nonzero correction and has non-negative empirical covariance with every child left at the constraint boundary.*

*Proof.* Equation equation 11 is equivalent to

$$\min_{\beta \leq 0} \frac{1}{2}\|\widetilde{y} - \widetilde{U}\beta\|_2^2.$$

Introduce Lagrange multipliers $\lambda \geq 0$ for the constraints $\beta \leq 0$. The KKT conditions at $\widehat{\beta}$ are

$$-\widetilde{U}^{\top}\left(\widetilde{y} - \widetilde{U}\widehat{\beta}\right) + \lambda = 0, \qquad \lambda_j\widehat{\beta}_j = 0, \qquad \lambda_j \geq 0.$$

By Eq. equation 14, $\lambda = \widetilde{U}^{\top}\widetilde{r} \geq 0$. If $\widehat{\beta}_j < 0$, complementary slackness gives $\lambda_j = 0$, and therefore $\widetilde{U}_{:,j}^{\top}\widetilde{r} = 0$. If $\widehat{\beta}_j = 0$, dual feasibility gives $\widetilde{U}_{:,j}^{\top}\widetilde{r} = \lambda_j \geq 0$. □

Proposition 2 characterizes the fitted correction through first-order optimality conditions. A child receiving a nonzero correction has zero residual linear association with the repaired parent on the fitting set; for a child at the constraint boundary, no negative first-order association remains that can be reduced by an admissible coefficient change. These are properties of the fitted activation model, rather than guarantees about the true semantic content of a latent.

**Reconstruction non-identifiability.** The optimality conditions select coefficients using activation statistics. Such an additional criterion is necessary because all members of the paired DAR family have the same reconstruction loss.

**Corollary 1** (Reconstruction non-identifiability)**.** *For a fixed SAE code and supplied group, the reconstruction loss is constant over all coefficient vectors applied through the paired DAR transformation:*

$$\mathcal{L}_{\mathrm{rec}}(z', D') = \mathcal{L}_{\mathrm{rec}}(z, D). \tag{15}$$

*Consequently, reconstruction alone cannot determine the transfer coefficients or distinguish different parent-child allocations within this family.*

*Proof.* The claim follows directly from the reconstruction identity in Proposition 1. □

Corollary 1 explains why the reconstruction objective cannot select a particular allocation of the parent contribution. DAR instead uses the activation-based criterion in Eq. equation 11 to choose the transfer coefficients. Nevertheless, an unchanged reconstruction does not imply that other properties of the sparse representation remain unchanged.

**Interaction with sparsity.** In particular, increasing the parent activation can increase a sparsity term even though the decoded representation is preserved. The following result makes this trade-off explicit.

**Corollary 2** (Tension with monotone sparsity penalties)**.** *Suppose the sparsity functional $S(z)$ is coordinatewise non-decreasing on $\mathbb{R}^m_{\geq 0}$. Then, for every admissible DAR correction $\beta \leq 0$,*

$$S(z') \geq S(z), \tag{16}$$

*while $\mathcal{L}_{\mathrm{rec}}$ remains unchanged. Hence the uncorrected member $\beta = 0$ weakly minimizes such a sparsity term within the DAR family.*

*Proof.* By Proposition 1, $z' \geq z$ coordinatewise. Coordinatewise monotonicity of $S$ therefore gives $S(z') \geq S(z)$. Reconstruction invariance follows from Proposition 1(1). □

Thus, for coordinatewise non-decreasing sparsity penalties, the DAR correction cannot reduce the sparsity term. This includes the $\ell_1$ mass of a non-negative code and thresholded support size. The result does not imply that every SAE training objective causes absorption; rather, it identifies a possible tension between parent restoration and sparsity when reconstruction is held fixed.

**Population identification of the transfer.** The preceding results establish what DAR preserves and what it may change. We next examine when the activation-based estimator recovers the transfer coefficients of an underlying additive absorption model.

**Proposition 3** (Population identification under additive absorption)**.** *Assume that, conditional on $B_p$, the parent activation satisfies*

$$y = \mu - u^\top \alpha^\star + \epsilon, \qquad \alpha^\star \in \mathbb{R}^k_{\geq 0}, \qquad \mathbb{E}[\epsilon \mid u, B_p] = 0, \tag{17}$$

*and that*

$$\Sigma_u := \mathrm{Cov}(u \mid B_p) \succ 0. \tag{18}$$

*Then*

$$\beta^\star = -\alpha^\star \tag{19}$$

*is the unique minimizer over $\beta \leq 0$ of the population counterpart of Eq. equation 11.*

*Proof.* After centering on $B_p$,

$$\widetilde{y} = -\widetilde{u}^\top \alpha^\star + \epsilon = \widetilde{u}^\top \beta^\star + \epsilon, \qquad \beta^\star = -\alpha^\star \leq 0.$$

For any feasible $\beta$,

$$\widetilde{y} - \widetilde{u}^\top \beta = \widetilde{u}^\top (\beta^\star - \beta) + \epsilon.$$

The conditional-mean assumption gives $\mathbb{E}[\widetilde{u}\,\epsilon \mid B_p] = 0$, and therefore

$$Q(\beta) = (\beta - \beta^\star)^\top \Sigma_u (\beta - \beta^\star) + \mathbb{E}[\epsilon^2 \mid B_p]. \tag{20}$$

Since $\Sigma_u \succ 0$, the quadratic term is strictly positive for every $\beta \neq \beta^\star$. Because $\beta^\star \leq 0$ is feasible, it is the unique constrained minimizer. □

Proposition 3 provides a population-level interpretation of the estimator. Under the stated additive-linear model, conditional-mean assumption, and positive-definite covariance condition, the constrained criterion identifies the corresponding transfer coefficients. These assumptions are substantive: the proposition does not guarantee the same identification when the activation relationship is misspecified or the supplied child activations are not sufficiently informative.

**Why conditioning uses child activity.** The fitting set $B_p$ is defined using the supplied child activations rather than the parent response. Consequently, selection is predictor-measurable. Under Eq. equation 17, conditioning on this event is compatible with the stated conditional-mean assumption. Selecting examples according to the parent response would generally require an additional assumption to retain the same exogeneity interpretation.

**Population specificity.** The identification result also has a useful limiting case: when the parent and supplied child activations are conditionally unrelated, the population criterion selects no correction.

**Corollary 3** (Population specificity)**.** *Suppose*

$$y \perp u \mid B_p \tag{21}$$

*and* $\mathrm{Cov}(u \mid B_p) \succ 0$*. Then the population minimizer of the DAR regression is*

$$\beta^\star = 0. \tag{22}$$

*Hence DAR reduces to the identity transformation at the population level for a conditionally unrelated supplied group.*

*Proof.* Conditional independence implies

$$\mathbb{E}[y \mid u, B_p] = \mathbb{E}[y \mid B_p].$$

Thus Eq. equation 17 holds with $\alpha^\star = 0$ and

$$\epsilon = y - \mathbb{E}[y \mid B_p],$$

for which $\mathbb{E}[\epsilon \mid u, B_p] = 0$. Proposition 3 therefore gives $\beta^\star = 0$. □

Corollary 3 establishes that unrelated supplied groups receive no correction at the population level under its assumptions. This specificity property does not imply that finite-sample estimates are exactly zero, since sampling variation can produce nonzero fitted coefficients.

**Composition over multiple groups.** The preceding statements concern individual supplied groups. In practice, DAR may be applied to several parent–child groups within a hierarchy. The next proposition shows that the reconstruction and admissibility guarantees are maintained when these transformations are composed sequentially.

**Proposition 4** (Closure under sequential composition)**.** *Let* $(z^{(g)}, D^{(g)})$ *denote the representation after the* $g$*-th stored DAR transformation, with* $(z^{(0)}, D^{(0)}) = (z, D)$*. Then, for any finite sequence of groups,*

$$b + D^{(G)} z^{(G)} = b + Dz. \tag{23}$$

*If every stored coefficient vector is elementwise non-positive, then*

$$z^{(G)} \geq z^{(G-1)} \geq \cdots \geq z^{(0)} \geq 0 \tag{24}$$

*coordinatewise.*

*Proof.* By Proposition 1, each individual transformation satisfies

$$D^{(g)} z^{(g)} = D^{(g-1)} z^{(g-1)}.$$

Applying this identity recursively proves Eq. equation 23.

For admissibility, assume $z^{(g-1)} \geq 0$. The child activations at step $g$ are therefore non-negative and the stored coefficient vector is non-positive. Proposition 1(2) implies

$$z^{(g)} \geq z^{(g-1)} \geq 0.$$

Induction over $g$ proves Eq. equation 24. □

Thus, the guarantees of exact reconstruction preservation and non-negative repaired codes extend from individual groups to any finite sequence of stored DAR transformations. This does not imply order invariance: overlapping groups need not commute, and the processing order specified in Section A.2.1 remains part of the method.

**Monotonicity in correction strength.** Once the coefficients have been estimated, the correction strength controls how much activation is reassigned to the parent. The sign constraint makes this dependence monotone.

Define

$$q(x) := -u(x)^\top \widehat{\beta} \geq 0,$$

so that the strength-controlled parent activation is

$$z'_p(x;\gamma) = z_p(x) + \gamma\, q(x).$$

**Proposition 5** (Monotonicity in correction strength)**.** *For every input $x$, $z'_p(x;\gamma)$ is non-decreasing in $\gamma$. Moreover, for any fixed threshold $\theta_p$ independent of $\gamma$,*

$$\mathbf{1}\{z'_p(x;\gamma) > \theta_p\} \tag{25}$$

*is non-decreasing in $\gamma$.*

*Proof.* For $0 \leq \gamma_1 \leq \gamma_2 \leq 1$,

$$z'_p(x;\gamma_2) - z'_p(x;\gamma_1) = (\gamma_2 - \gamma_1)q(x) \geq 0.$$

Threshold monotonicity follows immediately. □

Proposition 5 guarantees that increasing the correction strength cannot decrease the repaired parent activation or its firing indicator under a fixed threshold. The increase need not be strict: if $u(x)^\top \widehat{\beta} = 0$, the parent activation remains unchanged for all strengths. This monotonicity concerns activation and firing, not a guarantee that semantic specificity improves at every strength.

**Exact representational cost.** Although reconstruction remains unchanged at every correction strength, restoring a suppressed parent can change the sparse code and the affected child decoder directions. The following proposition quantifies these changes for a single group.

For tolerance $\tau \geq 0$, define

$$\|z\|_{0,\tau} := \sum_{j=1}^{m} \mathbf{1}\{z_j > \tau\}.$$

**Proposition 6** (Exact representation cost)**.** *For a single DAR group and any $\gamma \in [0,1]$,*

$$\|z'(\gamma)\|_{0,\tau} - \|z\|_{0,\tau} = \mathbf{1}\left\{z_p \leq \tau,\ z_p + \gamma\, q(x) > \tau\right\}, \tag{26}$$

$$\|z'(\gamma)\|_1 - \|z\|_1 = \gamma\, q(x) \geq 0, \tag{27}$$

$$\|d'_{c_j}(\gamma)\|_2^2 = \|d_{c_j}\|_2^2 + 2\gamma\widehat{\beta}_j d_p^\top d_{c_j} + \gamma^2\widehat{\beta}_j^2\|d_p\|_2^2. \tag{28}$$

*Proof.* Only the parent code coordinate changes, from $z_p$ to $z_p + \gamma\, q(x)$, and this change is non-negative. The thresholded support therefore increases only when the parent was previously at or below the tolerance and crosses it after repair, giving Eq. equation 26.

Since the code is non-negative,

$$\|z'(\gamma)\|_1 - \|z\|_1 = z'_p - z_p = \gamma\, q(x),$$

which proves Eq. equation 27. Finally, substituting

$$d'_{c_j}(\gamma) = d_{c_j} + \gamma\widehat{\beta}_j d_p$$

and expanding the squared Euclidean norm gives Eq. equation 28. □

Proposition 6 makes the cost of parent restoration explicit: parent support can increase when the repaired activation crosses the chosen tolerance, the non-negative code's $\ell_1$ mass increases by the transferred amount, and the supplied child decoder norms change according to Eq. equation 28. These changes are compatible with exact reconstruction preservation; they describe properties of the representation that reconstruction fidelity alone does not capture.

**Algorithm 1** De-Absorption Reparameterization for one supplied group

**Require:** Frozen SAE $(E, D, b)$; supplied group $(p, C_p)$; calibration split $\mathcal{D}_{\text{fit}}$; tolerance $\tau$; correction strength $\gamma$
**Ensure:** Repaired decoder $D'$ and code map $z \mapsto z'$

**Calibration and coefficient estimation**

$Z \leftarrow \{E(x_i)\}_{i \in \mathcal{D}_{\text{fit}}}$ ▷ Frozen encoder
$B_p \leftarrow \{i : \exists\, c \in C_p,\ Z_{ic} > \tau\}$ ▷ Eq. equation 9
$C_p \leftarrow \{c \in C_p : |\{i \in B_p : Z_{ic} > \tau\}| \geq 2\}$
**if** $|B_p| < 2$ **or** $C_p = \varnothing$ **then**
    **return** Identity transformation
**end if**
$B_p \leftarrow \{i : \exists\, c \in C_p,\ Z_{ic} > \tau\}$
$N \leftarrow |B_p|$
**if** $N < 2$ **then**
    **return** Identity transformation
**end if**
$\mathbf{y} \leftarrow Z_{B_p,p}$; $U \leftarrow Z_{B_p,C_p}$
$\widetilde{\mathbf{y}} \leftarrow H_N \mathbf{y}$; $\widetilde{U} \leftarrow H_N U$
$\widehat{\beta} \leftarrow \arg\min_{\beta \leq 0} \left\| \widetilde{\mathbf{y}} - \widetilde{U}\beta \right\|_2^2$ ▷ Eq. equation 11

**Decoder reparameterization**

$D' \leftarrow D$
**for** $j = 1, \ldots, |C_p|$ **do**
    $d'_{c_j} \leftarrow d_{c_j} + s\widehat{\beta}_j d_p$
**end for**

**Inference**

Define the code map $z \mapsto z'$ as follows:
    $z \leftarrow E(x)$
    $z'_p \leftarrow z_p - \gamma\, u(x)^\top \widehat{\beta}$
    $z'_j \leftarrow z_j$ for all $j \neq p$
**return** $(D',\ z \mapsto z')$

## A.2 Experiments

### A.2.1 Inference and Hierarchical Composition

**Inference.** For a new input $x$, we first compute $z = E(x)$ using the original frozen encoder. The stored DAR transformation is then applied to obtain $z'$, which is decoded using $D'$. No optimization or coefficient estimation is performed at inference time.

**Multiple hierarchy groups.** For a hierarchy containing several parent-child groups, DAR applies the stored transformations sequentially. When the supplied hierarchy forms a directed acyclic graph, groups are processed from the deepest level toward the root, so that a latent acting as a child is repaired before it is used as a parent. Each transformation preserves reconstruction, and therefore their composition does so as well.

### A.2.2 Measuring the Repair

Since DAR preserves the reconstruction by construction, reconstruction error does not measure the effectiveness of the repair; instead, it serves as a correctness check. We therefore evaluate the repaired representation along four complementary axes on held-out data.

**Parent recovery.** Recovery measures the fraction of held-out absorbed examples for which the repaired parent activation exceeds the original SAE's native firing cutoff. The cutoff is determined from the original SAE and kept fixed across all correction strengths. This measures whether DAR restores the parent strongly enough to be considered active under the original SAE's firing criterion.

**Leakage.** Leakage measures the firing rate of the repaired parent on negative-control examples for which the parent concept is absent according to ground truth. Because DAR can only increase the parent activation, recovery is always reported together with leakage to distinguish useful restoration from indiscriminate activation increases.

**Absorption reduction.** To determine whether DAR removes the dependence underlying absorption rather than merely increasing the parent magnitude, we compare the association between parent and child activations before and after repair. Successful de-absorption should weaken the negative association characteristic of parent suppression when a child is active.

**Representation cost.** We additionally report the relative reconstruction change, the change in the number of active latents per example, and the change in the $\ell_1$ mass of the latent code. The reconstruction change is expected to remain at numerical precision, whereas the latter two quantities characterize the representational cost of restoring the parent latent.

### A.2.3 Scope of DAR

DAR is a post-hoc reparameterization method that operates on supplied parent-child relations. It preserves the decoded reconstruction exactly, maintains non-negative parent activations, and never decreases the parent activation. The modification is confined to the specified hierarchy groups and requires neither encoder retraining nor per-example optimization at inference.

DAR does not discover parent-child relations and does not preserve the original sparsity penalty, decoder norms, or complete SAE training objective. Moreover, the repaired code need not itself be an output of the original encoder's native gating rule. DAR should therefore be viewed as a post-hoc semantic re-attribution procedure rather than an alternative SAE training objective or an absorption-detection method.

### A.2.4 Supplied hierarchies.

DAR operates on predefined parent-child relations rather than discovering them from SAE activations. To evaluate the method across different types of semantic hierarchies, we obtain these relations from three sources: spelling-based feature absorption, visual object taxonomies, and lexical semantic hierarchies.

*(i) Spelling.* We use the first-letter absorption benchmark of (Chanin et al., 2025). Parent concepts represent the initial letters of tokens, while child concepts correspond to specific tokens associated with those letters. The corresponding child features are identified using the benchmark's ablation-based absorption procedure.

*(ii) Visual taxonomy.* We construct visual concept hierarchies using the ImageNet synset taxonomy (Deng et al., 2009; Russakovsky et al., 2015), obtained from the official ILSVRC2012 development kit.[1] Parent concepts correspond to higher-level internal synsets representing broad object categories, while child concepts correspond to more specific descendant synsets. Specifically, each child represents a subset of the object classes covered by its parent, establishing a hierarchical relationship between general and specific visual concepts.

*(iii) Lexical semantics.* We construct lexical hierarchies using WordNet noun hypernym relations (Miller, 1995) restricted to concepts occurring in SemCor (Miller et al., 1993). Hypernyms represent broader parent concepts, while their hyponyms represent more specific child concepts. These semantic relations are defined independently of the SAE and subsequently mapped to the corresponding SAE latents.

Across all three settings, the supplied parent-child relations are fixed before DAR evaluation. DAR uses these relations to estimate transfer coefficients without adding, removing, or selecting relations based on the resulting evaluation performance.

**Data splits.** To prevent data leakage, we use disjoint data partitions for concept-to-latent mapping, coefficient estimation, and evaluation. For lexical-semantic experiments, SemCor is partitioned by

[1] https://image-net.org/challenges/LSVRC/2012/2012-downloads

source file into mapping ($\mathcal{D}_{\text{map}}$), fitting ($\mathcal{D}_{\text{fit}}$), and testing ($\mathcal{D}_{\text{test}}$) splits in a 40/30/30 ratio. The vision experiments follow the same partitioning strategy at the image level, ensuring that all patches from a given image belong to the same split. Concept-to-latent mapping is performed exclusively on $\mathcal{D}_{\text{map}}$, DAR transfer coefficients are estimated using $\mathcal{D}_{\text{fit}}$, and all evaluation metrics are computed on the held-out $\mathcal{D}_{\text{test}}$.

For the spelling experiments, we follow a 50/50 calibration–test split, using separate halves for coefficient estimation and held-out evaluation.

### A.2.5 PROTOCOL

**Absorbed and negative-control examples.** For each supplied group $(p, C_p)$, an *absorbed example* is a held-out input on which the parent concept is present according to ground truth, at least one supplied child latent fires, and the corresponding parent latent remains inactive. *Negative-control examples* are held-out inputs on which the parent concept is absent. Both sets are determined using the original SAE and benchmark annotations before applying DAR.

**Firing criterion.** We evaluate parent recovery using fixed, architecture-specific activation thresholds. For JumpReLU SAEs, the repaired parent activation is compared with the original learned threshold of the corresponding latent. For TopK SAEs, we use the $k$-th largest pre-activation of the original input, with the threshold additionally clamped at zero for the GPT-2 TopK SAE. For BatchTopK SAEs, we use the original SAE's estimated global inference threshold.

All evaluation thresholds are determined before applying DAR and remain fixed across correction strengths. Unless otherwise specified, we use the full correction strength $\gamma = 1$. We additionally consider $z'_p(x) > 0$ as a secondary diagnostic rather than the primary recovery criterion.

**Evaluation metrics.** Recovery is reported using both macro-averaging across parent groups and example-weighted averaging to account for differences in the number of absorbed examples. For semantic experiments, we additionally report *addressable recovery*, restricting evaluation to absorbed examples for which the relevant parent and child concepts map to distinct SAE latents.

To quantify changes in the parent-child activation relationship, we compute Spearman's rank correlation (Spearman, 1904):

$$s_{\text{cc}} = \text{Spearman}\left(z_p(x), \max_{c \in C_p} z_c(x)\right), \tag{29}$$

over held-out child-active examples. We compare the correlation before and after DAR using the same examples. A negative correlation indicates an inverse association between parent and child activations, consistent with the suppression pattern associated with feature absorption.

We additionally measure the parent-given-child activation rate, denoted by $s_{\text{ai}}$, as the proportion of held-out child-active examples on which the parent latent is active under the corresponding evaluation threshold.

Reconstruction preservation is verified using the relative reconstruction difference

$$\Delta_{\text{rec}} = \frac{\|\widehat{X}' - \widehat{X}\|_F}{\|\widehat{X}\|_F}, \tag{30}$$

where $\widehat{X}$ and $\widehat{X}'$ denote the original and repaired reconstructions, respectively. We quantify the representational cost through changes in the number of active latents per example and the $\ell_1$ mass of the latent code.

**Causal evaluation.** We evaluate the effects of DAR through parent steering and child-to-parent spillover, while examining whether the child's original causal effect is retained. Statistical uncertainty is quantified using percentile bootstrap confidence intervals, with resampling at the parent-group level where applicable.

**Reproducibility.** All experimental protocols, group-selection rules, and evaluation criteria are fixed before DAR evaluation. Supplied parent-child groups are not added, removed, or selected

Table 4: Generalization to vision using CLIP ViT-L/14-336 at layer 22 with a BatchTopK SAE and ImageNet synset hierarchy.

| Quantity | Before → After |
|---|---|
| Row-weighted recovery | $0.000 \rightarrow \mathbf{0.262}$ |
| Recovery, random edges | $0.000 \rightarrow 1.5 \times 10^{-5}$ |
| Signature $s_{\mathrm{cc}}$ | $+0.135 \rightarrow \mathbf{+0.634}$ |
| Leakage | $0.00419 \rightarrow 0.00439$ |
| Reconstruction change | $5.24 \times 10^{-17}$ |

Table 5: Cross-family generalization using GPT-2 Small at layer 6 with the OpenAI v5 TopK SAE and WordNet relations.

| Quantity | Before → After |
|---|---|
| Addressable recovery | $0.000 \rightarrow \mathbf{0.369}$ |
| Row-weighted recovery | $0.000 \rightarrow 0.351$ |
| Absorption prevalence | $0.142 \rightarrow \mathbf{0.080}$ |
| Parent-ablation effect | $0.000 \rightarrow \mathbf{-1.982}$ |
| Leakage increase | $+0.00059$ |
| Recon. change, SAE space | $6.53 \times 10^{-17}$ |
| Recon. change, model space | $9.75 \times 10^{-17}$ |

according to DAR outcomes. We record SHA-256 hashes of the core DAR implementation and frozen benchmark artifacts, together with runtime consistency checks where implemented, to verify that the method and benchmark definitions remain unchanged across experiments.

### A.2.6 Does the Repair Generalize?

We investigate whether DAR generalizes beyond the model, SAE architecture, and benchmark used in the preceding experiments. Specifically, we evaluate DAR in two additional settings: a vision model with a BatchTopK SAE and a different language model with a TopK SAE. In both settings, we retain the same DAR procedure and coefficient-estimation protocol without modifying its selection rules or tuning the method on the held-out evaluation data. Transfer coefficients are estimated separately using the corresponding calibration splits.

**Generalization to vision.** We first evaluate DAR using CLIP ViT-L/14-336 at layer 22 with a 65,536-feature BatchTopK SAE ($k = 20$). parent-child relations are constructed from the ImageNet synset hierarchy, and evaluation is performed on held-out images. As shown in Table 4, DAR achieves a row-weighted recovery of $0.262$ on previously absorbed patches. The absorption signature $s_{\mathrm{cc}}$ increases from $+0.135$ to $+0.634$, indicating a stronger positive association between parent and child activations after repair. Meanwhile, leakage increases only slightly, from $0.00419$ to $0.00439$, corresponding to an absolute increase of $2.0 \times 10^{-4}$.

To examine whether these improvements depend on meaningful parent-child relations, we conduct a control experiment using $50$ groups constructed from random non-hypernym edges. Applying the same DAR procedure to these groups yields a row-weighted recovery of only $1.5 \times 10^{-5}$, compared with $0.262$ for the supplied ImageNet hierarchy. Furthermore, patches on which the parent was already active retain a firing rate of $1.000$ after repair. The relative reconstruction difference remains $5.24 \times 10^{-17}$, confirming reconstruction preservation to numerical precision in the vision setting.

**Generalization across language-model and SAE families.** We further evaluate DAR using GPT-2 Small with the OpenAI v5 TopK SAE at layer 6. Unlike the preceding experiments, the parent-child relations are constructed using WordNet hypernym relations over SemCor, with data partitioned by source file. As reported in Table 5, DAR achieves an addressable recovery of $0.369$ (95% CI $[0.178, 0.558]$) and a row-weighted recovery of $0.351$. The prevalence of absorption decreases from $0.142$ to $0.080$, indicating fewer absorption cases after repair. These improvements are accompanied by an absolute leakage increase of only $5.9 \times 10^{-4}$. The relative reconstruction differences are $6.53 \times 10^{-17}$ in the SAE's internal reconstruction space and $9.75 \times 10^{-17}$ in the model's activation space. These measurements confirm that reconstruction preservation is maintained numerically despite the SAE's input-normalization procedure.

We additionally examine the causal effect of the repaired parent through latent ablation. The measured parent-ablation effect changes from $0.000$ before repair to $-1.982$ after repair, with a 95% confidence interval of $[-3.180, -0.885]$ for the reported post-repair effect. This result provides additional evidence that the repaired parent becomes an effective intervention target in the cross-family setting.

Overall, these experiments demonstrate that DAR can restore parent activations across different model families, SAE architectures, and data modalities. The substantially lower recovery obtained from random non-hypernym relations suggests that the improvements depend on the supplied semantic hierarchy. Across both settings, DAR preserves the original reconstruction to numerical precision while improving parent recovery, with the accompanying leakage measurements quantifying unintended activations.

### A.2.7 Does a Repaired Latent Matter?

We investigate whether the repaired parent latent becomes more useful for downstream interpretability tasks. In particular, we examine whether DAR strengthens the parent's causal influence on model predictions and improves its ability to identify the intended concept. Both evaluations use the supplied hierarchy of $24$ parent latents introduced in Section 5.1.

**Causal intervention.** We assess the causal effect of the repaired parent by ablating it on held-out tokens affected by absorption and measuring the resulting changes in model outputs. For each intervention, the modified SAE reconstruction is patched into the model's residual stream, with the original SAE reconstruction error added back to isolate the intervention from reconstruction loss. Because DAR preserves the original reconstruction, the unablated repaired representation reproduces the original model activation with a numerical difference of $3.6 \times 10^{-15}$. Thus, the reparameterization itself leaves the reconstructed activation unchanged, allowing us to examine the effect of interventions on the repaired coordinates. As shown in Table 6, ablating the parent changes the target logit by $-0.038$ before repair and by $-0.596$ after repair. The magnitude of the ablation effect therefore increases by approximately $16$ times, with larger effects observed for $22$ of the $24$ parent latents. Similarly, the parent-ablation effect on a held-out semantic probe increases in magnitude from $-0.348$ to $-3.664$.

We further examine whether DAR reduces the parent-related influence carried by child latents. When amplifying a child latent, the corresponding parent-probe response decreases from $8.456$ to $5.521$, representing a $34.7\%$ reduction. The corresponding logit-response slope decreases from $0.885$ to $0.454$, a reduction of $48.8\%$. Since DAR leaves the child activation unchanged while modifying its decoder direction, these results are consistent with a redistribution of parent-related information from the child decoder to the repaired parent latent.

**Intended-latent readout.** We next examine whether the repaired parent provides a more accurate representation of its intended concept. All $24$ hierarchy-group transformations are composed into a single repaired code matrix and decoder. The resulting representation contains no negative latent activations and preserves reconstruction with a relative difference of $5.6 \times 10^{-17}$. We evaluate the intended parent latents on a held-out set of $15{,}031$ tokens, including $596$ tokens identified as absorbed, using each token's first-letter concept as the evaluation target. As reported in Table 6, the intended-latent AUROC on absorbed tokens increases from $0.535$ to $0.941$, while classification accuracy increases from $0.079$ to $0.543$. These improvements indicate that the repaired parent activations provide substantially more informative signals for identifying their intended concepts on absorption-affected inputs.

Importantly, accuracy on non-absorbed tokens also increases, from $0.781$ to $0.824$, while overall accuracy improves from $0.752$ to $0.813$. Thus, the improvements on absorbed tokens are accompanied by improved readout accuracy on the remaining held-out examples.

Overall, these experiments demonstrate that DAR improves both the causal influence and the predictive usefulness of repaired parent latents. Following repair, parent ablation produces a substantially larger effect on model outputs, while child-to-parent spillover is reduced. At the same time, the repaired parent activations provide more accurate signals for identifying their intended concepts, without changing the original SAE reconstruction.

## A.3 Ablation Studies

We conduct ablation studies to examine the contributions of three key components of DAR: the sign constraint on the transfer coefficients, the paired code–decoder transformation, and the correction

Table 6: Downstream evaluation of DAR on the supplied hierarchy of 24 parents using Gemma-2-2B at layer 12. The upper block reports causal interventions on absorption-affected tokens, while the lower block reports intended-latent readout performance on held-out tokens.

| Measurement | Before → After | Change |
|---|---|---|
| *Causal intervention* | | |
| Parent ablation, target logit | $-0.038 \rightarrow \mathbf{-0.596}$ | $-0.558$ |
| Parent ablation, probe score | $-0.348 \rightarrow \mathbf{-3.664}$ | $-3.316$ |
| Child-to-parent spillover, probe | $8.456 \rightarrow \mathbf{5.521}$ | $-34.7\%$ |
| Child-to-parent spillover, logit | $0.885 \rightarrow \mathbf{0.454}$ | $-48.8\%$ |
| *Intended-latent readout* | | |
| AUROC, absorbed tokens | $0.535 \rightarrow \mathbf{0.941}$ | $+0.406$ |
| Accuracy, absorbed tokens | $0.079 \rightarrow \mathbf{0.543}$ | $+0.464$ |
| Accuracy, non-absorbed tokens | $0.781 \rightarrow 0.824$ | $+0.043$ |
| Accuracy, overall | $0.752 \rightarrow 0.813$ | $+0.061$ |

strength. Unless otherwise specified, experiments use the supplied hierarchy of 24 parents and 245 child edges with Gemma-2-2B at layer 12 and are evaluated on held-out data.

**Effect of the sign constraint.** We investigate whether the sign constraint is necessary to maintain non-negative repaired latent codes. We compare three coefficient-estimation strategies: unconstrained least squares, least squares with non-negativity constraints imposed on the calibration examples, and the proposed sign-constrained estimator. As reported in Table 7, unconstrained least squares improves 22 of the 24 groups but produces negative parent activations in 10 groups on both the calibration and held-out splits.

Imposing non-negativity constraints on the calibration examples improves all 24 groups and eliminates negative activations on those examples. However, negative parent activations still occur in four groups on the held-out split, reaching a minimum of $-18.85$. In contrast, the proposed sign-constrained estimator improves all 24 groups without producing negative parent activations on either split. Unlike constraints imposed only on calibration examples, the condition $\widehat{\beta} \leq 0$ guarantees that $z'_p(x) \geq z_p(x) \geq 0$ for any non-negative SAE code. Thus, the sign constraint ensures that DAR preserves non-negativity even on previously unseen inputs.

**Effect of the paired transformation.** We examine the contributions of the code and decoder updates by applying each independently. We also compare the proposed joint estimation of a parent's transfer coefficients with fitting each child separately. As shown in Table 9, applying only the code update produces the same reported recovery (0.431), leakage (0.0292), and absorption signature ($+0.047$) as the complete DAR transformation. However, without the compensating decoder update, the relative reconstruction difference increases from $4.1 \times 10^{-18}$ to $1.5 \times 10^{-2}$. Conversely, applying only the decoder update leaves the latent activations unchanged, resulting in the original recovery (0.085), leakage (0.0185), and absorption signature ($-0.204$). Nevertheless, this modification changes the reconstruction by $1.8 \times 10^{-2}$.

These results demonstrate the complementary roles of the two updates. The code update increases parent activation, while the corresponding decoder update compensates for this change to preserve the original reconstruction. Both are therefore necessary for the reconstruction-preserving reparameterization. We further compare joint coefficient estimation with independent fitting for individual children. Joint estimation achieves a reported recovery of 0.431, compared with 0.396 for individual fits, while reducing leakage from 0.0350 to 0.0292. The absorption signature also increases from $+0.022$ to $+0.047$. These results indicate that jointly accounting for multiple child activations produces more effective corrections under the evaluated conditions.

**Effect of correction strength.** We examine the sensitivity of DAR to the correction strength $s \in [0, 1]$ across 168 parent groups and 14 experimental configurations. The transfer coefficients are estimated once for each configuration and subsequently scaled without refitting.

Table 7: Effect of coefficient constraints. Negative parent codes are counted as the number of groups producing at least one.

| Estimator | Improved | Neg. (cal.) | Neg. (held out) |
|---|---|---|---|
| Unconstrained | 22/24 | 10/24 | 10/24 |
| Row-wise | 24/24 | 0/24 | 4/24 |
| **Sign-constrained** | **24/24** | **0/24** | **0/24** |

Table 8: Effect of correction strength across 168 groups and 14 configurations.

| $\gamma$ | Recovery | Row-wtd. | Leakage increase |
|---|---|---|---|
| 0.00 | 0.158 | 0.090 | 0.0000 |
| 0.25 | 0.198 | 0.119 | **0.0000** |
| 0.50 | 0.320 | 0.224 | +0.0022 |
| 0.75 | 0.386 | 0.336 | +0.0050 |
| 1.00 | **0.426** | 0.418 | +0.0092 |

Table 9: Ablation of the paired code–decoder transformation and coefficient-estimation strategy on the supplied hierarchy. Without repair, $s_{\text{cc}} = -0.204$.

| Variant | Recovery | Leakage | Recon. change | $s_{\text{cc}}$ |
|---|---|---|---|---|
| No repair | 0.085 | 0.0185 | 0 | −0.204 |
| **DAR, grouped** | **0.431** | **0.0292** | $\mathbf{4.1 \times 10^{-18}}$ | **+0.102** |
| DAR, single-child fits | 0.396 | 0.0350 | $4.5 \times 10^{-18}$ | +0.022 |
| Code update only | 0.431 | 0.0292 | $1.5 \times 10^{-2}$ | +0.047 |
| Decoder update only | 0.085 | 0.0185 | $1.8 \times 10^{-2}$ | −0.204 |

Table 8 summarizes the results. As correction strength increases from 0 to 1, the reported recovery increases from 0.158 to 0.426, while row-weighted recovery increases from 0.090 to 0.418. The reported leakage increase reaches 0.0092 at full correction strength. At $\gamma = 0.25$, recovery increases from 0.158 to 0.198, without a measurable increase in leakage. Further increases in correction strength produce additional recovery accompanied by higher leakage, illustrating the trade-off between parent restoration and unintended activations. For any fixed input and fitted coefficient vector, DAR guarantees that the repaired parent activation is non-decreasing with correction strength. Since the evaluation thresholds remain fixed, both recovery and leakage are consequently non-decreasing with $s$. This behavior is also observed across all 14 experimental configurations.

The set of nonzero repaired codes remains unchanged for $\gamma > 0$ in the evaluated configurations. Moreover, the maximum observed relative reconstruction difference is $1.24 \times 10^{-16}$, and no negative repaired codes are observed at any correction strength. These findings show that correction strength provides control over the recovery-leakage trade-off while maintaining reconstruction preservation and code non-negativity. Unless otherwise specified, we use $\gamma = 1$ throughout the experiments.

**Causal steering and specificity.** DAR makes previously absorbed parent features substantially more controllable: the parent-steering slope increases from 0.044 to 0.505 on held-out absorbed examples (Fig. 5a). Parent-positive examples remain strongly affected, while letter-negative examples remain near zero (Fig. 5b). The child's own steering effect is largely preserved, retaining 97.5% of its original magnitude ($9.545 \rightarrow 9.305$; Fig. 5c). At the same time, the effect of child steering on the parent-related output decreases from 0.853 to 0.332, a 61.1% reduction (Fig. 5d). Thus, DAR strengthens control through the parent feature while preserving most of the child's effect and reducing child-to-parent spillover.

## B LIMITATIONS

Despite its reconstruction-preserving guarantees and empirical improvements, DAR has several limitations. First, it requires externally supplied parent-child relations and does not automatically discover feature hierarchies. Its effectiveness therefore depends on the quality of the supplied relations and the correspondence between semantic concepts and SAE latents. Second, transfer coefficients are estimated from observed parent-child activation patterns rather than direct measurements of absorbed semantic information. Consequently, their reliability may be affected by limited calibration data, correlated child activations, or distribution shifts. The sign constraint guarantees non-negative repaired codes and monotone parent activation, but it does not guarantee complete semantic recovery.

Furthermore, restoring parent activations introduces a trade-off between recovery and unintended activations. Although DAR preserves reconstruction exactly, it can increase latent sparsity costs and




Figure 4: Additional qualitative examples of DAR repair using Gemma-2-2B at layer 12 with the Gemma Scope 16K JumpReLU SAE. In each example, the parent code is initially silent, while the child decoder atom contains a component along the parent direction. DAR increases the parent code and reduces this component in the updated child decoder atom, while leaving the child code unchanged.

modify child decoder directions. The repaired codes also need not satisfy the original encoder's native gating rule. Finally, our experiments cover selected language and vision models, SAE architectures, and semantic hierarchies. The effectiveness of DAR on substantially larger, overlapping hierarchies and other model families remains to be investigated.

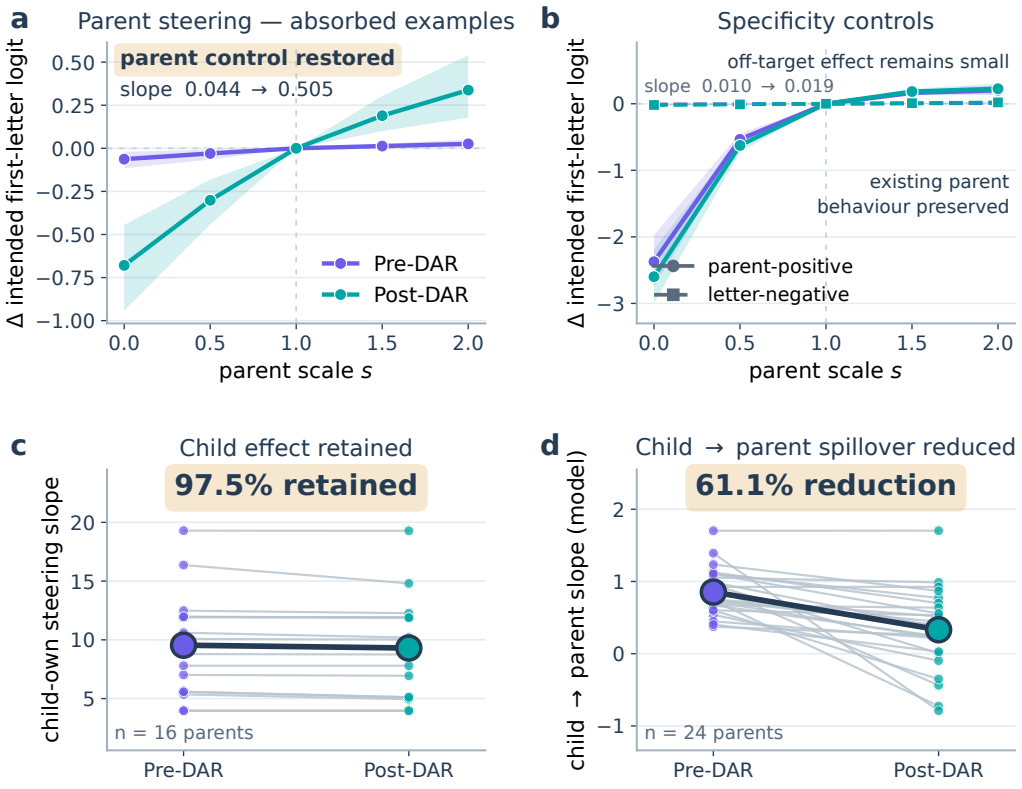


Figure 5: **Parent and child steering before and after DAR.** (a) Parent steering on held-out absorbed spelling examples, measured by the intended first-letter logit ($n = 24$ parents, 266 tokens). $s = 1$ is the original representation, $s = 0$ suppresses the parent, and $s > 1$ amplifies it. (b) The same intervention on parent-positive and letter-negative examples. (c) Child steering on the SemCor/WordNet benchmark using a frozen child-semantic probe ($n = 16$ parents, 553 held-out rows). (d) Effect of child steering on the parent-related spelling output ($n = 24$ parents). Thin lines in (c–d) denote individual parent groups and large markers denote their means.