Same Concept, Different Directions: Cross-Modal Feature Heterogeneity in Sparse Autoencoders

Chungpa Lee [*1] **Jihoon Kwon** [*2] **Kyle Min** [3] **Jy-yong Sohn** [1]

Abstract

Vision–language models map images and text into a joint embedding space. However, these embeddings often entangle multiple semantic features, which limits their interpretability and controllability. While sparse autoencoders have emerged as a useful tool for decomposing these embeddings into monosemantic features, their application to joint embedding spaces has largely relied on an implicit, untested assumption that semantically corresponding features share the same directions across modalities. In this paper, we challenge this assumption by identifying discrepancies in feature directions for the same concept across image and text modalities, a phenomenon we term *cross-modal feature heterogeneity*. We demonstrate that this heterogeneity is a key driver of the *modality split*, where a shared concept activates different latents depending on the modality. This finding further reveals why aligning latent activations alone is insufficient to resolve the underlying feature mismatch. Motivated by this observation, we propose an approach that trains modality-specific sparse autoencoders to preserve each modality's feature geometry, and then aligns corresponding features post hoc. Our method improves reconstruction fidelity and enhances performance in cross-modal retrieval and concept steering.

1. Introduction

Vision-language models (VLMs) map images and text into a joint representation space, which supports a wide range of downstream tasks (Radford et al., 2021; Yu et al., 2022; Chen et al., 2024) and serves as the foundation for generative VLMs (Liu et al., 2023; Rombach et al., 2022). Despite their empirical success, the joint representation remains difficult to interpret (Oikarinen & Weng, 2023). These embeddings are typically *polysemantic*, as they encode multiple semantic concepts in a single vector, and it remains unclear how each concept is encoded and shared across modalities.

The linear representation hypothesis (Elhage et al., 2022; Park et al., 2024) provides a useful framework for studying this question, positing that each embedding can be expressed as a linear combination of a small number of *monosemantic* features. Sparse autoencoders (SAEs) operationalize this hypothesis by mapping each embedding to a *sparse latent code*, whose active coordinates indicate which monosemantic features are present in the embedding. Building on their success in unimodal settings (Huben et al., 2023; Templeton et al., 2024; Gao et al., 2025; Marks et al., 2025), recent work has applied SAEs to joint embedding spaces to recover monosemantic features shared across image and text modalities (Costa et al., 2026; Zaigrajew et al., 2025; Zhang et al., 2025; Papadimitriou et al., 2025; Pach et al., 2026; Kaushik et al., 2026; DHIMOÏLA et al., 2026).

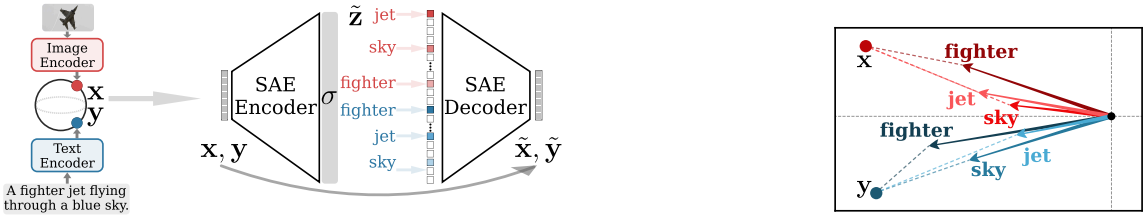
When SAEs are applied to joint embeddings of VLMs, prior work (Papadimitriou et al., 2025) observes a phenomenon known as *modality split*, where the same concept activates different latent coordinates across modalities. This split makes latent codes difficult to use across modalities, motivating prior approaches to align latent activations across modalities (Kaushik et al., 2026; DHIMOÏLA et al., 2026). However, these approaches treat modality split as a mismatch in latent activations, while implicitly assuming that semantically corresponding features share the same directions across modalities.

In this paper, we revisit this assumption by asking whether semantically corresponding feature directions are indeed aligned across modalities in the learned embedding space. We characterize *cross-modal feature heterogeneity*, a phenomenon in which the same semantic concept can be represented by different feature directions across image and text modalities, as illustrated in Figure 1.

Accounting for this heterogeneity is crucial as it redefines our interpretation of the modality split. When corresponding image and text features have different directions in the joint embedding space, an SAE naturally assigns them to

[*]Equal contribution [1]Yonsei University, Korea [2]Seoul National University, Korea [3]Oracle, USA. Correspondence to: Jy-yong Sohn <jysohn1108@yonsei.ac.kr>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea. 2026.
Copyright 2026 by the author(s).



(a) Modality split in sparse autoencoder

(b) Cross-modal feature heterogeneity

Figure 1. Illustration of modality split and cross-modal feature heterogeneity on joint embedding spaces. (a) An SAE encodes an image or text embedding ($\mathbf{x}$ or $\mathbf{y}$) from a VLM into a sparse latent code $\tilde{\mathbf{z}}$ and then reconstructs the original input as $\tilde{\mathbf{x}}$ or $\tilde{\mathbf{y}}$. Ideally, the same concept should activate the same coordinate in $\tilde{\mathbf{z}}$ across modalities. However, prior work observes the *modality split*, where the same concept (*e.g., sky*, *jet*, or *fighter*) activates different coordinates for images (red) and text (blue). (b) We show that this split is driven by *cross-modal feature heterogeneity*, where corresponding features (shown as arrows) fail to align directionally across modalities.

different latent coordinates to ensure precise reconstruction. Consequently, while forcing these latent activations to align across different modalities might mitigate the modality split, it risks degrading feature recovery by collapsing geometrically distinct directions into a single coordinate. This motivates our approach, which trains modality-specific SAEs to preserve each modality's feature geometry and then aligns corresponding features post hoc.

Our contributions are summarized below.

- In Section 4, we characterize *cross-modal feature heterogeneity*, where the same concept can have different feature directions across modalities in the joint embedding space.
- In Section 5, we show that this heterogeneity can explain *modality split* in multimodal SAEs and analyze how existing alignment approaches trade reconstruction quality for latent alignment.
- In Section 6, we propose an approach that preserves the unique feature geometry of each modality and aligns corresponding latent coordinates post hoc without sacrificing reconstruction quality.
- In Section 7, we show that our approach improves reconstruction fidelity and cross-modal retrieval, and enables more effective concept steering on real VLM embeddings.

2. Related Work

Mechanistic interpretability and linear representation hypothesis. Mechanistic interpretability aims to reverse-engineer neural networks, often by interpreting their learned representations as human-understandable mechanisms (Conmy et al., 2023; Elhage et al., 2021; Wang et al., 2023; Chughtai et al., 2023). A key assumption underlying this line of work is the *linear representation hypothesis* (Elhage et al., 2022; Park et al., 2024), which posits that each semantically coherent concept is encoded as a unique direction in representation space, and that embeddings can be expressed as linear combinations of such directions. This hypothesis is supported by diverse empirical evidence from linear probing (Alain & Bengio, 2017; Hewitt & Manning, 2019) and steering interventions (Li et al., 2023; Arditi et al., 2024; Rimsky et al., 2024). Recent work has begun extending this hypothesis to multimodal representations (Patashnik et al., 2021; Brack et al., 2023; Liu et al., 2025), opening avenues for understanding feature structure across modalities.

Sparse autoencoders for interpreting representations. SAEs have shown great promise for extracting monosemantic features from embeddings of language models (Huben et al., 2023; Templeton et al., 2024; Gao et al., 2025; Marks et al., 2025; Härle et al., 2025). Building on these advances, SAEs have been further applied to VLMs (Rao et al., 2024; Zaigrajew et al., 2025; Pach et al., 2026). Recent work has further examined how these features align or separate across modalities (Costa et al., 2026; Zaigrajew et al., 2025; Zhang et al., 2025; Papadimitriou et al., 2025; Pach et al., 2026), motivating an analysis of their underlying structural properties in joint embedding spaces.

Joint representations of vision-language models. A prominent phenomenon in joint VLM representations is the *modality gap*, where image and text embeddings occupy disjoint regions of the shared space (Liang et al., 2022). Prior work has analyzed this gap from multiple geometric, distributional, and optimization-level perspectives (Liang et al., 2022; Zhang et al., 2024; Schrodi et al., 2025). One line of work argues that the modality gap partly reflects modality-specific features in multimodal representations, rather than merely a failure of alignment (Jiang et al., 2023; Ramasinghe et al., 2024; Qian et al., 2023). Another line focuses on cross-modal misalignment of shared features and studies how reducing such mismatch improves downstream task performance (Eslami & de Melo, 2025; Yamaguchi et al., 2025; Grassucci et al., 2026; Mistretta et al., 2025).

Recent SAE-based analyses further reveal *modality split* (Papadimitriou et al., 2025), where the same concept activates different latent codes across modalities. Existing remedies force shared latent activations (Kaushik et al., 2026; DHIMOÏLA et al., 2026), implicitly assuming feature directions are aligned (*i.e.,* $\phi_i = \psi_i$) and overlooking directional

misalignment. When shared concepts are encoded along distinct directions, forcing a single shared SAE inevitably degrades reconstruction. We instead preserve modality-specific directions by using modality-specific SAEs and align the corresponding features post hoc.

3. Preliminaries

We formalize joint embeddings as linear combinations of monosemantic features. We then introduce SAEs as a framework for recovering feature directions from embeddings.

Embeddings as linear combinations of features. Following the linear representation hypothesis (Arora et al., 2018; Elhage et al., 2022; Park et al., 2024; Cui et al., 2026), we assume that VLM embeddings $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ are polysemantic, meaning that each embedding can be expressed as a linear combination of monosemantic features. A monosemantic feature is a direction in the joint embedding space that represents a semantically coherent concept.

Formally, let n be the number of latent concepts represented in the joint embedding space, and define $[n] := \{1, 2, \cdots, n\}$. We assume the existence of a latent code $\mathbf{z} := (z_1 \ \cdots \ z_n)^\top \in \mathbb{R}_+^n$ and *feature matrices* $\mathbf{\Phi} := (\boldsymbol{\phi}_1 \ \cdots \ \boldsymbol{\phi}_n) \in \mathbb{R}^{d\times n}$ for images and $\mathbf{\Psi} := (\boldsymbol{\psi}_1 \ \cdots \ \boldsymbol{\psi}_n) \in \mathbb{R}^{d\times n}$ for text. The column vectors $\boldsymbol{\phi}_i, \boldsymbol{\psi}_i \in \mathbb{R}^d$ denote the *feature direction vectors* of the i-th latent concept for images and text, respectively. We assume that all feature directions have unit norm.

Under the linear representation hypothesis, an image embedding is written as $\mathbf{x} = \sum_{i\in[n]} z_i \boldsymbol{\phi}_i$, where each activation value $z_i \in \mathbb{R}_+$ indicates the strength of the corresponding feature, whereas $z_i = 0$ indicates that it is inactive. When an image–text pair is considered, we assume that the two embeddings share the same latent code $\mathbf{z}$, while their feature directions may differ across modalities:

$$\mathbf{x} = \mathbf{\Phi}\mathbf{z} = \textstyle\sum_{i\in[n]} z_i \boldsymbol{\phi}_i, \quad \mathbf{y} = \mathbf{\Psi}\mathbf{z} = \textstyle\sum_{i\in[n]} z_i \boldsymbol{\psi}_i. \tag{1}$$

Unlike prior work (Kaushik et al., 2026), we impose no constraint such as $\boldsymbol{\phi}_i = \boldsymbol{\psi}_i$. Thus, we allow $\boldsymbol{\phi}_i \neq \boldsymbol{\psi}_i$, capturing the possibility that the same concept is encoded along different directions across modalities.

We assume that the coordinates of the latent code $\mathbf{z}$ are independent and sparse. Specifically, the sparsity parameter s is defined as the probability that each coordinate is inactive,

$$\Pr(z_i = 0) = s, \qquad i \in [n], \tag{2}$$

where s is close to 1. This implies that only a small number of features are active in each embedding.

Sparse autoencoder. SAEs encode each embedding into a sparse latent code and decode this code to reconstruct the original embedding (Huben et al., 2023; Templeton et al., 2024; Gao et al., 2025; Rao et al., 2024; Zaigrajew et al., 2025; Cui et al., 2026). Ideally, each active coordinate of the latent code represents the strength of a monosemantic feature in the embedding, while the corresponding decoder column provides its feature direction.

Specifically, let m be the latent dimension of the SAE, and let $\mathbf{W} := [\mathbf{w}_1 \cdots \mathbf{w}_m] \in \mathbb{R}^{d\times m}$ denote its weight matrix. For simplicity, we omit bias terms and use the same weight matrix for the encoder and decoder. The encoder produces an *estimated latent code* $\tilde{\mathbf{z}} := (\tilde{z}_1 \ \cdots \ \tilde{z}_m)^\top \in \mathbb{R}_+^m$. For image and text embeddings, we write

$$\tilde{\mathbf{z}}(\mathbf{x}) := \sigma(\mathbf{W}^\top \mathbf{x}), \qquad \tilde{\mathbf{z}}(\mathbf{y}) := \sigma(\mathbf{W}^\top \mathbf{y}), \tag{3}$$

where σ denotes an activation function[1] that induces sparsity. The decoder then produces

$$\begin{aligned} \tilde{\mathbf{x}}(\mathbf{x}) &:= \mathbf{W}\tilde{\mathbf{z}}(\mathbf{x}) = \textstyle\sum_{j\in[m]} \tilde{z}_j(\mathbf{x})\, \mathbf{w}_j, \\ \tilde{\mathbf{y}}(\mathbf{y}) &:= \mathbf{W}\tilde{\mathbf{z}}(\mathbf{y}) = \textstyle\sum_{j\in[m]} \tilde{z}_j(\mathbf{y})\, \mathbf{w}_j. \end{aligned} \tag{4}$$

The SAE is trained to reconstruct embeddings from both modalities by minimizing

$$\begin{aligned} &\mathcal{L}_{\text{rec}}(\mathbf{W}; \mathbf{\Phi}) + \mathcal{L}_{\text{rec}}(\mathbf{W}; \mathbf{\Psi}) \\ &\quad := \mathbb{E}_{\mathbf{x}} \|\mathbf{x} - \tilde{\mathbf{x}}(\mathbf{x})\|_2^2 + \mathbb{E}_{\mathbf{y}} \|\mathbf{y} - \tilde{\mathbf{y}}(\mathbf{y})\|_2^2, \end{aligned} \tag{5}$$

where the first term denotes the reconstruction loss for image embeddings generated by $\mathbf{\Phi}$, and the second term denotes the reconstruction loss for text embeddings generated by $\mathbf{\Psi}$, as specified in (1). More explicitly, $\mathcal{L}_{\text{rec}}(\mathbf{W}; \mathbf{\Phi}) := \mathbb{E}_{\mathbf{x}} \|\mathbf{x} - \tilde{\mathbf{x}}(\mathbf{x})\|_2^2 = \mathbb{E}_{\mathbf{z}} \left\|\mathbf{\Phi}\mathbf{z} - \mathbf{W}\sigma\big(\mathbf{W}^\top \mathbf{\Phi}\mathbf{z}\big)\right\|_2^2$. After training the SAE, the encoder extracts latent codes $\tilde{\mathbf{z}}$, and each decoder column $\mathbf{w}_j$ provides an estimate of a feature direction (Cui et al., 2026).

Concept alignment across modalities in sparse autoencoders. When an SAE is trained on embeddings from both modalities, it is expected to *recover monosemantic features* within each modality and to *align shared concepts* by assigning corresponding image and text features to shared latent coordinates. However, each latent index j is tied to a single column $\mathbf{w}_j$ of the decoder weight matrix in (4). Assigning corresponding features to the same coordinate implicitly assumes that they share a common direction in the joint embedding space. We first examine whether this assumption holds in learned joint embedding spaces.

[1] For theory, we use the Top-1 operator, whose behavior can be approximated by Top-K activations with bias terms. In experiments, we use SAEs with Top-K activations for $K > 1$, bias terms, and separate encoder and decoder weights.

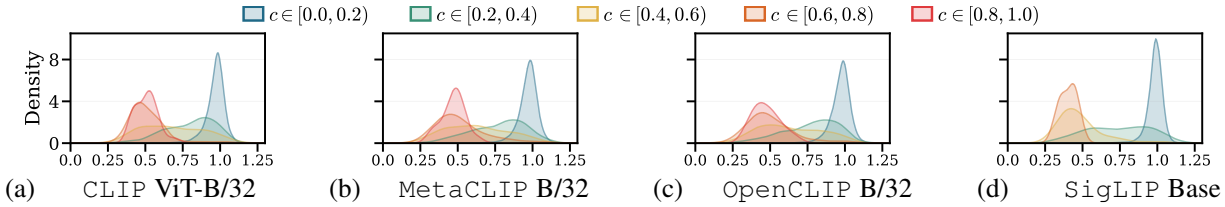


Figure 2. Distribution of cosine distances $d_{\cos}$ between image–text feature pairs estimated from embeddings of four VLMs. We group image–text feature pairs $(\hat{\phi}_i, \hat{\psi}_j)$ by their coactivation correlation $c_{i,j}$ in (6), shown in different colors. The value $c_{i,j}$ measures how strongly the i-th image feature and the j-th text feature coactivate on image–text embeddings, so pairs with larger correlations are more likely to represent the same shared concept. Across all models, the distribution of distances $d_{\cos}$ remains centered around a positive value, even for high-correlation pairs ($c \geq 0.8$, shown in blue). This observation supports the presence of cross-modal feature heterogeneity.

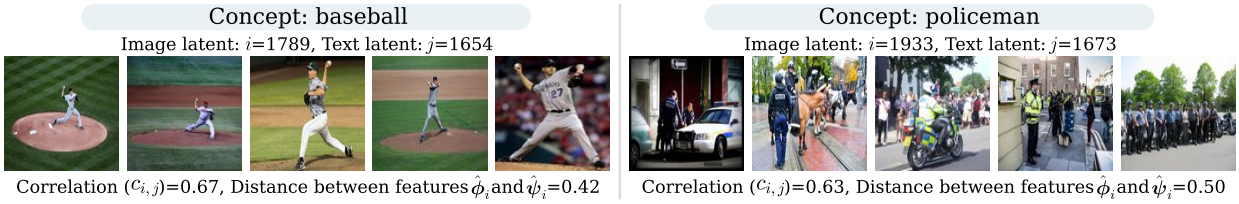


Figure 3. Image examples with high coactivation correlation $c_{i,j}$ from the CLIP setting in Figure 2, where i and j denote the image and text latent indices, respectively. The images reveal coherent concepts such as *baseball* and *policeman*, but the corresponding features $\hat{\phi}_i$ and $\hat{\psi}_j$ remain directionally separated, with feature distances of 0.42 and 0.50, respectively.

4. Cross-Modal Feature Heterogeneity in Joint Embedding Spaces

We examine whether semantically corresponding features share a common direction across image and text modalities in the joint embedding space. When they do not, we refer to this directional mismatch as *cross-modal feature heterogeneity*, which we formally define below.

Definition 1 (Cross-Modal Feature Heterogeneity)**.** *The i-th latent concept exhibits cross-modal feature heterogeneity if its image and text feature directions are not perfectly aligned, i.e., $\phi_i \neq \psi_i$.*

This definition separates semantic correspondence from directional identity in the joint embedding space. An image feature and a text feature may represent the same concept while occupying different directions in the joint embedding space. We next examine whether such directional differences appear in joint embedding spaces learned by VLMs.

Measuring differences in feature directions across modalities. Since feature directions are unobserved, we use decoder columns as their estimates (Cui et al., 2026), letting $\hat{\phi}_i$ and $\hat{\psi}_j$ denote the i-th and j-th decoder columns of SAEs trained on image and text embeddings, respectively.

The correspondences between image and text features are also unknown. We therefore use coactivation correlations on paired image and text embeddings as a proxy for semantic correspondence. Using the encoders of the SAEs trained on image and text embeddings, we compute latent codes $\tilde{\mathbf{z}}(\mathbf{x})$ and $\tilde{\mathbf{z}}(\mathbf{y})$ for paired embeddings $(\mathbf{x}, \mathbf{y})$ and form their correlation matrix over the training set:

$$\mathbf{C} := \mathrm{Corr}\big(\tilde{\mathbf{z}}(\mathbf{x}), \tilde{\mathbf{z}}(\mathbf{y})\big) \in \mathbb{R}^{m \times m}, \qquad c_{i,j} := [\mathbf{C}]_{i,j} \quad (6)$$

Each entry $c_{i,j}$ measures how strongly the i-th image latent and the j-th text latent coactivate on paired inputs. Larger values are treated as indicating more likely semantic correspondence. To test whether such likely corresponding features are also directionally aligned in the joint embedding space, we measure the distance between their estimated feature directions. Specifically, for each pair (i, j), we compute the cosine distance between $\hat{\phi}_i$ and $\hat{\psi}_j$, defined as $d_{\cos}(\hat{\phi}_i, \hat{\psi}_j) := 1 - \cos(\hat{\phi}_i, \hat{\psi}_j) \in [0, 2]$.

Figure 2 reports the distribution of cosine distances $d_{\cos}$ between estimated image and text feature directions, grouped by coactivation correlation c in (6). The results are computed on MS-COCO (Lin et al., 2014) embeddings extracted from four VLMs, including CLIP (Radford et al., 2021), MetaCLIP (Chuang et al., 2026), OpenCLIP (Cherti et al., 2023), and SigLIP2 (Tschannen et al., 2025). Feature pairs with larger coactivation correlation tend to have smaller distances, but they do not concentrate near zero distance. This provides empirical support for cross-modal feature heterogeneity, where semantically corresponding image and text features need not share identical directions in the joint embedding space. See Appendix D.1 for details.

Figure 3 shows examples where matched image and text latents strongly coactivate and respond to the same concept, such as *baseball*. In this case, the image latent with index $i = 1789$ and the text latent with index $j = 1654$ have a high coactivation correlation, $c_{i,j} = 0.67$. But the distance between the estimated features remains large, with $d_{\cos}(\hat{\phi}_i, \hat{\psi}_j) = 0.42$. These examples show that semantic

correspondence need not imply directional identity.

Appendix B discusses possible sources of this heterogeneity and relates it to the modality gap, showing that a nontrivial modality gap implies cross-modal feature heterogeneity.

5. Analysis of Sparse Autoencoders under Cross-Modal Feature Heterogeneity

We analyze whether an SAE can reconstruct image and text feature directions while assigning corresponding concepts to shared latent coordinates. Under cross-modal feature heterogeneity, these two goals conflict. Corollary 1 explains why reconstruction alone induces modality split, Theorem 1 shows how limited latent capacity collapses distinct feature directions, and Propositions 1 and 2 show why existing alignment losses reduce modality split only by sacrificing reconstruction quality. Proofs are provided in Appendix C.

Why reconstruction induces modality split. We consider an SAE trained with the reconstruction loss in (5). Since no alignment term is used, the objective only encourages accurate reconstruction.

Corollary 1. *Suppose the SAE in (3) and (4) has $m \geq 2n$ latent coordinates. Consider the loss $\mathcal{L}_{\mathrm{rec}}(\mathbf{W};\mathbf{\Phi}) + \mathcal{L}_{\mathrm{rec}}(\mathbf{W};\mathbf{\Psi})$ in (5). As the sparsity parameter s in (2) approaches 1, the loss is minimized to leading order by taking $\hat{\mathbf{W}} := [\mathbf{\Phi} \;\; \mathbf{\Psi} \;\; \mathbf{0}_{d\times(m-2n)}]\,\mathbf{P}$ for any permutation matrix $\mathbf{P} \in \mathbb{R}^{m\times m}$, and $\mathcal{L}_{\mathrm{rec}}(\hat{\mathbf{W}};\mathbf{\Phi})+\mathcal{L}_{\mathrm{rec}}(\hat{\mathbf{W}};\mathbf{\Psi}) = o(1-s)$.*

Corollary 1 shows that when the SAE has enough latent coordinates, the reconstruction loss is minimized to leading order by assigning a separate column of the decoder weight matrix $\mathbf{W}$ to each feature direction. Thus, even if the image feature $\boldsymbol{\phi}_i$ and the text feature $\boldsymbol{\psi}_i$ represent the same concept, the reconstruction objective does not force them to share a latent coordinate when they point in different directions. Instead, the SAE can reconstruct them by assigning them to different columns, say $\hat{\mathbf{w}}_a$ and $\hat{\mathbf{w}}_b$, which activate different latent coordinates, $\tilde{z}_a$ and $\tilde{z}_b$. This provides a mechanism for modality split: the same semantic concept appears at different latent indices across modalities because accurate reconstruction favors preserving distinct feature directions.

Why limited capacity collapses distinct features. Corollary 1 assumes enough latent coordinates to represent all feature directions separately, namely $m \geq 2n$. In practice, the effective number of coordinates can be smaller due to dead neurons (Templeton et al., 2024; Gao et al., 2025; Lu et al., 2020), *i.e.,* $\tilde{z}_j(\mathbf{x}) = 0$ for all $\mathbf{x}$. We therefore analyze the regime where the SAE has fewer latent coordinates than feature directions, $m < 2n$.

Theorem 1. *Suppose the SAE in (3) and (4) has $m < 2n$ latent coordinates. Consider the loss $\mathcal{L}_{\mathrm{rec}}(\mathbf{W};\mathbf{\Phi}) + \mathcal{L}_{\mathrm{rec}}(\mathbf{W};\mathbf{\Psi})$ in (5). Define $\mathbf{M}_i := \mathbb{E}\big[z_i^2 \mid z_i \neq 0\big]\,\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top \in \mathbb{R}^{d\times d}$ and $\mathbf{M}_{n+i} := \mathbb{E}\big[z_i^2 \mid z_i \neq 0\big]\,\boldsymbol{\psi}_i\boldsymbol{\psi}_i^\top \in \mathbb{R}^{d\times d}$ for $i \in [n]$. Let $(\mathbb{A}_1,\cdots,\mathbb{A}_m)$ be a partition of $[2n]$ that maximizes $\sum_{j\in[m]} \lambda_{\max}\Big(\sum_{i\in\mathbb{A}_j}\mathbf{M}_i\Big)$, where $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue. As the sparsity parameter s in (2) approaches 1, the loss is minimized to leading order by taking $\hat{\mathbf{W}} := [\hat{\mathbf{w}}_1 \;\cdots\; \hat{\mathbf{w}}_m]$, where each $\hat{\mathbf{w}}_j$ is a unit-norm top eigenvector of $\sum_{i\in\mathbb{A}_j}\mathbf{M}_i$. Moreover, if no two directions among $\{\boldsymbol{\phi}_i\}_{i\in[n]} \cup \{\boldsymbol{\psi}_i\}_{i\in[n]}$ are collinear, $\mathcal{L}_{\mathrm{rec}}(\hat{\mathbf{W}};\mathbf{\Phi})+\mathcal{L}_{\mathrm{rec}}(\hat{\mathbf{W}};\mathbf{\Psi}) > 0$ for s sufficiently close to 1.*

Theorem 1 analyzes the regime $m < 2n$, where the SAE has fewer latent coordinates than the total number of image and text feature directions. In this regime, some distinct feature directions must share a latent coordinate. The partition $(\mathbb{A}_1,\ldots,\mathbb{A}_m)$ describes this assignment over the set of all image and text feature directions. More precisely, all feature directions whose indices belong to the same set $\mathbb{A}_j$ are represented by the same column $\hat{\mathbf{w}}_j$ and therefore activate the same coordinate $\tilde{z}_j$.

For each group $\mathbb{A}_j$, the column $\hat{\mathbf{w}}_j$ is chosen as the best single direction for representing the directions in that group, namely the unit-norm top eigenvector of $\sum_{i\in\mathbb{A}_j}\mathbf{M}_i$, where each $\mathbf{M}_i$ is the weighted outer product of the corresponding feature direction. Thus, the partition is determined by which feature directions can be well represented by a common decoder column. With equal weights $\mathbb{E}[z_i^2 \mid z_i \neq 0]$, feature directions that are close in angle tend to be grouped together.

Consequently, two feature directions that are close in the embedding space may be represented by the same column $\hat{\mathbf{w}}_j$ and activate the same latent coordinate $\tilde{z}_j$, even if they correspond to different concepts. In a favorable case, if $\boldsymbol{\phi}_i$ and $\boldsymbol{\psi}_i$ represent the same concept and are also close in direction, limited capacity may group them into the same coordinate, reducing modality split. However, as discussed in Section 4, semantic correspondence does not guarantee directional alignment. Therefore, limited capacity alone does not provide a reliable solution to modality split, and an additional alignment procedure is needed to identify corresponding concepts across modalities.

Why alignment during training sacrifices reconstruction. The previous results suggest that reducing modality split requires placing corresponding image and text features on a shared latent coordinate. Existing methods encourage this behavior by adding an auxiliary alignment loss during training (DHIMOÏLA et al., 2026; Kaushik et al., 2026). We analyze how such losses improve alignment at the cost of reconstruction quality.

We characterize global minimizers $\hat{\mathbf{W}}$ for two alignment objectives, the group-sparse loss (Kaushik et al., 2026) and the Iso-Energy alignment loss (DHIMOÏLA et al., 2026).

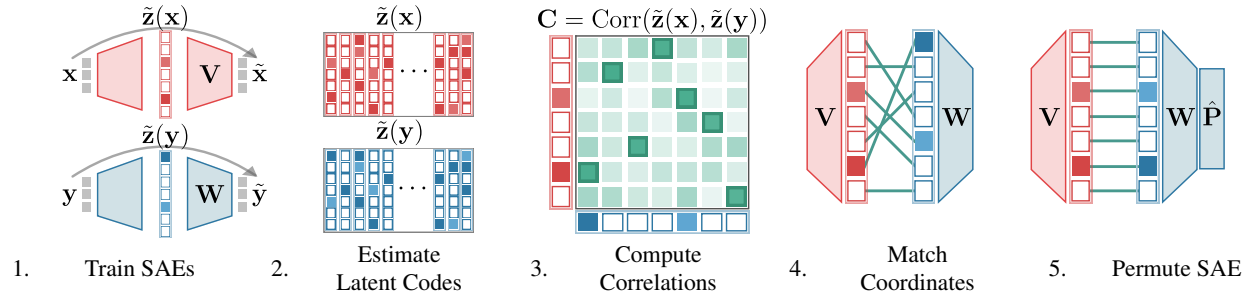


Figure 4. Overview of our approach. We first train modality-specific SAEs to reconstruct image and text embeddings, obtaining latent codes $\tilde{\mathbf{z}}(\mathbf{x})$ and $\tilde{\mathbf{z}}(\mathbf{y})$. We then compute their correlation matrix $\mathbf{C}$ in (6) over paired training data and apply the Hungarian algorithm to obtain the assignment $\hat{\mathbf{P}}$ in (7). Finally, we reindex the text latent coordinates using $\hat{\mathbf{P}}$, aligning corresponding image and text features to shared latent indices without altering the learned feature directions.

Each objective adds an auxiliary term weighted by λ, which controls the strength of alignment. Because the objectives are symmetric in the image and text modalities, the following propositions report only $\mathcal{L}_{\text{rec}}(\hat{\mathbf{W}};\boldsymbol{\phi})$, with the same value holding for $\mathcal{L}_{\text{rec}}(\hat{\mathbf{W}};\boldsymbol{\psi})$. In both cases, the form of the global minimizer changes at critical values of λ.

Proposition 1. *Consider the case $n = 1$, where the feature directions $\boldsymbol{\phi}, \boldsymbol{\psi} \in \mathbb{R}^d$ satisfy $\rho := \boldsymbol{\phi}^\top \boldsymbol{\psi} \in (0, 1)$. Suppose the SAE in (3) and (4) has $m \geq 2$ latent coordinates. Consider the loss (Kaushik et al., 2026)*

$$\mathcal{L}_{\text{rec}}(\mathbf{W};\boldsymbol{\phi}) + \mathcal{L}_{\text{rec}}(\mathbf{W};\boldsymbol{\psi}) + \lambda\, \mathbb{E}_z\Big[\textstyle\sum_{j\in[m]} \sqrt{\big[\sigma(\mathbf{W}^\top\boldsymbol{\phi}\, z)\big]_j^2 + \big[\sigma(\mathbf{W}^\top\boldsymbol{\psi}\, z)\big]_j^2}\Big].$$

Define $\lambda^\star(\rho) := \frac{(1-\rho)\mathbb{E}[z^2]}{(2-\sqrt{1+\rho})\cdot\,\mathbb{E}[z]}$ *and* $\lambda^{\star\star}(\rho) := \frac{\sqrt{1+\rho}\cdot\,\mathbb{E}[z^2]}{\mathbb{E}[z]}$. *Up to a permutation matrix $\mathbf{P}$, a global minimizer $\hat{\mathbf{W}}$ is:*

- *If $\lambda < \lambda^\star(\rho)$, then $\hat{\mathbf{W}} = [\boldsymbol{\phi}\ \boldsymbol{\psi}\ \mathbf{0}_{d\times(m-2)}]\,\mathbf{P}$, where $\mathcal{L}_{\text{rec}}(\hat{\mathbf{W}};\boldsymbol{\phi}) = 0$.*
- *If $\lambda \in (\lambda^\star(\rho), \lambda^{\star\star}(\rho))$, then $\hat{\mathbf{W}} = \big[\frac{\boldsymbol{\phi}+\boldsymbol{\psi}}{\|\boldsymbol{\phi}+\boldsymbol{\psi}\|}\ \mathbf{0}_{d\times(m-1)}\big]\,\mathbf{P}$, where $\mathcal{L}_{\text{rec}}(\hat{\mathbf{W}};\boldsymbol{\phi}) = \frac{1-\rho}{2}\,\mathbb{E}[z^2]$.*
- *If $\lambda > \lambda^{\star\star}(\rho)$, then $\hat{\mathbf{W}}$ satisfies $\sigma(\hat{\mathbf{W}}^\top\boldsymbol{\phi}) = \sigma(\hat{\mathbf{W}}^\top\boldsymbol{\psi}) = \mathbf{0}$, where $\mathcal{L}_{\text{rec}}(\hat{\mathbf{W}};\boldsymbol{\phi}) = \mathbb{E}[z^2]$.*

Proposition 2. *Consider the case $n = 1$, where the feature directions $\boldsymbol{\phi}, \boldsymbol{\psi} \in \mathbb{R}^d$ satisfy $\rho := \boldsymbol{\phi}^\top \boldsymbol{\psi} \in (0, 1)$. Suppose the SAE in (3) and (4) has $m \geq 2$ latent coordinates. Consider the loss (DHIMOÏLA et al., 2026)*

$$\mathcal{L}_{\text{rec}}(\mathbf{W};\boldsymbol{\phi}) + \mathcal{L}_{\text{rec}}(\mathbf{W};\boldsymbol{\psi}) - \lambda\mathbb{E}_z\Big[\big(\sigma(\mathbf{W}^\top\boldsymbol{\phi}z)\big)^\top\sigma(\mathbf{W}^\top\boldsymbol{\psi}z)\Big].$$

Define $\lambda^\star(\rho) := \frac{2(1-\rho)}{1+\rho}$. *Up to a permutation matrix $\mathbf{P}$, a global minimizer $\hat{\mathbf{W}}$ is:*

- *If $\lambda < \lambda^\star(\rho)$, then $\hat{\mathbf{W}} = [\boldsymbol{\phi}\ \boldsymbol{\psi}\ \mathbf{0}_{d\times(m-2)}]\,\mathbf{P}$, where $\mathcal{L}_{\text{rec}}(\hat{\mathbf{W}};\boldsymbol{\phi}) = 0$.*
- *If $\lambda > \lambda^\star(\rho)$, then $\hat{\mathbf{W}} = \big[\frac{\boldsymbol{\phi}+\boldsymbol{\psi}}{\|\boldsymbol{\phi}+\boldsymbol{\psi}\|}\ \mathbf{0}_{d\times(m-1)}\big]\mathbf{P}$, where $\mathcal{L}_{\text{rec}}(\hat{\mathbf{W}};\boldsymbol{\phi}) = \frac{1-\rho}{2}\,\mathbb{E}[z^2]$.*

Propositions 1 and 2 illustrate the trade-off between reconstruction and latent alignment. For both losses, a small λ keeps the image and text feature directions in two separate decoder columns, $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$, so they remain on distinct latent coordinates. Once λ exceeds $\lambda^\star(\rho)$, the two directions are represented by a single shared column, $\frac{\boldsymbol{\phi}+\boldsymbol{\psi}}{\|\boldsymbol{\phi}+\boldsymbol{\psi}\|}$, placing both on one coordinate. This enhances latent alignment but reduces reconstruction quality, with $\mathcal{L}_{\text{rec}}(\hat{\mathbf{W}};\boldsymbol{\phi}) = \frac{1-\rho}{2}\mathbb{E}[z^2]$. For the group-sparse loss, an even larger λ beyond $\lambda^{\star\star}(\rho)$ drives the SAE to a degenerate solution where both features become inactive.

6. Modality-Specific Sparse Autoencoders and Post-Hoc Alignment

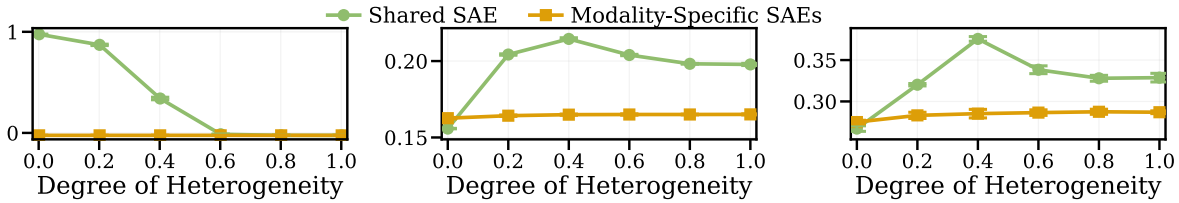
Building on Section 5, we propose a two-stage method. First, we train separate SAEs for image and text embeddings to preserve modality-specific feature directions. Then, we align their latent coordinates using activation correlations. Figure 4 provides an overview of this procedure.

Modality-specific sparse autoencoders. We train separate SAEs for image and text embeddings, with weight matrices $\mathbf{V}$ and $\mathbf{W}$, respectively, using the reconstruction loss in (5). We minimize $\mathcal{L}_{\text{rec}}(\mathbf{V};\boldsymbol{\Phi})$ and $\mathcal{L}_{\text{rec}}(\mathbf{W};\boldsymbol{\Psi})$ separately, without any cross-modal alignment constraint. This prevents image and text feature directions from competing for the same columns of the weight matrix, allowing each modality to preserve its own directions before cross-modal correspondences are identified.

Post-hoc alignment. Because the latent coordinates of the two SAEs are not naturally aligned, we use the correlation matrix $\mathbf{C}$ in (6) to identify corresponding image and text latents. We find a permutation matrix $\hat{\mathbf{P}}$ that maximizes the total correlation between aligned coordinates[2],

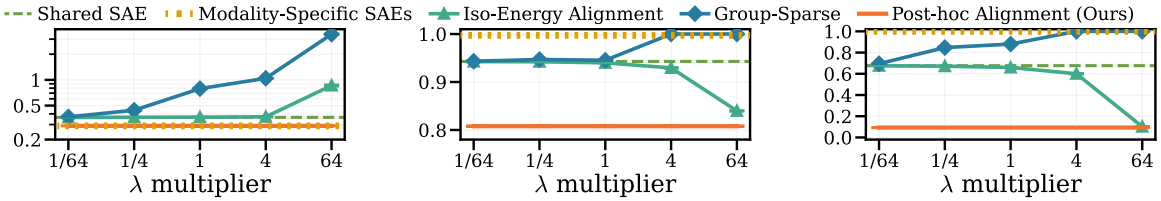
$$\hat{\mathbf{P}} \in \arg\max_{\mathbf{P}\in\mathcal{P}_m} \operatorname{tr}(\mathbf{C}\mathbf{P}), \tag{7}$$

[2]Dead latent coordinates are excluded in practice. Thus, (7) is applied to the corresponding submatrix of $\mathbf{C}$.



(a) Feature collapse rate (↓) *(b)* Reconstruction error (↓) *(c)* Feature recovery error (↓)

Figure 5. Comparison between a shared SAE trained on both modalities (green) and modality-specific SAEs trained separately for each modality (yellow) as the cross-modal feature distance $d_{\cos}$ varies. Each point corresponds to SAEs trained on synthetic embeddings sampled from ground-truth feature matrices $\mathbf{\Phi}$ and $\mathbf{\Psi}$, where the x-axis shows the cosine distance $d_{\cos}$ between corresponding image and text feature directions. When corresponding directions are distinct but geometrically close, the shared SAE tends to collapse them into a single learned direction, leading to higher reconstruction and feature recovery errors. In contrast, modality-specific SAEs avoid this collapse and better preserve feature directions under cross-modal feature heterogeneity, which corresponds to nonzero distance.



(a) Feature recovery error (↓) *(b)* Latent alignment error (↓) *(c)* Feature alignment error (↓)

Figure 6. Comparison between Post-hoc Alignment method (orange) and two auxiliary-loss baselines: Iso-Energy alignment (DHIMOÏLA et al., 2026) (teal) and group-sparse loss (Kaushik et al., 2026) (blue). The weight λ of the auxiliary loss is shown as a multiplier of the default value in the original papers. For reference, we also include the shared SAE (green) and modality-specific SAEs (yellow), neither of which depend on λ. Our method achieves both lower feature recovery error and lower alignment error than the baselines.

where $\mathcal{P}_m$ denotes the set of all $m \times m$ permutation matrices. The trace sums the correlations of image and text coordinates matched by $\mathbf{P}$. We compute this assignment using the Hungarian algorithm (Kuhn, 1955). We then apply the permutation only to the SAE for the text modality as

$$\tilde{\mathbf{z}}(\mathbf{y};\mathbf{W}) := \sigma\big(\hat{\mathbf{P}}^\top \mathbf{W}^\top \mathbf{y}\big), \quad \tilde{\mathbf{y}}(\mathbf{y};\mathbf{W}) := \mathbf{W}\hat{\mathbf{P}}\tilde{\mathbf{z}}(\mathbf{y};\mathbf{W}). \tag{8}$$

This alignment simply permutes the learned features of the text modality without changing their directions. Consequently, it preserves the feature geometry of each modality while assigning corresponding features from the image and text modalities to the same coordinates in the latent codes.

Inference. Embeddings are encoded into latent codes using modality-specific SAEs, and the permutation is applied to align their indices as in (8). This alignment enables cross-modal tasks such as retrieval in the aligned latent space, while encoding and decoding remain modality-specific.

7. Experiments

This section validates the theoretical findings in Section 5 and evaluates the approach proposed in Section 6. Section 7.1 uses synthetic embeddings where ground-truth feature directions and cross-modal correspondences are known, allowing us to directly test the theoretical predictions. Section 7.2 evaluates our method on real-world datasets to assess its effectiveness on cross-modal tasks.

7.1. Validation on Synthetic Embeddings

We validate two key design choices of the method proposed in Section 6. First, we test whether training modality-specific SAEs better preserves image and text feature directions than training a shared SAE. Second, we test whether post-hoc alignment can align corresponding latent coordinates without sacrificing feature recovery. These design choices are motivated by the analysis in Section 5.

Setup. To test the first choice, we compare a *shared SAE*, which uses one weight matrix for both modalities, with *modality-specific SAEs*, which use separate weight matrices for image and text embeddings. For a fair comparison, we match the total number of trainable parameters by using $m/2$ latent coordinates for each modality-specific SAE when the shared SAE uses m. Thus, any performance difference reflects how capacity is distributed across modalities, rather than model size.

To test the second choice, we compare our post-hoc alignment method with two auxiliary-loss baselines (DHIMOÏLA et al., 2026; Kaushik et al., 2026). We report the average over three independent runs. Details on synthetic embedding generation and training setup are provided in Appendix D.2.

Metrics. We evaluate each method along three axes: *Reconstruction Error* and *Feature Recovery Error* for reconstruction quality; *Latent Alignment Error* and *Feature Align-*

Table 1. Reconstruction and alignment quality of various methods. Reconstruction quality is measured by mean squared error between the input embedding and its reconstruction. Cross-modal alignment is evaluated by Recall@k for $k \in \{1, 5, 10\}$ on image-to-text and text-to-image retrieval tasks on `MS-COCO` (Lin et al., 2014), and by top-1 zero-shot classification accuracy on `ImageNet1K` (Deng et al., 2009). The best and second-best results for each metric are shown in bold and underlined, respectively.

	MS-COCO (Lin et al., 2014)							ImageNet1K (Deng et al., 2009)	
	Recon. (↓)	Image-to-Text (↑)			Text-to-Image (↑)			Recon. (↓)	Zero-shot (↑)
Methods	MSE	Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10	MSE	Accuracy
Shared SAE	0.090	6.1 (±0.3)	13.3 (±0.8)	17.8 (±0.8)	3.4 (±0.2)	9.5 (±0.5)	14.0 (±0.8)	0.118	15.7 (±1.3)
+ Iso-Energy alignment loss	0.091	4.7 (±1.9)	10.6 (±3.4)	14.6 (±4.4)	2.7 (±0.7)	7.4 (±2.1)	10.8 (±3.0)	0.118	13.0 (±2.3)
+ Group-sparse loss	0.105	7.1 (±0.1)	16.7 (±0.4)	23.8 (±0.3)	4.3 (±0.1)	12.2 (±0.1)	18.3 (±0.2)	0.134	**26.6** (±0.3)
Modality-Specific SAEs	**0.089**	0.0 (±0.0)	0.1 (±0.1)	0.2 (±0.1)	0.0 (±0.0)	0.1 (±0.0)	0.1 (±0.1)	**0.116**	0.1 (±0.0)
+ Post-hoc Alignment (Ours)	**0.089**	**16.0** (±0.5)	**34.1** (±1.4)	**44.5** (±1.6)	**11.4** (±0.4)	**27.2** (±1.2)	**37.0** (±1.3)	**0.116**	25.1 (±0.3)

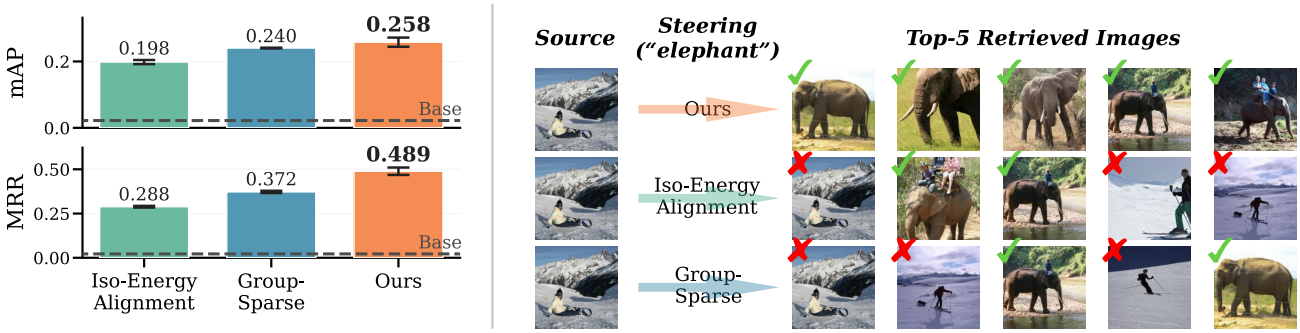


Figure 7. Image retrieval under concept latent steering on `MS-COCO`. We steer source image embeddings using target feature directions identified through aligned latent coordinates. Retrieval performance is measured by mean average precision (mAP) and mean reciprocal rank (MRR) over retrieved images containing the target concept. Qualitative examples show source images and their top retrieved images after steering. Our method steers images toward target concepts more effectively.

ment Error for cross-modal alignment; and *Feature Collapse Rate* for whether same-concept features across modalities are collapsed into a single learned direction. See Appendix D.3 for formal definitions.

Modality-specific sparse autoencoders better preserve feature directions under heterogeneity. Figure 5 reports the results obtained by varying the cosine distance $d_{\cos}$ between corresponding cross-modal features $\boldsymbol{\phi}_i$ and $\boldsymbol{\psi}_i$. In Figure 5a, the shared SAE exhibits feature collapse when image and text feature directions are geometrically close but not identical. Modality-specific SAEs avoid this collapse because each modality has its own SAE. Figures 5b and 5c show that this collapse leads to larger reconstruction and feature recovery errors. This is consistent with the analysis in Section 5, where representing different feature directions with a single decoder column increases reconstruction loss. At zero distance ($d_{\cos} = 0$), the image and text feature directions coincide ($\forall i \in [n], \boldsymbol{\phi}_i = \boldsymbol{\psi}_i$), so there is no cross-modal feature heterogeneity. In this idealized case, the shared SAE can represent the same concept with a single coordinate, which explains its lower reconstruction and recovery errors. Overall, these results show that modality-specific SAEs better preserve feature directions when semantically corresponding features differ across modalities, leading to better feature recovery.

Post-hoc alignment improves alignment without sacrificing feature recovery. Figure 6 reports the results obtained by varying the auxiliary-loss weight λ. We fix the cosine distance between corresponding cross-modal features at 0.5 to reflect the level of cross-modal feature heterogeneity observed in Figure 2. For the Iso-Energy alignment loss (DHIMOÏLA et al., 2026), increasing λ improves cross-modal alignment but degrades feature recovery, consistent with the reconstruction-alignment trade-off shown in Proposition 2. For the group-sparse loss (Kaushik et al., 2026), increasing λ eventually pushes the SAE into a degenerate regime where both feature recovery and alignment deteriorate, consistent with Proposition 1. In contrast, our post-hoc approach preserves the feature recovery achieved by modality-specific SAEs before alignment, while achieving better alignment than the auxiliary-loss baselines. These results support decoupling cross-modal alignment from the reconstruction objective.

7.2. Evaluation on Real-World Data

We evaluate our method on real-world embeddings extracted from a pre-trained VLM. We assess whether latent codes support cross-modal tasks and preserve semantically coherent activations. In particular, following prior work (Kaushik et al., 2026), we evaluate whether estimated latent codes $\tilde{\mathbf{z}}$ in (6) provide useful representations for cross-modal re-

trieval, concept steering, and monosemanticity evaluation.

Experimental protocol. We use `CC-3M` (Sharma et al., 2018) as a paired image-text dataset. From this dataset, we extract image and text embeddings using the `CLIP` model (Radford et al., 2021) with the ViT-B/32 architecture, and train SAEs on the resulting paired embeddings. We compare the method proposed in Section 6 with two auxiliary-loss baselines (DHIMOÏLA et al., 2026; Kaushik et al., 2026). To ensure a fair comparison between modality-specific and shared SAE approaches, all methods use the same total number of SAE parameters. We follow prior training protocols (Papadimitriou et al., 2025) and report averages over three runs. Details are provided in Appendix D.4.

Our method improves cross-modal performance while preserving reconstruction. We evaluate reconstruction quality and performance across two cross-modal tasks. Reconstruction quality is measured by the mean squared error between input embeddings and their reconstructions. For cross-modal retrieval on `MS-COCO`, we rank image and text latent codes by cosine similarity and report Recall@k. For zero-shot image classification on `ImageNet1K` (Deng et al., 2009), we report top-1 accuracy.

Table 1 reports the results for all methods. Our method achieves the lowest reconstruction error and the strongest cross-modal retrieval performance, while remaining competitive on zero-shot classification. In cross-modal retrieval, our method outperforms the strongest baseline by 8.9 points in image-to-text Recall@1 (16.0 vs. 7.1) and by 7.1 points in text-to-image Recall@1 (11.4 vs. 4.3). These results show that post-hoc alignment improves cross-modal performance without sacrificing the reconstruction quality of modality-specific SAEs. Additional results across a broader range of VLM sizes are provided in Appendix E.

Our method enables more effective cross-modal concept steering. We evaluate whether the aligned latent coordinates support controllable steering of object concepts. We treat each of the 80 object categories in `COCO` as a target concept. For each target concept, we compute the mean activation of each text latent coordinate on captions that mention the concept and on randomly sampled captions that do not. We then select the text latent coordinate with the largest difference between the two mean activations, which identifies the coordinate most associated with the target concept. Using the aligned image coordinate, we use the corresponding decoder column in the image SAE as the steering vector.

For each target concept, we select 100 source images from the test set that do not contain the target category. We add the steering vector to each source image embedding, producing a steered embedding intended to move it toward the target concept. We then rank all test image embeddings by cosine similarity to each steered embedding and retrieve the nearest images. Successful steering should retrieve images that contain the target concept.

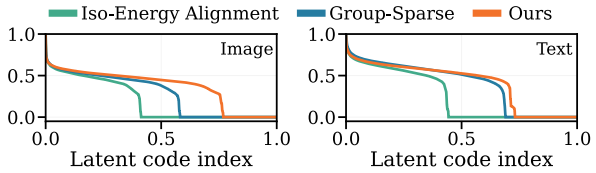


Figure 8. Monosemanticity scores for coordinates sorted in descending order, shown separately for image and text modalities.

Figure 7 reports the retrieval performance and qualitative examples of concept steering. Our method achieves the highest mean average precision and mean reciprocal rank over retrieved test images that contain the target concept. The qualitative examples further show that our method retrieves target-concept images more consistently. These results indicate that post-hoc latent coordinate alignment enables more reliable control than enforcing alignment through an auxiliary loss during training.

Our method preserves more semantically coherent latents. We examine whether the learned codes are activated by semantically coherent inputs. We use the monosemanticity score (Pach et al., 2026) on the validation set of `CC-3M`. This score measures whether inputs that activate the same latent are close to one another in an external embedding space. We use the `MetaCLIP` model (Chuang et al., 2026) as the external encoder. Larger values indicate that the latent responds to a more semantically coherent concept.

Figure 8 reports monosemanticity scores for each latent coordinate, sorted in descending order. Our method maintains higher scores across more latent coordinates in both modalities. This indicates that preserving modality-specific feature directions gives more semantically coherent latents.

8. Conclusion

We studied how features are organized across modalities in joint embedding spaces of vision–language models. We characterized *cross-modal feature heterogeneity*, where the same semantic concept can have different feature directions across image and text modalities. We showed that this heterogeneity can give rise to modality split in SAEs, and that existing auxiliary-loss approaches trade reconstruction quality for latent alignment. Motivated by this limitation, we proposed a simple approach that trains modality-specific SAEs and aligns their latent codes through coactivation. In experiments, our method preserves reconstruction quality and improves cross-modal alignment. In summary, effective alignment in joint embedding spaces should preserve each modality's feature geometry.

References

Alain, G. and Bengio, Y. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations*, 2017.

Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W., and Nanda, N. Refusal in language models is mediated by a single direction. In *Conference on Neural Information Processing Systems*, 2024.

Arora, S., Li, Y., Liang, Y., Ma, T., and Risteski, A. Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics*, 6:483–495, 07 2018.

Brack, M., Friedrich, F., Hintersdorf, D., Struppek, L., Schramowski, P., and Kersting, K. Sega: Instructing text-to-image models using semantic guidance. *Conference on Neural Information Processing Systems*, 2023.

Bussmann, B., Leask, P., and Nanda, N. Batchtopk sparse autoencoders. *arXiv preprint*, 2024.

Chen, Z., Deng, Y., Li, Y., and Gu, Q. Understanding transferable representation learning and zero-shot transfer in CLIP. In *International Conference on Learning Representations*, 2024.

Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., and Jitsev, J. Reproducible scaling laws for contrastive language-image learning. In *Conference on Computer Vision and Pattern Recognition*, 2023.

Chuang, Y.-S., Li, Y., Wang, D., Yeh, C.-F., Lyu, K., Raghavendra, R., Glass, J. R., HUANG, L., Weston, J. E., Zettlemoyer, L., Chen, X., Liu, Z., Xie, S., tau Yih, W., Li, S.-W., and Xu, H. Meta CLIP 2: A worldwide scaling recipe. In *Conference on Neural Information Processing Systems*, 2026.

Chughtai, B., Chan, L., and Nanda, N. A toy model of universality: Reverse engineering how networks learn group operations. In *International Conference on Machine Learning*, 2023.

Conmy, A., Mavor-Parker, A. N., Lynch, A., Heimersheim, S., and Garriga-Alonso, A. Towards automated circuit discovery for mechanistic interpretability. In *Conference on Neural Information Processing Systems*, 2023.

Costa, V., Fel, T., Lubana, E. S., Tolooshams, B., and Ba, D. E. From flat to hierarchical: Extracting sparse representations with matching pursuit. In *Conference on Neural Information Processing Systems*, 2026.

Cui, J., Zhang, Q., Wang, Y., and Wang, Y. On the limits of sparse autoencoders: A theoretical framework and reweighted remedy. In *International Conference on Learning Representations*, 2026.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition*, 2009.

DHIMOÏLA, G., Fel, T., Boutin, V., and Picard, A. M. Cross-modal redundancy and the geometry of vision–language embeddings. In *International Conference on Learning Representations*, 2026.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. A mathematical framework for transformer circuits, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., et al. Toy models of superposition. *arXiv preprint*, 2022.

Eslami, S. and de Melo, G. Mitigate the gap: Improving cross-modal alignment in CLIP. In *International Conference on Learning Representations*, 2025.

Gao, L., la Tour, T. D., Tillman, H., Goh, G., Troll, R., Radford, A., Sutskever, I., Leike, J., and Wu, J. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations*, 2025.

Goodfellow, I. J., Courville, A., and Bengio, Y. Large-scale feature learning with spike-and-slab sparse coding. In *International Coference on International Conference on Machine Learning*, 2012.

Grassucci, E., Cicchetti, G., Frasca, E., Uncini, A., and Comminiello, D. Closing the modality gap aligns group-wise semantics. In *International Conference on Learning Representations*, 2026.

Härle, R., Friedrich, F., Brack, M., Wäldchen, S., Deiseroth, B., Schramowski, P., and Kersting, K. Measuring and guiding monosemanticity. In *Conference on Neural Information Processing Systems*, 2025.

Hewitt, J. and Manning, C. D. A structural probe for finding syntax in word representations. In *North American Chapter of the Association for Computational Linguistics*, 2019.

Huben, R., Cunningham, H., Smith, L. R., Ewart, A., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, 2023.

Jiang, Q., Chen, C., Zhao, H., Chen, L., Ping, Q., Tran, S. D., Xu, Y., Zeng, B., and Chilimbi, T. Understanding and constructing latent modality structures in multi-modal representation learning. In *Conference on Computer Vision and Pattern Recognition*, 2023.

Kaushik, C., Barch, D., and Fanelli, A. Decomposing multimodal embedding spaces with group-sparse autoencoders. In *International Conference on Learning Representations*, 2026.

Kuhn, H. W. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

Li, K., Patel, O., Viégas, F., Pfister, H., and Wattenberg, M. Inference-time intervention: Eliciting truthful answers from a language model. In *Conference on Neural Information Processing Systems*, 2023.

Liang, W., Zhang, Y., Kwon, Y., Yeung, S., and Zou, J. Mind the gap: Understanding the modality gap in multimodal contrastive representation learning. In *Conference on Neural Information Processing Systems*, 2022.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. In *Conference on Neural Information Processing Systems*, 2023.

Liu, S., Ye, H., and Zou, J. Reducing hallucinations in large vision-language models via latent space steering. In *International Conference on Learning Representations*, 2025.

Lu, L., Shin, Y., Su, Y., and Karniadakis, G. E. Dying relu and initialization: Theory and numerical examples. *Communications in Computational Physics*, 28(5):1671–1706, 2020.

Makhzani, A. and Frey, B. K-sparse autoencoders. In *International Conference on Learning Representations*, 2014.

Marks, S., Rager, C., Michaud, E. J., Belinkov, Y., Bau, D., and Mueller, A. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *International Conference on Learning Representations*, 2025.

Mistretta, M., Baldrati, A., Agnolucci, L., Bertini, M., and Bagdanov, A. D. Cross the gap: Exposing the intra-modal misalignment in CLIP via modality inversion. In *International Conference on Learning Representations*, 2025.

Oikarinen, T. and Weng, T.-W. CLIP-dissect: Automatic description of neuron representations in deep vision networks. In *International Conference on Learning Representations*, 2023.

Pach, M., Karthik, S., Bouniot, Q., Belongie, S., and Akata, Z. Sparse autoencoders learn monosemantic features in vision-language models. In *Conference on Neural Information Processing Systems*, 2026.

Papadimitriou, I., Su, H., Fel, T., Kakade, S. M., and Gil, S. Interpreting the linear structure of vision-language model embedding spaces. In *Conference on Language Modeling*, 2025.

Park, K., Choe, Y. J., and Veitch, V. The linear representation hypothesis and the geometry of large language models. In *International Conference on Machine Learning*, 2024.

Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., and Lischinski, D. Styleclip: Text-driven manipulation of stylegan imagery. In *International Conference on Computer Vision*, 2021.

Qian, Q., Xu, Y., and Hu, J. Intra-modal proxy learning for zero-shot visual categorization with clip. In *Conference on Neural Information Processing Systems*, 2023.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Ramasinghe, S., Shevchenko, V., Avraham, G., and Thalaiyasingam, A. Accept the modality gap: An exploration in the hyperbolic space. In *Conference on Computer Vision and Pattern Recognition*, 2024.

Rao, S., Mahajan, S., Böhle, M., and Schiele, B. Discover-then-name: Task-agnostic concept bottlenecks via automated concept discovery. In *European Conference on Computer Vision*, 2024.

Rimsky, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., and Turner, A. Steering llama 2 via contrastive activation addition. In *Annual Meeting of the Association for Computational Linguistics*, 2024.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent

diffusion models. In *Conference on Computer Vision and Pattern Recognition*, 2022.

Schrodi, S., Hoffmann, D. T., Argus, M., Fischer, V., and Brox, T. Two effects, one trigger: On the modality gap, object bias, and information imbalance in contrastive vision-language models. In *International Conference on Learning Representations*, 2025.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Association for Computational Linguistics*, 2018.

Sheikh, A.-S., Shelton, J. A., and Lücke, J. A truncated em approach for spike-and-slab sparse coding. *Journal of Machine Learning Research*, 15(1):2653–2687, 2014.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C., and Henighan, T. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Tschannen, M., Gritsenko, A., Wang, X., Naeem, M. F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint*, 2025.

Wang, K. R., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations*, 2023.

Yamaguchi, S., Feng, D., Kanai, S., Adachi, K., and Chijiwa, D. Post-pre-training for modality alignment in vision-language foundation models. In *Conference on Computer Vision and Pattern Recognition*, 2025.

Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., and Wu, Y. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022.

Zaigrajew, V., Baniecki, H., and Biecek, P. Interpreting CLIP with hierarchical sparse autoencoders. In *International Conference on Machine Learning*, 2025.

Zhang, K., Shen, Y., Li, B., and Liu, Z. Large multi-modal models can interpret features in large multi-modal models. In *International Conference on Computer Vision*, 2025.

Zhang, Y., Sui, E., and Yeung, S. Connect, collapse, corrupt: Learning cross-modal tasks with uni-modal data. In *International Conference on Learning Representations*, 2024.

A. Limitations and Future Work

Our analysis relies on the linear representation hypothesis and simplified generative assumptions, which may not fully capture the complexity of real-world VLM embeddings. While our post-hoc alignment improves empirical performance, it assumes that feature correspondence can be inferred from coactivation statistics, which may become less accurate in noisy or weakly aligned settings.

Future work includes extending the analysis beyond the linear representation hypothesis and studying alignment under more complex forms of heterogeneity. While our current experiments focus on CLIP-like contrastive VLMs, another important direction is to evaluate whether the proposed method remains effective for representations extracted from broader VLM architectures, including LLM-based autoregressive VLMs such as LLaVA. Applying these ideas to larger-scale VLMs and downstream tasks may further clarify the role of feature-level alignment in multimodal systems. Moreover, post-hoc alignment could be used as an initialization for subsequent fine-tuning.

B. Additional Discussion on Cross-Modal Feature Heterogeneity

A possible source of cross-modal feature heterogeneity. We next discuss why cross-modal feature directions may be unevenly aligned in VLMs trained with objectives such as contrastive learning. Such objectives encourage cross-modal similarity for paired samples, typically by increasing $\mathbb{E}[\mathbf{x}^\top \mathbf{y}]$. As the sparsity level $s \to 1$, we have

$$\mathbb{E}\big[\mathbf{x}^\top \mathbf{y}\big] = \mathrm{tr}\big(\boldsymbol{\Phi}^\top \boldsymbol{\Psi}\, \mathbb{E}[\mathbf{z}\mathbf{z}^\top]\big) = s^{n-1}(1-s)\textstyle\sum_{i\in[n]} \boldsymbol{\phi}_i^\top \boldsymbol{\psi}_i \cdot \mathbb{E}\big[z_i^2 \big| z_i \neq 0\big] + o(1-s).$$

This expression suggests that each feature pair contributes to positive similarity in proportion to the second moment of its latent coordinate. Consequently, under finite data or finite model capacity, frequently activated features may receive stronger alignment pressure, whereas rarer features may remain less aligned. This provides one possible source of cross-modal feature heterogeneity.

Relation to modality gap. The following proposition shows that cross-modal feature heterogeneity is unavoidable whenever the embedding pair exhibits a nontrivial modality gap.

Proposition 3. *If there exists a modality gap between an embedding pair, i.e.,* $\mathbb{E}[\cos(\mathbf{x}, \mathbf{y})] < 1$*, then there exists at least one latent concept that exhibits cross-modal feature heterogeneity.*

Proof. For contradiction, assume that there is no cross-modal feature heterogeneity. Then $\boldsymbol{\phi}_i = \boldsymbol{\psi}_i$ for all $i \in [n]$, which implies $\boldsymbol{\Phi} = \boldsymbol{\Psi}$. Hence, the embeddings satisfy $\mathbf{x} = \boldsymbol{\Phi}\mathbf{z} = \boldsymbol{\Psi}\mathbf{z} = \mathbf{y}$ almost surely. This gives $\cos(\mathbf{x}, \mathbf{y}) = 1$ almost surely, contradicting the condition that $\mathbb{E}[\cos(\mathbf{x}, \mathbf{y})] < 1$. Therefore, there must exist at least one $i \in [n]$ such that $\boldsymbol{\phi}_i \neq \boldsymbol{\psi}_i$. □

This result formalizes that a modality gap at the embedding must be reflected by at least one directional mismatch between corresponding image and text features under the linear representation hypothesis. The main question studied in this paper is how such heterogeneity affects multimodal SAEs.

C. Proofs of Theoretical Results

We begin by clarifying notation. Let $\mathbf{0}_d$ and $\mathbf{0}_{m\times n}$ denote the zero vector in $\mathbb{R}^d$ and the $m \times n$ zero matrix, respectively. For a matrix $\mathbf{Z}$, $[\mathbf{Z}]_{[:,i]}$ denotes its i-th column.

Corollary 1. *Suppose the SAE in (3) and (4) has $m \geq 2n$ latent coordinates. Consider the loss $\mathcal{L}_{\mathrm{rec}}(\mathbf{W}; \boldsymbol{\Phi}) + \mathcal{L}_{\mathrm{rec}}(\mathbf{W}; \boldsymbol{\Psi})$ in (5). As the sparsity parameter s in (2) approaches 1, the loss is minimized to leading order by taking $\hat{\mathbf{W}} := [\boldsymbol{\Phi} \;\; \boldsymbol{\Psi} \;\; \mathbf{0}_{d\times(m-2n)}]\,\mathbf{P}$ for any permutation matrix $\mathbf{P} \in \mathbb{R}^{m\times m}$, and $\mathcal{L}_{\mathrm{rec}}(\hat{\mathbf{W}}; \boldsymbol{\Phi}) + \mathcal{L}_{\mathrm{rec}}(\hat{\mathbf{W}}; \boldsymbol{\Psi}) = o(1-s)$.*

Proof. Although Theorem 1 is stated for the regime $m < 2n$, the same argument extends beyond this restriction. When $m \geq 2n$, it gives the present corollary. We give a direct proof to make the construction explicit.

For $\mathbf{z} \in \mathbb{R}^n_+$, define the event $\mathbb{S}_k(\mathbf{z}) := \{\#\{i \in [n] : z_i = 0\} = k\}$ for $k \in \{0\} \cup [n]$, which denotes that exactly k entries of $\mathbf{z}$ are zero. From (2), $\Pr(\mathbb{S}_k(\mathbf{z})) = \binom{n}{k} s^k (1-s)^{n-k}$. In the sparsity regime $s \to 1$, the loss admits the expansion

$$\mathcal{L}_{\mathrm{rec}}(\mathbf{W};\, \boldsymbol{\Phi}) = s^{n-1}(1-s) \sum_{i\in[n]} \mu_i' \, \|\boldsymbol{\phi}_i - \mathbf{W}\sigma(\mathbf{W}^\top \boldsymbol{\phi}_i)\|_2^2 + o(1-s),$$

where $\mu'_i := \mathbb{E}[z_i^2 \mid z_i \neq 0] > 0$ for $i \in [n]$. An analogous expansion holds for $\mathcal{L}_{\text{rec}}(\mathbf{W};\, \mathbf{\Psi})$. Defining $\theta_i := \boldsymbol{\phi}_i$, $\theta_{n+i} := \boldsymbol{\psi}_i$, and $\mu'_{n+i} := \mu'_i$ for $i \in [n]$, we obtain

$$\mathcal{L}_{\text{rec}}(\mathbf{W};\, \mathbf{\Phi}) + \mathcal{L}_{\text{rec}}(\mathbf{W};\, \mathbf{\Psi}) = s^{n-1}(1-s) \sum_{i\in[2n]} \mu'_i \, \|\theta_i - \mathbf{W}\sigma(\mathbf{W}^\top \theta_i)\|_2^2 + o(1-s). \tag{9}$$

Since $s^{n-1}(1-s) > 0$ and $\mu'_i > 0$, the leading term in (9) is minimized to zero precisely when

$$\theta_i = \mathbf{W}\sigma(\mathbf{W}^\top \theta_i) \quad \text{for all } i \in [2n]. \tag{10}$$

For any permutation matrix $\mathbf{P} \in \mathbb{R}^{m\times m}$, substituting $\hat{\mathbf{W}} = [\mathbf{\Phi} \;\; \mathbf{\Psi} \;\; \mathbf{0}_{d\times(m-2n)}]\mathbf{P}$ into (10) gives, for all $i \in [2n]$,

$$[\mathbf{\Phi} \;\; \mathbf{\Psi} \;\; \mathbf{0}_{d\times(m-2n)}]\mathbf{P}\, \sigma\big(\mathbf{P}^\top [\mathbf{\Phi} \;\; \mathbf{\Psi} \;\; \mathbf{0}_{d\times(m-2n)}]^\top \theta_i\big) = [\mathbf{\Phi} \;\; \mathbf{\Psi}]\, \sigma\big([\mathbf{\Phi} \;\; \mathbf{\Psi}]^\top \theta_i\big) = \theta_i.$$

Therefore, $\hat{\mathbf{W}}$ minimizes the leading term in (9), and $\mathcal{L}_{\text{rec}}(\hat{\mathbf{W}};\, \mathbf{\Phi}) + \mathcal{L}_{\text{rec}}(\hat{\mathbf{W}};\, \mathbf{\Psi}) = o(1-s)$. □

Theorem 1. *Suppose the SAE in (3) and (4) has $m < 2n$ latent coordinates. Consider the loss $\mathcal{L}_{\text{rec}}(\mathbf{W};\mathbf{\Phi}) + \mathcal{L}_{\text{rec}}(\mathbf{W};\mathbf{\Psi})$ in (5). Define $\mathbf{M}_i := \mathbb{E}\big[z_i^2 \mid z_i \neq 0\big]\, \boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top \in \mathbb{R}^{d\times d}$ and $\mathbf{M}_{n+i} := \mathbb{E}\big[z_i^2 \mid z_i \neq 0\big]\, \boldsymbol{\psi}_i\boldsymbol{\psi}_i^\top \in \mathbb{R}^{d\times d}$ for $i \in [n]$. Let $(\mathbb{A}_1, \cdots, \mathbb{A}_m)$ be a partition of $[2n]$ that maximizes $\sum_{j\in[m]} \lambda_{\max}\Big(\sum_{i\in\mathbb{A}_j} \mathbf{M}_i\Big)$, where $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue. As the sparsity parameter s in (2) approaches 1, the loss is minimized to leading order by taking $\hat{\mathbf{W}} := [\hat{\mathbf{w}}_1 \;\; \cdots \;\; \hat{\mathbf{w}}_m]$, where each $\hat{\mathbf{w}}_j$ is a unit-norm top eigenvector of $\sum_{i\in\mathbb{A}_j} \mathbf{M}_i$. Moreover, if no two directions among $\{\boldsymbol{\phi}_i\}_{i\in[n]} \cup \{\boldsymbol{\psi}_i\}_{i\in[n]}$ are collinear, $\mathcal{L}_{\text{rec}}(\hat{\mathbf{W}};\mathbf{\Phi}) + \mathcal{L}_{\text{rec}}(\hat{\mathbf{W}};\mathbf{\Psi}) > 0$ for s sufficiently close to 1.*

Proof. For each $i \in [n]$, let $\theta_i := \boldsymbol{\phi}_i$, $\theta_{n+i} := \boldsymbol{\psi}_i$, and $\mu'_i := \mathbb{E}[z_i^2 \mid z_i \neq 0] > 0$, and let $\mu'_{n+i} := \mu'_i$. By the same expansion as in (9), as $s \to 1$, we have

$$\mathcal{L}_{\text{rec}}(\mathbf{W};\, \mathbf{\Phi}) + \mathcal{L}_{\text{rec}}(\mathbf{W};\, \mathbf{\Psi}) = s^{n-1}(1-s) \sum_{i\in[2n]} \mu'_i \, \|\theta_i - \mathbf{W}\sigma(\mathbf{W}^\top \theta_i)\|_2^2 + o(1-s). \tag{11}$$

We first fix $[\mathbf{w}_1 \cdots \mathbf{w}_m]$. For each $i \in [2n]$, let $\hat{\mathbf{w}}_i \in \arg\max_{\mathbf{w}\in\{\mathbf{w}_j\}_{j\in[m]}} \mathbf{w}^\top \theta_i$. Then

$$\sum_{i\in[2n]} \mu'_i \|\theta_i - \mathbf{W}\sigma(\mathbf{W}^\top \theta_i)\|_2^2 = \sum_{i\in[2n]} \mu'_i \|\theta_i - \hat{\mathbf{w}}_i\hat{\mathbf{w}}_i^\top \theta_i\|_2^2 = \sum_{i\in[2n]} \mu'_i - \sum_{i\in[2n]} \mu'_i (\theta_i^\top \hat{\mathbf{w}}_i)^2$$

Since $\sum_{i\in[2n]} \mu'_i$ is constant with respect to $\mathbf{W}$, minimizing the leading term in (11) is equivalent to maximizing

$$\mathcal{J}(\mathbf{W}) := \sum_{i\in[2n]} \mu'_i (\theta_i^\top \hat{\mathbf{w}}_i)^2 = \sum_{i\in[2n]} \hat{\mathbf{w}}_i^\top \mathbf{M}_i \, \hat{\mathbf{w}}_i,$$

where, for each $i \in [n]$, we define $\mathbf{M}_i := \mu'_i \boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top \in \mathbb{R}^{d\times d}$ and $\mathbf{M}_{n+i} := \mu'_i \boldsymbol{\psi}_i\boldsymbol{\psi}_i^\top \in \mathbb{R}^{d\times d}$ for $i \in [n]$. For each $j \in [m]$, define

$$\mathbb{A}_j := \{i \in [2n] | \hat{\mathbf{w}}_i = \mathbf{w}_j\},$$

where $\mathbb{A}_j$ contains the feature indices whose corresponding feature vectors are extracted by the column $\mathbf{w}_j$. Then $(\mathbb{A}_1, \cdots, \mathbb{A}_m)$ forms a partition of $[2n]$, and

$$\mathcal{J}(\mathbf{w}) = \sum_{j\in[m]} \mathbf{w}_j^\top \left(\sum_{i\in\mathbb{A}_j} \mathbf{M}_i\right) \mathbf{w}_j.$$

For a fixed partition $(\mathbb{A}_1, \ldots, \mathbb{A}_m)$, maximizing $\mathcal{J}(\mathbf{W})$ decouples over the columns of $\mathbf{W}$. Hence, for each $j \in [m]$,

$$\max_{\|\mathbf{w}_j\|_2^2=1} \mathbf{w}_j^\top \left(\sum_{i\in\mathbb{A}_j} \mathbf{M}_i\right) \mathbf{w}_j = \lambda_{\max}\left(\sum_{i\in\mathbb{A}_j} \mathbf{M}_i\right),$$

where $\lambda_{\max}$ denotes the largest eigenvalue. The maximum is attained at a unit eigenvector corresponding to this largest eigenvalue. Therefore, minimizing the leading term of the loss in (11) is equivalent to maximizing

$$\sum_{j\in[m]} \lambda_{\max}\Big(\sum_{i\in\mathbb{A}_j} \mathbf{M}_i\Big)$$

over all partitions $(\mathbb{A}_1,\dots,\mathbb{A}_m)$ of $[2n]$. Thus, choosing each $\hat{\mathbf{w}}_j$ as a unit eigenvector corresponding to the largest eigenvalue of $\sum_{i\in\mathbb{A}_j}\mathbf{M}_i$ yields a global minimizer $\hat{\mathbf{W}}$ of (11).

Substituting back into (11), together with $\sum_{i\in[2n]}\mu_i' = 2\sum_{i\in[n]}\mu_i'$, gives

$$\mathcal{L}_{\text{rec}}(\hat{\mathbf{W}};\,\mathbf{\Phi}) + \mathcal{L}_{\text{rec}}(\hat{\mathbf{W}};\,\mathbf{\Psi}) = s^{n-1}(1-s)\Big(2\sum_{i\in[n]}\mu_i' - \sum_{j\in[m]}\lambda_{\max}\Big(\sum_{i\in\mathbb{A}_j}\mathbf{M}_i\Big)\Big) + o(1-s).$$

It remains to note that the leading coefficient is nonnegative. Since each $\mathbf{M}_i$ is positive semidefinite,

$$\lambda_{\max}\Big(\sum_{i\in\mathbb{A}_j}\mathbf{M}_i\Big) \le \operatorname{tr}\Big(\sum_{i\in\mathbb{A}_j}\mathbf{M}_i\Big)$$

for each $j\in[m]$. Therefore, using the fact that $(\mathbb{A}_1,\dots,\mathbb{A}_m)$ is a partition of $[2n]$ and that $\|\boldsymbol{\phi}_i\|_2 = \|\boldsymbol{\psi}_i\|_2 = 1$, we obtain

$$\sum_{j\in[m]}\lambda_{\max}\Big(\sum_{i\in\mathbb{A}_j}\mathbf{M}_i\Big) \le \sum_{j\in[m]}\operatorname{tr}\Big(\sum_{i\in\mathbb{A}_j}\mathbf{M}_i\Big) = \sum_{i\in[2n]}\operatorname{tr}(\mathbf{M}_i) = 2\sum_{i\in[n]}\mu_i'.$$

Moreover, assume that no two feature directions are collinear, *i.e.,* $\theta_i \ne \pm\theta_j$ for all distinct $i,j\in[2n]$, then the above inequality is strict. Since $m<2n$, at least one set $\mathbb{A}_j$ contains two distinct indices. For this set, the matrix $\sum_{i\in\mathbb{A}_j}\mathbf{M}_i$ has rank at least two, and hence its largest eigenvalue is strictly smaller than its trace. Therefore,

$$\sum_{j\in[m]}\lambda_{\max}\Big(\sum_{i\in\mathbb{A}_j}\mathbf{M}_i\Big) < 2\sum_{i\in[n]}\mu_i'.$$

Thus the leading coefficient of the loss is strictly positive. Consequently,

$$\mathcal{L}_{\text{rec}}(\hat{\mathbf{W}};\,\mathbf{\Phi}) + \mathcal{L}_{\text{rec}}(\hat{\mathbf{W}};\,\mathbf{\Psi}) > 0$$

for all sufficiently large $s<1$. □

Proposition 1. *Consider the case $n=1$, where the feature directions $\boldsymbol{\phi},\boldsymbol{\psi}\in\mathbb{R}^d$ satisfy $\rho := \boldsymbol{\phi}^\top\boldsymbol{\psi}\in(0,1)$. Suppose the SAE in (3) and (4) has $m\ge 2$ latent coordinates. Consider the loss (Kaushik et al., 2026)*

$$\begin{aligned}&\mathcal{L}_{\text{rec}}(\mathbf{W};\boldsymbol{\phi}) + \mathcal{L}_{\text{rec}}(\mathbf{W};\boldsymbol{\psi})\\ &\quad + \lambda\,\mathbb{E}_z\Big[\textstyle\sum_{j\in[m]}\sqrt{\big[\sigma\big(\mathbf{W}^\top\boldsymbol{\phi}\,z\big)\big]_j^2 + \big[\sigma\big(\mathbf{W}^\top\boldsymbol{\psi}\,z\big)\big]_j^2}\Big].\end{aligned}$$

Define $\lambda^\star(\rho) := \frac{(1-\rho)\mathbb{E}[z^2]}{(2-\sqrt{1+\rho})\cdot\mathbb{E}[z]}$ and $\lambda^{\star\star}(\rho) := \frac{\sqrt{1+\rho}\cdot\mathbb{E}[z^2]}{\mathbb{E}[z]}$. Up to a permutation matrix $\mathbf{P}$, a global minimizer $\hat{\mathbf{W}}$ is:

- *If $\lambda<\lambda^\star(\rho)$, then $\hat{\mathbf{W}} = [\boldsymbol{\phi}\ \boldsymbol{\psi}\ \mathbf{0}_{d\times(m-2)}]\,\mathbf{P}$, where $\mathcal{L}_{\text{rec}}(\hat{\mathbf{W}};\boldsymbol{\phi}) = 0$.*
- *If $\lambda\in\big(\lambda^\star(\rho),\,\lambda^{\star\star}(\rho)\big)$, then $\hat{\mathbf{W}} = \big[\frac{\boldsymbol{\phi}+\boldsymbol{\psi}}{\|\boldsymbol{\phi}+\boldsymbol{\psi}\|}\ \mathbf{0}_{d\times(m-1)}\big]\,\mathbf{P}$, where $\mathcal{L}_{\text{rec}}(\hat{\mathbf{W}};\boldsymbol{\phi}) = \frac{1-\rho}{2}\,\mathbb{E}[z^2]$.*
- *If $\lambda>\lambda^{\star\star}(\rho)$, then $\hat{\mathbf{W}}$ satisfies $\sigma(\hat{\mathbf{W}}^\top\boldsymbol{\phi}) = \sigma(\hat{\mathbf{W}}^\top\boldsymbol{\psi}) = \mathbf{0}$, where $\mathcal{L}_{\text{rec}}(\hat{\mathbf{W}};\boldsymbol{\phi}) = \mathbb{E}[z^2]$.*

Proof. Since $z\ge 0$, we have $\sigma(\mathbf{W}^\top\boldsymbol{\phi}z) = z\,\sigma(\mathbf{W}^\top\boldsymbol{\phi})$ and $\sigma(\mathbf{W}^\top\boldsymbol{\psi}z) = z\,\sigma(\mathbf{W}^\top\boldsymbol{\psi})$. Hence

$$\mathcal{L}_{\text{rec}}(\mathbf{W};\boldsymbol{\phi}) + \mathcal{L}_{\text{rec}}(\mathbf{W};\boldsymbol{\psi}) = \mathbb{E}[z^2]\Big(\|\boldsymbol{\phi}-\mathbf{W}\sigma(\mathbf{W}^\top\boldsymbol{\phi})\|^2 + \|\boldsymbol{\psi}-\mathbf{W}\sigma(\mathbf{W}^\top\boldsymbol{\psi})\|^2\Big),$$

$$\mathbb{E}_z\Big[\sum_{j\in[m]}\sqrt{[\sigma(\mathbf{W}^\top\boldsymbol{\phi}z)]_j^2 + [\sigma(\mathbf{W}^\top\boldsymbol{\psi}z)]_j^2}\Big] = \mathbb{E}[z]\sum_{j\in[m]}\sqrt{[\sigma(\mathbf{W}^\top\boldsymbol{\phi})]_j^2 + [\sigma(\mathbf{W}^\top\boldsymbol{\psi})]_j^2}.$$

Setting $\mathbf{u} := \sigma(\mathbf{W}^\top \boldsymbol{\phi})$, $\mathbf{v} := \sigma(\mathbf{W}^\top \boldsymbol{\psi})$, and $\tilde{\lambda} := \lambda\,\mathbb{E}[z]/\mathbb{E}[z^2]$, dividing through by $\mathbb{E}[z^2]$ reduces the problem to minimizing

$$\mathcal{J}(\mathbf{W}) := \|\boldsymbol{\phi} - \mathbf{W}\mathbf{u}\|^2 + \|\boldsymbol{\psi} - \mathbf{W}\mathbf{v}\|^2 + \tilde{\lambda} \sum_{j\in[m]} \sqrt{[\mathbf{u}]_j^2 + [\mathbf{v}]_j^2}.$$

If $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ activate different columns $i \neq j$, setting $[\mathbf{W}]_{[:,i]} = \boldsymbol{\phi}$ and $[\mathbf{W}]_{[:,j]} = \boldsymbol{\psi}$ gives zero reconstruction error and group penalty 2. Hence,

$$L_{\text{sep}} := \mathcal{J}(\mathbf{W}) = 2\tilde{\lambda}.$$

If instead both activate the same unit column $\mathbf{w}$, writing $a := \mathbf{w}^\top \boldsymbol{\phi} \geq 0$ and $b := \mathbf{w}^\top \boldsymbol{\psi} \geq 0$,

$$\mathcal{J}([\mathbf{w}]) = 2 - (a^2 + b^2) + \tilde{\lambda}\sqrt{a^2 + b^2}.$$

With $\mathbf{M} := \boldsymbol{\phi}\boldsymbol{\phi}^\top + \boldsymbol{\psi}\boldsymbol{\psi}^\top$, we have $a^2 + b^2 = \mathbf{w}^\top \mathbf{M}\mathbf{w}$. Then the top eigenvector of $\mathbf{M}$ is

$$\hat{\mathbf{w}} := \frac{\boldsymbol{\phi} + \boldsymbol{\psi}}{\|\boldsymbol{\phi} + \boldsymbol{\psi}\|}, \qquad a = b = \sqrt{\tfrac{1+\rho}{2}},$$

with eigenvalue $1 + \rho$. By the Rayleigh quotient, the feasible range of $a^2 + b^2$ over unit $\mathbf{w}$ satisfies $a^2 + b^2 \leq 1 + \rho$, with equality at $\hat{\mathbf{w}}$. Since $\mathcal{J}([\mathbf{w}])$ depends on $\mathbf{w}$ through $r := \sqrt{a^2 + b^2}$, define

$$f(r) := 2 - r^2 + \tilde{\lambda} r, \qquad r \in [0, \sqrt{1+\rho}].$$

Since f is concave in r, its minimum over $[0, \sqrt{1+\rho}]$ is attained at an endpoint. Evaluating at $r = 0$ and $r = \sqrt{1+\rho}$ gives

$$f(0) = 2, \qquad f(\sqrt{1+\rho}) = (1-\rho) + \tilde{\lambda}\sqrt{1+\rho}.$$

Thus

$$L_{\text{sh}} := \min_{\mathbf{w}} \mathcal{J}([\mathbf{w}]) = \min\left\{2,\ (1-\rho) + \tilde{\lambda}\sqrt{1+\rho}\right\}.$$

Finally, consider the inactive solution where $\sigma(\mathbf{W}^\top \boldsymbol{\phi}) = \mathbf{0}$ and $\sigma(\mathbf{W}^\top \boldsymbol{\psi}) = \mathbf{0}$. In this case,

$$L_{\text{dead}} := \mathcal{J}(\mathbf{W}) = \|\boldsymbol{\phi}\|^2 + \|\boldsymbol{\psi}\|^2 = 2.$$

We now compare the three values:

$$L_{\text{sep}} = 2\tilde{\lambda}, \quad L_{\text{sh}} = \min\left\{2,\ (1-\rho) + \tilde{\lambda}\sqrt{1+\rho}\right\}, \quad L_{\text{dead}} = 2.$$

If $\tilde{\lambda} < \frac{1-\rho}{2-\sqrt{1+\rho}}$, then $2\tilde{\lambda} < (1-\rho) + \tilde{\lambda}\sqrt{1+\rho}$ and $2\tilde{\lambda} < 2$; hence $L_{\text{sep}} < L_{\text{sh}}$ and $L_{\text{sep}} < L_{\text{dead}}$. If $\frac{1-\rho}{2-\sqrt{1+\rho}} < \tilde{\lambda} < \sqrt{1+\rho}$, then $(1-\rho) + \tilde{\lambda}\sqrt{1+\rho} < 2\tilde{\lambda}$ and $(1-\rho) + \tilde{\lambda}\sqrt{1+\rho} < 2$; hence $L_{\text{sh}} < L_{\text{sep}}$ and $L_{\text{sh}} < L_{\text{dead}}$. If $\tilde{\lambda} > \sqrt{1+\rho}$, then $2 < 2\tilde{\lambda}$ and $2 < (1-\rho) + \tilde{\lambda}\sqrt{1+\rho}$; hence $L_{\text{dead}} < L_{\text{sep}}$ and $L_{\text{dead}} < L_{\text{sh}}$.

Substituting $\tilde{\lambda} = \lambda\,\mathbb{E}[z]/\mathbb{E}[z^2]$ yields the result. When the shared solution is optimal, the reconstruction errors are

$$\mathcal{L}_{\text{rec}}(\hat{\mathbf{W}}; \boldsymbol{\phi}) = \mathbb{E}[z^2](1 - a^2) = \tfrac{1-\rho}{2}\,\mathbb{E}[z^2], \qquad \mathcal{L}_{\text{rec}}(\hat{\mathbf{W}}; \boldsymbol{\psi}) = \mathbb{E}[z^2](1 - b^2) = \tfrac{1-\rho}{2}\,\mathbb{E}[z^2],$$

and when the inactive solution is optimal,

$$\mathcal{L}_{\text{rec}}(\hat{\mathbf{W}}; \boldsymbol{\phi}) = \mathcal{L}_{\text{rec}}(\hat{\mathbf{W}}; \boldsymbol{\psi}) = \mathbb{E}[z^2].$$

□

Proposition 2. *Consider the case $n = 1$, where the feature directions $\boldsymbol{\phi}, \boldsymbol{\psi} \in \mathbb{R}^d$ satisfy $\rho := \boldsymbol{\phi}^\top \boldsymbol{\psi} \in (0, 1)$. Suppose the SAE in (3) and (4) has $m \geq 2$ latent coordinates. Consider the loss (DHIMOÏLA et al., 2026)*

$$\mathcal{L}_{\text{rec}}(\mathbf{W};\boldsymbol{\phi})+\mathcal{L}_{\text{rec}}(\mathbf{W};\boldsymbol{\psi})-\lambda\mathbb{E}_z\Big[\sigma\big(\mathbf{W}^\top\boldsymbol{\phi}z\big)^\top\sigma\big(\mathbf{W}^\top\boldsymbol{\psi}z\big)\Big].$$

Define $\lambda^\star(\rho) := \frac{2(1-\rho)}{1+\rho}$. Up to a permutation matrix $\mathbf{P}$, a global minimizer $\hat{\mathbf{W}}$ is:

- *If $\lambda < \lambda^\star(\rho)$, then $\hat{\mathbf{W}} = [\boldsymbol{\phi}\ \boldsymbol{\psi}\ \mathbf{0}_{d\times(m-2)}]\,\mathbf{P}$, where $\mathcal{L}_{\text{rec}}(\hat{\mathbf{W}};\boldsymbol{\phi}) = 0$.*
- *If $\lambda > \lambda^\star(\rho)$, then $\hat{\mathbf{W}} = \big[\frac{\boldsymbol{\phi}+\boldsymbol{\psi}}{\|\boldsymbol{\phi}+\boldsymbol{\psi}\|}\ \mathbf{0}_{d\times(m-1)}\big]\mathbf{P}$, where $\mathcal{L}_{\text{rec}}(\hat{\mathbf{W}};\boldsymbol{\phi}) = \frac{1-\rho}{2}\,\mathbb{E}[z^2]$.*

Proof. Since $z \geq 0$, we have $\sigma(\mathbf{W}^\top\boldsymbol{\phi}z) = z\,\sigma(\mathbf{W}^\top\boldsymbol{\phi})$ and $\sigma(\mathbf{W}^\top\boldsymbol{\psi}z) = z\,\sigma(\mathbf{W}^\top\boldsymbol{\psi})$. Hence

$$\begin{aligned}\mathcal{L}_{\text{rec}}(\mathbf{W};\boldsymbol{\phi}) + \mathcal{L}_{\text{rec}}(\mathbf{W};\boldsymbol{\psi}) &= \mathbb{E}[z^2]\Big(\|\boldsymbol{\phi} - \mathbf{W}\sigma(\mathbf{W}^\top\boldsymbol{\phi})\|^2 + \|\boldsymbol{\psi} - \mathbf{W}\sigma(\mathbf{W}^\top\boldsymbol{\psi})\|^2\Big),\\ \mathbb{E}_z\Big[\sigma(\mathbf{W}^\top\boldsymbol{\phi}z)^\top\sigma(\mathbf{W}^\top\boldsymbol{\psi}z)\Big] &= \mathbb{E}[z^2]\,\sigma(\mathbf{W}^\top\boldsymbol{\phi})^\top\sigma(\mathbf{W}^\top\boldsymbol{\psi}).\end{aligned}$$

Setting $\mathbf{u} := \sigma(\mathbf{W}^\top\boldsymbol{\phi})$ and $\mathbf{v} := \sigma(\mathbf{W}^\top\boldsymbol{\psi})$, the loss factors as $\mathbb{E}[z^2]\,\mathcal{J}(\mathbf{W})$, where

$$\mathcal{J}(\mathbf{W}) := \|\boldsymbol{\phi} - \mathbf{W}\mathbf{u}\|^2 + \|\boldsymbol{\psi} - \mathbf{W}\mathbf{v}\|^2 - \lambda\,\mathbf{u}^\top\mathbf{v}.$$

If $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ activate different columns $j_a \neq j_b$, setting $[\mathbf{W}]_{[:,j_a]} = \boldsymbol{\phi}$ and $[\mathbf{W}]_{[:,j_b]} = \boldsymbol{\psi}$ gives zero reconstruction error and $\mathbf{u}^\top\mathbf{v} = 0$; hence

$$L_{\text{sep}} := \mathcal{J}(\mathbf{W}) = 0.$$

If instead both activate the same unit column $\mathbf{w}$, writing $a := \mathbf{w}^\top\boldsymbol{\phi} \geq 0$ and $b := \mathbf{w}^\top\boldsymbol{\psi} \geq 0$,

$$\mathcal{J}([\mathbf{w}]) = 2 - a^2 - b^2 - \lambda\,ab = 2 - \mathbf{w}^\top\mathbf{M}_\lambda\mathbf{w},$$

where $\mathbf{M}_\lambda := \boldsymbol{\phi}\boldsymbol{\phi}^\top + \boldsymbol{\psi}\boldsymbol{\psi}^\top + \frac{\lambda}{2}(\boldsymbol{\phi}\boldsymbol{\psi}^\top + \boldsymbol{\psi}\boldsymbol{\phi}^\top)$. Under $\rho \in (0, 1)$ and $\lambda \geq 0$, the top eigenvector is

$$\hat{\mathbf{w}} := \frac{\boldsymbol{\phi} + \boldsymbol{\psi}}{\|\boldsymbol{\phi} + \boldsymbol{\psi}\|}, \qquad a = b = \sqrt{\tfrac{1+\rho}{2}}.$$

By the Rayleigh quotient, $\mathbf{w}^\top\mathbf{M}_\lambda\mathbf{w} \leq (1 + \rho)(1 + \frac{\lambda}{2})$ for unit $\mathbf{w}$, with equality at $\hat{\mathbf{w}}$. Therefore

$$L_{\text{sh}} := \min_{\mathbf{w}} \mathcal{J}([\mathbf{w}]) = 2 - (1 + \rho)(1 + \tfrac{\lambda}{2}) = (1 - \rho) - \lambda\,\tfrac{1+\rho}{2}.$$

Finally, consider the inactive solution where $\sigma(\mathbf{W}^\top\boldsymbol{\phi}) = \mathbf{0}$ and $\sigma(\mathbf{W}^\top\boldsymbol{\psi}) = \mathbf{0}$. In this case,

$$L_{\text{dead}} := \mathcal{J}(\mathbf{W}) = \|\boldsymbol{\phi}\|^2 + \|\boldsymbol{\psi}\|^2 = 2.$$

We now compare the three values:

$$L_{\text{sep}} = 0, \quad L_{\text{sh}} = (1 - \rho) - \lambda\,\tfrac{1+\rho}{2}, \quad L_{\text{dead}} = 2.$$

Since $\lambda \geq 0$, we have $L_{\text{dead}} = 2 > 0 = L_{\text{sep}}$, so the dead solution is never optimal. Comparing the remaining two, $L_{\text{sh}} < L_{\text{sep}}$ if and only if $\lambda > \frac{2(1-\rho)}{1+\rho} = \lambda^\star(\rho)$.

When $\lambda > \lambda^\star(\rho)$, the reconstruction errors at the shared minimizer are

$$\mathcal{L}_{\text{rec}}(\hat{\mathbf{W}};\boldsymbol{\phi}) = \mathbb{E}[z^2](1 - a^2) = \tfrac{1-\rho}{2}\,\mathbb{E}[z^2], \qquad \mathcal{L}_{\text{rec}}(\hat{\mathbf{W}};\boldsymbol{\psi}) = \mathbb{E}[z^2](1 - b^2) = \tfrac{1-\rho}{2}\,\mathbb{E}[z^2].$$

□

D. Experimental Details

This section provides details for the empirical analyses in Sections 4 and 7. The code is available at https://github.com/JiH00nKw0n/cross_modal_feature_heterogeneity, and all experiments are conducted on a single NVIDIA A100 GPU.

D.1. Experimental Details for Measuring Cross-Modal Feature Heterogeneity

This section provides experimental details for Figure 2 in Section 4, which reports the distribution of cosine distances between estimated image and text feature directions grouped by coactivation correlation. Recall that we use decoder columns $\hat{\boldsymbol{\phi}}_i$ and $\hat{\boldsymbol{\psi}}_j$ of modality-specific SAEs as estimates of image and text feature directions, and that the coactivation correlation $c_{i,j} := [\mathbf{C}]_{i,j}$ between the i-th image latent and the j-th text latent on paired embeddings (6) serves as a proxy for semantic correspondence. For each pair (i, j), we measure the cosine distance between $\hat{\boldsymbol{\phi}}_i$ and $\hat{\boldsymbol{\psi}}_j$.

We group all m^2 index pairs (i, j)[3] by their correlation value $c_{i,j}$ and examine the distribution of cosine distances between the corresponding feature vectors. A larger $c_{i,j}$ indicates that the corresponding image and text features are more likely to represent the same shared concept. Under perfect cross-modal feature alignment, their cosine distances would concentrate near zero. We train modality-specific SAEs with latent dimension $m = 8192$, using the Top-K (Makhzani & Frey, 2014) activation function with $K = 8$. We follow the training protocol of Papadimitriou et al. (2025). Specifically, we use the AdamW optimizer with learning rate 5×10^{-4} and weight decay 10^{-5}, batch size 1024, and a cosine schedule with 5% linear warmup over 30 training epochs.

D.2. Implementation Details for Synthetic Data Experiments

This section provides the data-generation specification, SAE architectures, and optimization hyperparameters used in Section 7.1. We repeat each experiment over three runs, reporting the average.

Dataset. We generate each pair of synthetic embeddings, following prior work (Goodfellow et al., 2012; Sheikh et al., 2014). We first sample a sparse latent code $\mathbf{z}$ from a Bernoulli–Exponential process, and then form $(\mathbf{x}, \mathbf{y})$ from the ground-truth feature matrices $(\boldsymbol{\Phi}, \boldsymbol{\Psi})$ via (1). The embedding dimension is $d = 256$, and the feature matrices $\boldsymbol{\Phi}, \boldsymbol{\Psi} \in \mathbb{R}^{d\times n}$ contain $n = n_S + n_I + n_T = 2048$ columns in total, where $n_S = 1024$ is the number of shared concepts present in both modalities, $n_I = 512$ is the number of image-only concepts (with $\boldsymbol{\psi}_i = \mathbf{0}$), and $n_T = 512$ is the number of text-only concepts (with $\boldsymbol{\phi}_i = \mathbf{0}$). For each shared concept $i \in [n_S]$, we sample $\boldsymbol{\phi}_i$ and $\boldsymbol{\psi}_i$ as unit-norm vectors satisfying $\cos(\boldsymbol{\phi}_i, \boldsymbol{\psi}_i) = \alpha$ for a prescribed cross-modal feature alignment α.

All feature directions are unit norm. Apart from the prescribed alignment $\cos(\boldsymbol{\phi}_i, \boldsymbol{\psi}_i) = \alpha$ between matched shared concepts, every pair of directions belonging to distinct concepts $i \neq j$ has bounded pairwise interference, $|\cos(\boldsymbol{\phi}_i, \boldsymbol{\phi}_j)|, |\cos(\boldsymbol{\psi}_i, \boldsymbol{\psi}_j)|, |\cos(\boldsymbol{\phi}_i, \boldsymbol{\psi}_j)| \leq 0.30$. Each coordinate of the latent code $\mathbf{z} \in \mathbb{R}^n_+$ is independently set to zero with probability $s = 0.99$, and otherwise drawn i.i.d. from $\mathrm{Exp}(\beta)$ with rate $\beta = 1$. The paired embeddings are then formed as $\mathbf{x} = \boldsymbol{\Phi}\mathbf{z} + \boldsymbol{\epsilon}_I$ and $\mathbf{y} = \boldsymbol{\Psi}\mathbf{z} + \boldsymbol{\epsilon}_T$ via (1), with independent observation noise $\boldsymbol{\epsilon}_I, \boldsymbol{\epsilon}_T \sim \mathcal{N}(\mathbf{0}, \sigma^2_{\mathrm{obs}}\mathbf{I})$ of standard deviation $\sigma_{\mathrm{obs}} = 0.05$. We generate 50,000 paired embeddings for training and 10,000 for evaluation.

Models. The shared SAE, which uses a single SAE across both modalities (shown in Figures 5 and 6), has latent dimension $m = 8192$ and uses the Top-K activation function (Makhzani & Frey, 2014) with $K = 16$. Each modality-specific SAE, which uses a separate SAE for each modality, uses latent dimension $m = 4096$, so that the total number of learnable parameters matches that of the shared SAE. The two baselines (Iso-Energy Alignment (DHIMOÏLA et al., 2026) and Group-Sparse (Kaushik et al., 2026)), shown in Figure 6, augment the shared SAE by training with an auxiliary loss.

Training. All SAEs are trained with AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$) at a constant learning rate of 5×10^{-4} with no warmup. We use a batch size of 256 and train for 10 epochs.

D.3. Evaluation Metrics for Synthetic Data Experiments

We provide formal definitions of the five metrics introduced in Section 7.1. We assess each method using these metrics: (i) and (ii) measure how well the SAE reconstructs embeddings and recovers features, (iii) and (iv) measure how well it aligns

[3] We exclude dead latent coordinates, which are zero across all inputs, so fewer than m^2 pairs remain in practice.

latent codes across modalities, and (v) measures whether cross-modal features collapse into a single feature.

(i) *Reconstruction Error.* We measure the mean squared error (MSE) between input embeddings $\mathbf{x}$ (or $\mathbf{y}$) and their reconstructions $\tilde{\mathbf{x}}$ (or $\tilde{\mathbf{y}}$) on an evaluation set, averaged over both modalities.

(ii) *Feature Recovery Error.* Similar to Reconstruction Error, but using the feature direction vectors $\boldsymbol{\phi}_i$ (or $\boldsymbol{\psi}_i$) as inputs, we measure their reconstruction MSE averaged over both modalities.

(iii) *Latent Alignment Error.* We measure the cosine distance between the image and text latent codes $\tilde{\mathbf{z}}_{\mathrm{I}}(\mathbf{x})$ and $\tilde{\mathbf{z}}_{\mathrm{T}}(\mathbf{y})$ obtained from paired embeddings $(\mathbf{x}, \mathbf{y})$ on an evaluation set.

(iv) *Feature Alignment Error.* Similar to Alignment Error, but using features of the same concept $(\boldsymbol{\phi}_i, \boldsymbol{\psi}_i)$ as inputs, we measure the cosine distance between the latent codes $\tilde{\mathbf{z}}_{\mathrm{I}}(\boldsymbol{\phi}_i)$ and $\tilde{\mathbf{z}}_{\mathrm{T}}(\boldsymbol{\psi}_i)$.

(v) *(Cross-Modal) Feature Collapse Rate.* We measure the fraction of feature pairs $(\boldsymbol{\phi}_i, \boldsymbol{\psi}_i)$ that represent the same concept across modalities and are assigned to the same latent coordinate, indicating whether the SAE merges two modality-specific features into a single learned feature.

Note that metrics (ii), (iv), and (v) have to know the ground-truth features $(\boldsymbol{\Phi}, \boldsymbol{\Psi})$ and thus apply only in the synthetic setting. The mathematical definitions are given below.

(i) *Reconstruction Error:*

$$\frac{1}{2}\,\mathbb{E}\Big[\,\|\mathbf{x}-\tilde{\mathbf{x}}\|_2^2 + \|\mathbf{y}-\tilde{\mathbf{y}}\|_2^2\Big] = \frac{1}{2}\,\mathbb{E}\Big[\,\big\|\mathbf{x}-\mathbf{W}\sigma\big(\mathbf{W}^\top\mathbf{x}\big)\big\|_2^2 + \big\|\mathbf{y}-\mathbf{W}\sigma\big(\mathbf{W}^\top\mathbf{y}\big)\big\|_2^2\Big].$$

(ii) *Feature Recovery Error:*

$$\frac{1}{2n_{\mathrm{S}}}\sum_{i\in[n_{\mathrm{S}}]}\Big(\,\big\|\boldsymbol{\psi}_i-\mathbf{W}\sigma\big(\mathbf{W}^\top\boldsymbol{\psi}_i\big)\big\|_2^2 + \big\|\boldsymbol{\phi}_i-\mathbf{W}\sigma\big(\mathbf{W}^\top\boldsymbol{\phi}_i\big)\big\|_2^2\Big).$$

(iii) *Latent Alignment Error:*

$$1-\mathbb{E}\left[\cos\left(\tilde{\mathbf{z}}(\mathbf{x}),\tilde{\mathbf{z}}(\mathbf{y})\right)\right] = 1-\mathbb{E}\left[\cos\left(\sigma(\mathbf{W}^\top\mathbf{x}),\sigma(\mathbf{W}^\top\mathbf{y})\right)\right].$$

(iv) *Feature Alignment Error:*

$$1-\frac{1}{n_{\mathrm{S}}}\sum_{i\in[n_{\mathrm{S}}]}\cos\big(\sigma(\mathbf{W}^\top\boldsymbol{\phi}_i),\ \sigma(\mathbf{W}^\top\boldsymbol{\psi}_i)\big).$$

(v) *(Cross-Modal) Feature Collapse Rate:*

$$\frac{1}{n_{\mathrm{S}}}\sum_{i\in[n_{\mathrm{S}}]}\mathbf{1}\left[\,\arg\max_{j\in[m]}\cos\big(\mathbf{w}_j,\ \boldsymbol{\phi}_i\big)\ =\ \arg\max_{j\in[m]}\cos\big(\mathbf{w}_j,\ \boldsymbol{\psi}_i\big)\,\right]\ \in\ [0,1].$$

D.4. Implementation Details for Real-World Experiments

Implementation Details. We use the `CC-3M` (Sharma et al., 2018) dataset having approximately 2.82M image-caption pairs after dropping invalid URLs. We optimize all SAEs with AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$) at learning rate 5×10^{-4}, weight decay 10^{-5}, batch size 1024, gradient clipping at norm 1.0, and a cosine schedule with 5% linear warmup followed by cosine decay to zero. We use latent dimension $m = 8192$ and Top-K activation (Makhzani & Frey, 2014) with $K = 32$, and we adopt the training protocol of Papadimitriou et al. (2025).

Licenses of datasets. The `CC-3M` dataset is released by Google for free use under the terms of its repository license. The `MS-COCO` dataset is released under `CC BY 4.0`.

E. Additional Experimental Results

E.1. Evidence for Cross-Modal Feature Heterogeneity on Embeddings of Larger VLMs

To check whether the observed cross-modal feature heterogeneity persists at larger scales, we repeat the same analysis (shown in Figure 2) on the corresponding large-size models (ViT-L/14) of `CLIP`, `MetaCLIP`, `OpenCLIP`, and `SigLIP2`, and report the results in Figure 9 (Figure 2 in the main body shows results for the base-size models from the four VLM families). The overall trend is consistent with Figure 2: feature pairs with larger coactivation correlation tend to have smaller cosine distances, but they do not concentrate near zero. This indicates that cross-modal feature heterogeneity persists across embedding models of different sizes.

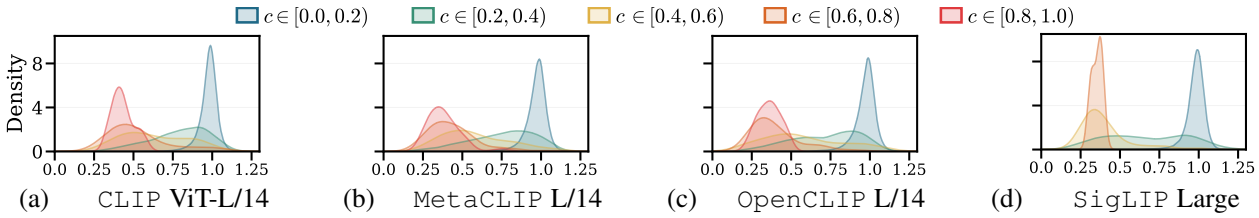


Figure 9. Distribution of cosine distances between image–text feature pairs estimated from embeddings of four VLMs. This figure presents the same analysis as Figure 2, extended to larger models.

E.2. Real-World Results Across Experimental Configurations

In Section 7.2, Table 1 reports results for a single VLM embedding and activation setup (`CLIP` ViT-B/32 embeddings with Top-K activation at $K = 32$). To verify whether the conclusions hold across diverse setups, we conduct extended experiments along three orthogonal aspects: (i) the sparsity level K in Top-K activation, (ii) the pre-trained VLM backbone (varying both architecture scale and pre-training corpus), and (iii) the choice of sparsifying activation function. All other settings follow Section 7.2, and we report averages over three independent runs. Across all setups, our method matches the lowest reconstruction error of modality-specific SAEs while delivering the strongest cross-modal retrieval performance, same with the conclusions of Table 1.

Robustness to sparsity level K. We first vary the Top-K sparsity level on `CLIP` ViT-B/32 embeddings. Specifically, we sweep $K \in \{16, 32, 64\}$ while keeping all other hyperparameters identical to the main experiment in Section 7.2; the $K = 32$ result is reported in Table 1, and the remaining levels in Table 2. Across all three levels, our method preserves the lowest reconstruction error inherited from modality-specific SAEs, and $K = 32$ emerges as the best operating point for cross-modal retrieval, where our method peaks at Image-to-Text Recall@1 of 16.0. The gain over the strongest baseline is largest in the sparser regime: at $K = 16$, our method more than doubles its performance in both retrieval directions, whereas at $K = 64$ the auxiliary-loss baselines benefit from the larger active budget and narrow the gap – though this reflects our own scores dropping from the $K = 32$ peak rather than the baselines surpassing us, as our method still leads on Text-to-Image Recall and on top-10 retrieval in both directions. Overall, $K = 32$ is the optimal operating point for cross-modal retrieval, and at this level our method consistently outperforms all baselines.

Robustness across vision-language model backbones. We next vary the pre-trained VLM backbone while fixing $K = 32$ with Top-K activation. Specifically, we span two architecture scales (ViT-B/32 and ViT-L/14) and two pre-training corpora (`CLIP` (Radford et al., 2021) and `OpenCLIP` (Cherti et al., 2023)). We keep all other hyperparameters identical to the main experiment in Section 7.2. Specifically, for each backbone we scale the SAE width m with the embedding dimension d to hold the expansion ratio m/d fixed, so that comparisons are not confounded by SAE capacity. The results are reported in Table 3, with the `CLIP` ViT-B/32 setting in Table 1. Across all backbones, our method retains the lowest reconstruction error while attaining the strongest cross-modal retrieval, confirming that the effect is not specific to a single architecture or pre-training corpus. Therefore, these results show that the advantage of our method holds across different VLM backbones.

Robustness to activation function. We replace the sparsifying activation function, changing from Top-K to BatchTop-K (Bussmann et al., 2024), which selects the top-K activations across the entire batch rather than per sample. We keep all other hyperparameters identical to the main experiment in Section 7.2. The results are reported in Table 4. Our method again attains the lowest reconstruction error and the strongest cross-modal retrieval. This confirms that the advantage of our method is robust to the choice of sparsifying activation function.

Table 2. Effect of Top-K sparsity level K on CLIP ViT-B/32 embeddings. Each section corresponds to a different value of K.

	MS-COCO (Lin et al., 2014)							ImageNet1K (Deng et al., 2009)	
	Recon. (↓)	Image-to-Text (↑)			Text-to-Image (↑)			Recon. (↓)	Zero-shot (↑)
Methods	MSE	R@1	R@5	R@10	R@1	R@5	R@10	MSE	Accuracy
$K = 16$									
Shared SAE	0.131	3.0 (±1.9)	7.7 (±3.3)	10.9 (±4.1)	3.0 (±0.3)	7.8 (±0.7)	11.3 (±1.1)	0.169	12.6 (±4.1)
+ Iso-Energy alignment loss	0.131	4.2 (±1.2)	9.6 (±2.8)	13.6 (±3.8)	3.1 (±0.2)	8.2 (±0.8)	12.1 (±1.3)	0.168	15.3 (±3.8)
+ Group-sparse loss	0.134	5.1 (±0.1)	12.2 (±0.2)	17.2 (±0.2)	2.9 (±0.1)	8.8 (±0.5)	12.9 (±0.5)	0.170	20.3 (±0.6)
Modality-Specific SAEs	**0.127**	0.0 (±0.0)	0.1 (±0.1)	0.2 (±0.1)	0.0 (±0.0)	0.1 (±0.1)	0.2 (±0.1)	**0.165**	0.1 (±0.1)
+ Post-hoc Alignment (Ours)	**0.127**	**13.3** (±0.7)	**29.4** (±0.3)	**38.9** (±0.7)	**9.6** (±0.3)	**23.0** (±0.5)	**31.8** (±0.5)	**0.165**	**22.9** (±0.5)
$K = 64$									
Shared SAE	0.063	9.8 (±1.3)	21.2 (±2.0)	28.0 (±2.9)	6.0 (±0.6)	15.2 (±1.2)	21.3 (±1.5)	**0.083**	15.0 (±1.5)
+ Iso-Energy alignment loss	0.063	9.0 (±0.8)	19.9 (±0.4)	26.7 (±0.5)	5.3 (±0.6)	13.4 (±1.5)	18.8 (±2.0)	0.084	16.3 (±1.2)
+ Group-sparse loss	0.091	**11.8** (±0.5)	**26.0** (±0.8)	34.6 (±1.1)	7.0 (±0.3)	18.4 (±0.5)	26.4 (±0.5)	0.116	**33.0** (±0.4)
Modality-Specific SAEs	**0.062**	0.0 (±0.0)	0.1 (±0.0)	0.2 (±0.1)	0.0 (±0.0)	0.1 (±0.0)	0.3 (±0.1)	**0.083**	0.1 (±0.0)
+ Post-hoc Alignment (Ours)	**0.062**	11.2 (±0.8)	26.0 (±1.6)	**34.9** (±2.2)	**8.5** (±0.8)	**21.5** (±2.1)	**29.7** (±2.8)	**0.083**	19.3 (±1.2)

Table 3. Effect of the size and type of the pre-trained VLM backbone. Each section corresponds to a different backbone, with the latent dimension of SAE m is shown in the header.

	MS-COCO (Lin et al., 2014)							ImageNet1K (Deng et al., 2009)	
	Recon. (↓)	Image-to-Text (↑)			Text-to-Image (↑)			Recon. (↓)	Zero-shot (↑)
Methods	MSE	R@1	R@5	R@10	R@1	R@5	R@10	MSE	Accuracy
CLIP ViT-L/14 ($m = 12288$)									
Shared SAE	0.119	18.2 (±0.2)	33.6 (±0.4)	41.8 (±0.7)	9.9 (±0.8)	22.9 (±0.9)	30.7 (±1.2)	0.155	38.9 (±0.6)
+ Iso-Energy alignment loss	0.119	16.8 (±1.1)	31.2 (±1.2)	39.3 (±1.7)	9.4 (±0.5)	21.3 (±0.9)	28.9 (±1.2)	**0.154**	39.4 (±0.8)
+ Group-sparse loss	0.137	14.2 (±0.2)	29.6 (±0.7)	38.5 (±1.0)	8.2 (±0.4)	20.3 (±0.4)	28.4 (±0.3)	0.176	**45.0** (±0.3)
Modality-Specific SAEs	**0.117**	0.0 (±0.0)	0.1 (±0.0)	0.2 (±0.0)	0.0 (±0.0)	0.1 (±0.0)	0.2 (±0.1)	**0.154**	0.1 (±0.0)
+ Post-hoc Alignment (Ours)	**0.117**	**22.6** (±0.7)	**43.6** (±1.2)	**54.9** (±1.2)	**15.3** (±0.4)	**33.0** (±0.4)	**42.9** (±0.3)	**0.154**	35.9 (±1.1)
OpenCLIP ViT-B/32 ($m = 8192$)									
Shared SAE	0.116	8.9 (±0.3)	18.5 (±0.3)	24.4 (±0.5)	4.8 (±0.2)	12.2 (±0.2)	17.4 (±0.2)	0.149	18.7 (±0.7)
+ Iso-Energy alignment loss	0.117	9.2 (±0.8)	18.6 (±1.3)	24.5 (±1.2)	4.8 (±0.5)	12.3 (±0.8)	17.4 (±1.1)	0.149	18.9 (±1.3)
+ Group-sparse loss	0.141	10.1 (±0.4)	23.1 (±0.4)	31.8 (±0.5)	6.1 (±0.0)	17.1 (±0.2)	25.4 (±0.3)	0.176	**34.3** (±0.4)
Modality-Specific SAEs	**0.115**	0.0 (±0.0)	0.1 (±0.0)	0.2 (±0.1)	0.0 (±0.0)	0.1 (±0.0)	0.2 (±0.1)	**0.149**	0.1 (±0.0)
+ Post-hoc Alignment (Ours)	**0.115**	**21.0** (±0.9)	**41.2** (±1.1)	**52.0** (±1.0)	**11.2** (±0.1)	**26.5** (±0.6)	**36.1** (±0.9)	**0.149**	29.2 (±0.7)
OpenCLIP ViT-L/14 ($m = 12288$)									
Shared SAE	0.133	26.5 (±0.4)	45.5 (±1.8)	54.5 (±2.1)	15.8 (±0.8)	32.0 (±1.5)	40.6 (±1.7)	**0.164**	50.9 (±0.7)
+ Iso-Energy alignment loss	0.134	27.1 (±0.4)	46.2 (±1.1)	55.5 (±0.7)	16.2 (±0.5)	32.9 (±0.9)	41.6 (±0.8)	0.166	50.3 (±0.5)
+ Group-sparse loss	0.162	23.7 (±0.2)	43.3 (±0.5)	53.9 (±0.1)	13.1 (±0.3)	29.1 (±0.4)	38.6 (±0.5)	0.196	**53.9** (±0.3)
Modality-Specific SAEs	**0.129**	0.1 (±0.0)	0.1 (±0.1)	0.2 (±0.1)	0.0 (±0.0)	0.1 (±0.1)	0.2 (±0.1)	**0.164**	0.1 (±0.0)
+ Post-hoc Alignment (Ours)	**0.129**	**27.4** (±2.2)	**50.6** (±1.8)	**62.0** (±1.4)	**18.1** (±2.0)	**37.6** (±3.0)	**47.9** (±3.1)	**0.164**	42.4 (±0.4)

Table 4. Effect of sparsifying activation function on CLIP ViT-B/32 ($L = 8192$, $K = 32$). The Top-K setting is reported in Table 1; this table reports the BatchTop-K (Bussmann et al., 2024) variant.

	MS-COCO (Lin et al., 2014)							ImageNet1K (Deng et al., 2009)	
	Recon. (↓)	Image-to-Text (↑)			Text-to-Image (↑)			Recon. (↓)	Zero-shot (↑)
Methods	MSE	R@1	R@5	R@10	R@1	R@5	R@10	MSE	Accuracy
BatchTop-K									
Shared SAE	0.089	6.3 (±0.6)	15.9 (±1.5)	22.0 (±1.7)	3.9 (±0.6)	11.3 (±0.9)	16.8 (±1.0)	0.115	16.4 (±3.1)
+ Iso-Energy alignment loss	0.089	5.8 (±0.8)	14.8 (±0.8)	20.4 (±0.9)	3.3 (±1.4)	9.3 (±3.3)	13.9 (±3.9)	0.115	13.0 (±1.8)
+ Group-sparse loss	0.105	7.8 (±0.3)	19.0 (±0.4)	26.2 (±0.4)	4.9 (±0.0)	13.7 (±0.1)	20.6 (±0.1)	0.132	**28.0** (±0.7)
Modality-Specific SAEs	**0.088**	0.0 (±0.0)	0.0 (±0.0)	0.1 (±0.0)	0.0 (±0.0)	0.1 (±0.0)	0.2 (±0.0)	**0.114**	0.1 (±0.0)
+ Post-hoc Alignment (Ours)	**0.088**	**16.2** (±0.8)	**34.6** (±1.0)	**44.7** (±1.4)	**10.5** (±0.6)	**25.5** (±0.8)	**35.0** (±0.9)	**0.114**	22.8 (±1.0)