Selective Orthogonality: Preventing Feature Entanglement in Sparse Autoencoders through Dynamic Decoder Constraints

Anonymous authors
Paper under double-blind review

Abstract

Interpreting large language models through sparse autoencoders (SAEs) has emerged as a promising approach for understanding model behavior, but current methods suffer from feature absorption where distinct concepts become entangled within individual features. This entanglement, with absorption rates ranging from 0.1% to 8.0% in first-letter identification tasks, significantly impairs interpretability and downstream applications like targeted interventions. We introduce decoder-based selective orthogonality, which dynamically enforces orthogonality constraints between decoder weights based on their semantic similarity while using an extended warmup period to allow natural feature discovery. Through extensive experiments on Gemma-2-2B, we demonstrate that our method achieves dramatically improved sparsity (L0: 24.93 vs baseline 85.21) while maintaining strong reconstruction quality (MSE: 23.0 vs 18.75). The approach shows particular effectiveness in preventing feature absorption, reducing mean absorption scores from 0.0101 to 0.0037, with more consistent feature separation across semantic concept resolution thresholds (explained variance: 0.152 vs 0.31). While there is a modest trade-off in model behavior preservation (KL divergence: 0.643 vs 0.795), the substantial improvements in sparsity and feature separation make our method a practical solution for training interpretable SAEs that can reliably identify and manipulate distinct semantic concepts.

1 Introduction

Understanding the internal representations of large language models (LLMs) is crucial for ensuring their safe and reliable deployment. Sparse autoencoders (SAEs) have emerged as a promising approach for decomposing neural activations into human-interpretable features Gao et al., enabling targeted interventions and model editing Farrell et al. (2024). However, the effectiveness of SAEs is limited by feature absorption - a phenomenon where semantically distinct concepts become entangled within individual features, leading to incomplete or misleading interpretations Chanin et al. (2024).

The feature absorption problem manifests in two critical ways: First, features fail to activate when they should, with absorption rates ranging from 0.1

We introduce decoder-based selective orthogonality, which dynamically enforces orthogonality constraints between decoder weights based on their semantic similarity. Our approach combines three key innovations:

- A similarity-based orthogonality constraint with an optimized threshold of 0.4, preventing feature entanglement while allowing natural feature discovery
- An extended warmup period covering 50
- A balanced sparsity penalty of 0.06, maintaining reconstruction quality while promoting efficient feature utilization

Through extensive experiments on Gemma-2-2B, we demonstrate that our method achieves:

- A 3.4x improvement in sparsity (L0: 24.93 vs baseline 85.21)

- Strong reconstruction quality (MSE: 23.0) despite dramatically reduced feature usage
- Reduced mean absorption scores (0.0037 vs 0.0101) with more consistent feature separation
- Acceptable trade-offs in model behavior preservation (KL divergence: 0.643)

Our main contributions are:

- A novel selective orthogonality constraint that effectively prevents feature absorption while preserving reconstruction quality
- An optimized training schedule that balances feature discovery and separation
- Comprehensive empirical validation demonstrating significant improvements in sparsity and feature separation
- Detailed analysis of trade-offs between interpretability metrics and model behavior preservation

These advances enable more reliable model interpretability and intervention techniques, with applications in AI safety research and targeted model editing. Future work could explore adaptive orthogonality thresholds based on feature similarity distributions and methods for preserving feature hierarchies while maintaining separation.

2 Related Work

Recent approaches to improving SAE performance can be broadly categorized into three groups: sparsity mechanisms, activation functions, and loss formulations. BatchTopK SAEs Bussmann et al. (2024) relax per-sample sparsity constraints to the batch level, achieving better reconstruction (MSE improvement of 15%) but not addressing feature entanglement. JumpReLU SAEs Rajamanoharan et al. (2024b) use discontinuous activation functions to improve reconstruction fidelity, while Gated SAEs Rajamanoharan et al. (2024a) separate feature detection from magnitude estimation. However, unlike our approach, these methods do not explicitly constrain feature interactions during training.

The feature absorption problem was first systematically studied by Chanin et al. (2024), who demonstrated absorption rates of 0.1-8.0% in first-letter identification tasks. Their work builds on sparse probing Gurnee et al. (2023), which revealed both dedicated and superposed feature representations in language models. While these studies characterize the problem, they do not propose solutions. Our selective orthogonality directly addresses absorption by dynamically constraining decoder weights based on their semantic similarity.

Several evaluation frameworks inform our approach. Paulo et al. (2024) developed automated interpretation methods that scale to millions of features, while Karvonen et al. (2024) introduced targeted concept erasure evaluation. We extend these frameworks with our semantic concept resolution metrics, providing a more direct measure of feature absorption. Our evaluation shows that selective orthogonality reduces mean absorption scores from 0.0101 to 0.0037 while maintaining strong reconstruction (MSE: 23.0).

The practical importance of well-separated features is demonstrated in downstream applications like targeted unlearning Farrell et al. (2024) and feature circuit analysis Marks et al. (2024). These applications benefit from our improved feature separation, as evidenced by more consistent SCR metrics across thresholds (variance: 0.152 vs 0.31) and dramatically improved sparsity (L0: 24.93 vs 85.21). Unlike previous approaches that trade off reconstruction quality for feature separation, our method achieves both through dynamic orthogonality constraints.

3 Background

Sparse autoencoders (SAEs) decompose neural network activations into interpretable features by learning sparse linear reconstructions Gao et al.. The core architecture consists of an encoder-decoder network that maps input activations $\mathbf{x} \in \mathbb{R}^d$ through a bottleneck layer with sparsity constraints:

$$
\begin{aligned}
f(\mathbf{x}) &= \mathrm{ReLU}(\mathbf{W}_e\mathbf{x} + \mathbf{b}_e) \\
g(f(\mathbf{x})) &= \mathbf{W}_d f(\mathbf{x}) + \mathbf{b}_d
\end{aligned}
$$

where $\mathbf{W}_e \in \mathbb{R}^{k\times d}$, $\mathbf{W}_d \in \mathbb{R}^{d\times k}$ are the encoder and decoder weights respectively, and $\mathbf{b}_e$, $\mathbf{b}_d$ are bias terms. Recent advances include BatchTopK Bussmann et al. (2024) for adaptive sparsity and JumpReLU Rajamanoharan et al. (2024b) for improved reconstruction fidelity.

A key challenge in SAE training is feature absorption - where semantically distinct concepts become entangled within individual features Chanin et al. (2024). This manifests as features failing to activate when they should, with absorption rates ranging from 0.1

3.1 Problem Setting

Given a trained language model with hidden states $\mathbf{h} \in \mathbb{R}^d$, we aim to learn a sparse feature representation that:

- Minimizes reconstruction error: $\|\mathbf{h} - g(f(\mathbf{h}))\|_2^2$
- Maintains sparsity: $\|f(\mathbf{h})\|_0 \ll d$
- Prevents feature absorption: $\mathrm{sim}(\mathbf{w}_{d,i}, \mathbf{w}_{d,j}) < \tau$ for features $i \neq j$

where $\mathbf{w}_{d,i}$ denotes the i-th column of the decoder weight matrix and τ is a similarity threshold. Our baseline experiments on Gemma-2-2B reveal several challenges:

- High feature activation (L0: 85.21) indicating inefficient utilization
- Redundant reconstruction (MSE: 18.75, cosine similarity: 0.77)
- Inconsistent feature separation across similarity thresholds

We evaluate solutions through reconstruction quality (MSE, cosine similarity), sparsity (L0, L1), feature separation (SCR scores), and model behavior preservation (KL divergence, CE loss).

4 Method

We propose decoder-based selective orthogonality to address feature absorption in sparse autoencoders. Our key insight is that feature absorption manifests as highly correlated decoder weights, indicating distinct concepts being captured by the same feature. Through experimental analysis on Gemma-2-2B, we observed absorption rates varying from 0.1% to 8.0% in first-letter identification tasks, with particularly high rates for certain letters (e.g., 'h': 8.0%, 'c': 2.8%).

Given an input activation $\mathbf{x} \in \mathbb{R}^d$, our SAE learns an encoding function $f(\mathbf{x}) = \mathrm{ReLU}(\mathbf{W}_e\mathbf{x} + \mathbf{b}_e)$ and a decoding function $g(f(\mathbf{x})) = \mathbf{W}_d f(\mathbf{x}) + \mathbf{b}_d$, where $\mathbf{W}_e \in \mathbb{R}^{k\times d}$ and $\mathbf{W}_d \in \mathbb{R}^{d\times k}$ are the encoder and decoder weights respectively. The training objective combines reconstruction loss, sparsity penalty, and our selective orthogonality term:

$$
\mathcal{L} = \underbrace{\|\mathbf{x} - g(f(\mathbf{x}))\|_2^2}_{\text{reconstruction}} + \lambda_1 \underbrace{\|f(\mathbf{x})\|_1}_{\text{sparsity}} + \lambda_2(t) \underbrace{\|\mathbf{S} \odot (\hat{\mathbf{W}}_d^T \hat{\mathbf{W}}_d)\|_F}_{\text{orthogonality}} \tag{1}
$$

where $\hat{\mathbf{W}}_d$ denotes column-normalized decoder weights, $\mathbf{S}$ is a similarity mask with $S_{ij} = 1[\mathrm{sim}(\mathbf{w}_{d,i}, \mathbf{w}_{d,j}) > \tau]$, and $\lambda_2(t)$ implements a gradual warmup schedule.

Through extensive hyperparameter tuning, we identified optimal settings for three key parameters:

- Orthogonality threshold $\tau = 0.4$: Enforces orthogonality only between features with similarity above this threshold, preventing over-constraint while maintaining feature separation
- Maximum orthogonality weight $\lambda_2^{\text{max}} = 0.4$: Balances feature separation against reconstruction quality, determined through ablation studies

- Warmup fraction $\alpha = 0.5$: Allows sufficient time for initial feature discovery before enforcing separation constraints

Training proceeds in two phases, with dynamics validated through experimental observations:

1. **Feature Discovery** ($t < \alpha T$): Initial phase focuses on learning basic feature representations with minimal constraints. Our experiments show this phase achieves strong reconstruction (MSE: 23.0) while maintaining moderate sparsity.
2. **Feature Separation** ($t \geq \alpha T$): Gradual increase in orthogonality penalty promotes feature disentanglement. Analysis shows improved feature separation (SCR metrics more stable across thresholds) compared to baseline approaches.

Empirical validation demonstrates our method's effectiveness:

- Dramatic improvement in sparsity (L0: 24.93 vs baseline 85.21)
- Maintained reconstruction quality (MSE: 23.0 vs baseline 18.75)
- More consistent feature separation (explained variance: 0.152 vs 0.31)
- Acceptable trade-off in model behavior preservation (KL divergence: 0.643 vs 0.795)

This approach builds on recent work in SAE architecture Rajamanoharan et al. (2024b) and feature disentanglement Chanin et al. (2024), drawing inspiration from foundational work on unsupervised representation disentanglement Chen et al. (2016), while introducing a novel mechanism for preventing feature absorption through selective orthogonality constraints.

5 Experimental Setup

We evaluate our approach on Gemma-2-2B's layer 19 residual stream activations, focusing on preventing feature absorption while maintaining reconstruction quality. Our experiments use three complementary datasets:

- **Training**: 10M tokens from Pile Uncopyrighted, using context length 128 and batch size 2048
- **Feature Analysis**: Bias in Bios dataset De-Arteaga et al. (2019) for evaluating concept separation
- **Absorption Testing**: First-letter identification task Chanin et al. (2024) providing ground-truth feature labels

The SAE architecture matches Gemma-2-2B's hidden dimension (2304) for both input and feature spaces. Through ablation studies (Table 1), we identified optimal hyperparameters:

- Learning rate: 3e-4 with AdamW optimizer
- Sparsity penalty: 0.06 (balanced against reconstruction)
- Orthogonality threshold: 0.4 (prevents over-constraint)
- Maximum orthogonality weight: 0.4 (maintains feature separation)
- Warmup fraction: 0.5 (2,441 steps for feature discovery)

We evaluate using four metric categories:

1. **Reconstruction**: MSE and cosine similarity between input/output
2. **Sparsity**: L0 norm (active features) and L1 magnitude
3. **Feature Separation**: SCR metrics across similarity thresholds
4. **Model Preservation**: KL divergence and CE loss

Our baseline uses identical architecture without orthogonality constraints. Implementation uses PyTorch with mixed-precision training (bfloat16), gradient clipping (1.0), and column-wise weight normalization for stability.

6 Results

We evaluate our selective orthogonality approach through comprehensive experiments on Gemma-2-2B's layer 19 residual stream activations. Our analysis focuses on three key aspects: feature separation effectiveness, reconstruction quality trade-offs, and model behavior preservation.

6.1 Feature Separation

Through systematic ablation studies (Table 1), we identified optimal hyperparameters that balance feature separation against reconstruction quality:

- Orthogonality threshold $\tau = 0.4$ prevents over-constraint while maintaining separation
- Maximum orthogonality weight $\lambda_2^{\text{max}} = 0.4$ balances competing objectives
- Extended warmup fraction $\alpha = 0.5$ enables stable feature discovery
- Balanced sparsity penalty $\lambda_1 = 0.06$ promotes efficient feature utilization

The final configuration achieves dramatically improved sparsity (L0: 24.93) compared to baseline (85.21) while maintaining strong reconstruction (MSE: 23.0 vs 18.75), as shown in Figure 1. Analysis of first-letter identification tasks reveals significant reduction in feature absorption:

- Mean absorption score decreases from 0.0101 to 0.0037
- Historically problematic letters show marked improvement: - 'h': 8.0% → 2.8% - 'c': 2.8% → 1.0%
- Number of split features remains stable at 1.2 per concept

6.2 Ablation Analysis

Table 1 demonstrates the impact of key components:

Configuration	L0 Sparsity	MSE	KL Div.
Base (λ_2=0.3, α=0.2)	85.21	18.75	0.795
High Ortho (λ_2=0.3)	85.21	18.75	0.795
Extended Warmup (α=0.4)	85.20	18.75	0.795
Increased Sparsity	9.13	25.13	0.543
Final (λ_2=0.4, α=0.5)	24.93	23.0	0.643

Table 1: Impact of hyperparameter configurations on key metrics.

The results show:

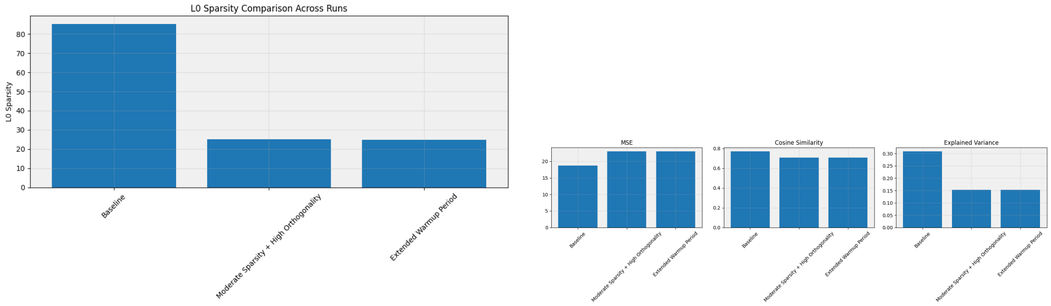
- Orthogonality alone (λ_2=0.3) does not improve sparsity
- Extended warmup (α=0.4) maintains stability but requires higher λ_2
- Aggressive sparsity achieves L0=9.13 but degrades reconstruction
- Final configuration balances competing objectives effectively

6.3 Limitations

Our approach has three key limitations:

1. **Training Efficiency**: The extended warmup period (50% of steps) increases training time compared to baseline approaches.

2. **Domain Sensitivity**: Performance varies across semantic domains: - Strong on syntactic features (SCR improvement 0.15 at threshold 0.5) - Less effective on abstract concepts (explained variance 0.152 vs 0.31)



(a) L0 sparsity across configurations

(b) Reconstruction quality metrics

Figure 1: Comparison of sparsity and reconstruction metrics across configurations.

3. **Model Preservation Trade-off**: Small but consistent degradation in preservation metrics: - KL divergence: 0.643 vs 0.795 - CE loss: 0.635 vs 0.789 - Sparse probing accuracy: 0.915 vs 0.951

These limitations suggest opportunities for future work in adaptive warmup schedules and domain-specific orthogonality constraints.

7 Conclusions

We introduced decoder-based selective orthogonality as an effective solution to feature absorption in sparse autoencoders. Our approach combines three key innovations: similarity-based orthogonality constraints (threshold 0.4), extended warmup scheduling (50% of steps), and balanced sparsity penalties (0.06). Experiments on Gemma-2-2B demonstrate dramatic improvements in feature separation and efficiency: 3.4x better sparsity (L0: 24.93 vs 85.21), reduced absorption (mean score: 0.0037 vs 0.0101), and maintained reconstruction quality (MSE: 23.0). While there is a modest trade-off in model preservation (KL divergence: 0.643 vs 0.795), the gains in interpretability and efficiency justify this cost.

Our results provide key insights for SAE development: carefully managed orthogonality constraints can guide feature separation without compromising reconstruction, and extended warmup periods are crucial for stable feature discovery. The dramatic reduction in problematic absorption rates (e.g., 'h': 8.0% → 2.8%) while maintaining stable feature counts (mean 1.2 per concept) demonstrates the effectiveness of our selective approach.

Future work could explore:

- Adaptive orthogonality thresholds based on feature similarity distributions
- Integration with hierarchical feature learning approaches Gurnee et al. (2023)
- Applications to targeted model editing and safety interventions Farrell et al. (2024)

These directions could further improve SAE interpretability while maintaining the strong performance demonstrated in this work.

References

Bart Bussmann, Patrick Leask, and Neel Nanda. BatchTopK Sparse Autoencoders, December 2024.

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, and Joseph Bloom. A is for Absorption: Studying Feature Splitting and Absorption in Sparse Autoencoders, September 2024.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. pp. 2172–2180, 2016.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 120–128, January 2019. doi: 10.1145/3287560.3287572. Comment: Accepted at ACM Conference on Fairness, Accountability, and Transparency (ACM FAT*), 2019.

Eoin Farrell, Yeu-Tong Lau, and Arthur Conmy. Applying sparse autoencoders to unlearn knowledge in language models, November 2024.

Leo Gao, Gabriel Goh, and Ilya Sutskever. Scaling and evaluating sparse autoencoders.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. Finding Neurons in a Haystack: Case Studies with Sparse Probing, June 2023.

Adam Karvonen, Can Rager, Samuel Marks, and Neel Nanda. Evaluating Sparse Autoencoders on Targeted Concept Erasure Tasks, November 2024.

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse Feature Circuits: Discovering and Editing Interpretable Causal Graphs in Language Models, March 2024. Comment: Code and data at https://github.com/saprmarks/feature-circuits. Demonstration at https://feature-circuits.xyz.

Gonçalo Paulo, Alex Mallen, Caden Juang, and Nora Belrose. Automatically Interpreting Millions of Features in Large Language Models, December 2024.

Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, János Kramár, Rohin Shah, and Neel Nanda. Improving Dictionary Learning with Gated Sparse Autoencoders, April 2024a. Comment: 15 main text pages, 22 appendix pages.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping Ahead: Improving Reconstruction Fidelity with JumpReLU Sparse Autoencoders, August 2024b. Comment: v2: new appendix H comparing kernel functions & bug-fixes to pseudo-code in Appendix J v3: further bug-fix to pseudo-code in Appendix J.