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

Anonymous Author(s)
Affiliation
Address
email

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. To address this misalignment, we propose an approach that trains sparse autoencoders to preserve the unique feature geometry of each modality and aligns corresponding features post hoc. Our method improves reconstruction fidelity and enhances performance in cross-modal retrieval and concept steering.

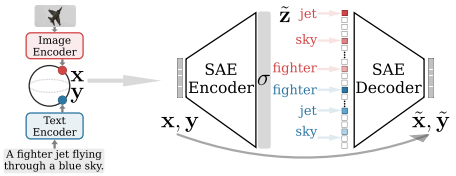
1 Introduction

Vision-language models (VLMs) map images and text into a joint representation space, which supports a wide range of downstream tasks [42, 54, 6] and serves as the foundation for generative VLMs [30, 46]. Despite their empirical success, the joint representation remains difficult to interpret [36]. 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.

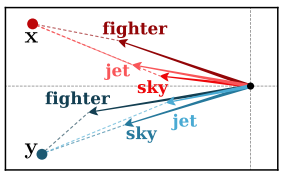
The linear representation hypothesis [16, 39] provides a useful framework for studying this question, positing that each embedding can be expressed as a linear combination of a small number of *monosemantic* features. Sparse autoencoders (SAEs) operationalize this hypothesis by mapping each embedding to a *sparse latent code*, whose active coordinates indicate which monosemantic features are present in the embedding. Building on their success in unimodal settings [23, 50, 18, 34], recent work has applied SAEs to joint embedding spaces to recover monosemantic features shared across image and text modalities [11, 55, 56, 38, 37, 25, 14].

When applied to joint embeddings of VLMs, prior work [25] 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 [25, 14]. 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.

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.



(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 within the embedding space.

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 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 SAEs to preserve the unique feature geometry of each modality and then aligns the corresponding features post hoc.

Our contributions are summarized below.

- In Section 3, we characterize *cross-modal feature heterogeneity*, where the same concept can have different feature directions across modalities in the joint embedding space.
- In Section 4, 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 5, 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 6, we evaluate our approach, showing that preserving reconstruction quality is important for improving performance on cross-modal tasks, including retrieval and concept steering.

Due to space constraints, a detailed comparison with related work is provided in Appendix A.

2 Preliminaries

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

Embeddings as linear combinations of features. Following the linear representation hypothesis [3, 16, 39, 12], 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, 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, and we assume that they 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:

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

Unlike prior work [25], 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 further assume that the coordinates of the latent code $\mathbf{z}$ are independent and sparse. Specifically, the sparse factor 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 [23, 50, 18, 44, 55, 12]. 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 $\mathbf{W} := [\mathbf{w}_1 \cdots \mathbf{w}_m] \in \mathbb{R}^{d\times m}$ denote the weight matrix of an SAE. 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

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

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

$$\tilde{\mathbf{x}}(\mathbf{x}) := \mathbf{W}\tilde{\mathbf{z}}(\mathbf{x}) = \textstyle\sum_{j\in[m]} \tilde{z}_j(\mathbf{x})\ \mathbf{w}_j, \qquad \tilde{\mathbf{y}}(\mathbf{y}) := \mathbf{W}\tilde{\mathbf{z}}(\mathbf{y}) = \textstyle\sum_{j\in[m]} \tilde{z}_j(\mathbf{y})\ \mathbf{w}_j. \tag{3}$$

The SAE is trained to reconstruct each embedding using a reconstruction loss. For image embeddings,

$$\mathcal{L}_{\mathrm{rec}}(\mathbf{W}; \mathbf{\Phi}) := \mathbb{E}_{\mathbf{x}} \|\mathbf{x} - \tilde{\mathbf{x}}(\mathbf{x})\|_2^2 = \mathbb{E}_{\mathbf{z}} \left\| \mathbf{\Phi}\mathbf{z} - \mathbf{W}\sigma\big(\mathbf{W}^\top \mathbf{\Phi}\mathbf{z}\big) \right\|_2^2, \tag{4}$$

and we use $\mathcal{L}_{\mathrm{rec}}(\mathbf{W}; \mathbf{\Psi})$ for text embeddings. After optimizing the SAE, the encoder extracts sparse latent codes, and the decoder columns provide estimates of monosemantic feature directions [12].

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 (3). 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.

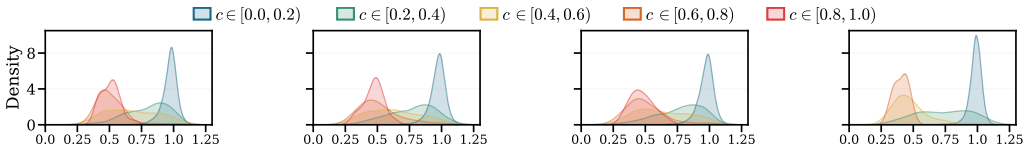
3 Cross-Modal Feature Heterogeneity in Joint Embedding Spaces

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., $\boldsymbol{\phi}_i \neq \boldsymbol{\psi}_i$.*

This definition separates semantic correspondence from directional identity in the joint embedding space. In other words, 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.

[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.



(a) CLIP [42] VIT-B/32 (b) MetaCLIP [8] B/32 (c) OpenCLIP [7] B/32 (d) SigLIP [51] Base

Figure 2: Distribution of cosine distances between image–text feature pairs estimated from embeddings of four VLMs. We group image–text feature pairs by their coactivation correlation $c_{i,j}$ in (5), 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 paired image–text embeddings, so pairs with larger correlations are more likely to represent the same shared concept. Across all models, the distribution does not concentrate near 0 but remains centered around a positive value, even for high-correlation pairs ($c \in [0.8, 1.0)$, shown in blue). This observation supports the presence of cross-modal feature heterogeneity.

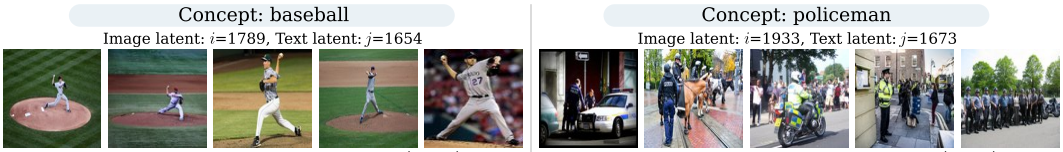


Figure 3: Image examples with high coactivation correlation $c_{i,j}$ from the CLIP setting in Figure 2, where i and j denote the image and text latent indices, respectively. The images reveal coherent concepts such as *baseball* and *policeman*, but the features $\hat{\phi}_i$ and $\hat{\psi}_j$ remain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

Measuring differences in feature directions across modalities. Since the true feature directions are unobserved, we use decoder columns as their estimates, 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:

$$\mathbf{C} := \mathrm{Corr}\big(\tilde{\mathbf{z}}(\mathbf{x}), \tilde{\mathbf{z}}(\mathbf{y})\big) \in \mathbb{R}^{m \times m}, \qquad c_{i,j} := [\mathbf{C}]_{i,j}. \tag{5}$$

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. For each pair (i, j), we measure the cosine distance between the estimated feature directions $\hat{\phi}_i$ and $\hat{\psi}_j$.

Figure 2 reports the distribution of cosine distances between estimated image and text feature directions, grouped by coactivation correlation. The results are computed on MS-COCO [29] embeddings extracted from four VLMs, including CLIP [42], MetaCLIP [8], OpenCLIP [7], and SigLIP2 [51]. 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 E.1 for experimental details.

Figure 3 shows examples where matched image and text latents strongly coactivate and respond to the same concept, such as *baseball* or *policeman*. For example, the *baseball* concept activates the image latent (with index $i = 1789$) and the text latent (with index $j = 1654$), yielding a high coactivation correlation, $c_{i,j} = 0.67$. But the cosine distance between the estimated features $\hat{\phi}_i$ and $\hat{\psi}_j$ remains large at 0.42. These examples show that semantic correspondence need not imply directional identity.

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

4 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 across modalities to shared latent coordinates. These goals can conflict under cross-modal feature heterogeneity. Corollary 1 explains why reconstruction induces modality split,

Theorem 1 studies the limited capacity case, and Proposition 1 shows why existing alignment methods reduce modality split by sacrificing reconstruction quality. Proofs are provided in Appendix D.

Why reconstruction induces modality split. We consider an SAE trained with the reconstruction loss (4) on both image and text embeddings. Since no alignment term is used, the objective only encourages accurate reconstruction.

Corollary 1. *Suppose an SAE has $m \geq 2n$ latent coordinates and is trained by minimizing $\mathcal{L}_{\mathrm{rec}}(\mathbf{W}; \mathbf{\Phi}) + \mathcal{L}_{\mathrm{rec}}(\mathbf{W}; \mathbf{\Psi})$. As the sparse factor s in (2) approaches 1, for any permutation matrix $\mathbf{P} \in \mathbb{R}^{m \times m}$, the weight matrix $\hat{\mathbf{W}} := [\mathbf{\Phi} \quad \mathbf{\Psi} \quad \mathbf{0}_{d \times (m-2n)}]\,\mathbf{P}$ is a global minimizer, and $\mathcal{L}_{\mathrm{rec}}(\hat{\mathbf{W}}; \mathbf{\Phi}) + \mathcal{L}_{\mathrm{rec}}(\hat{\mathbf{W}}; \mathbf{\Psi}) = 0$.*

Corollary 1 shows that reconstruction is minimized by assigning a separate column of the decoder weight matrix $\mathbf{W}$ to each feature direction. Thus, if the image feature $\boldsymbol{\phi}_i$ and the text feature $\boldsymbol{\psi}_i$ represent the same concept but point in different directions, they are assigned to different columns, say $\hat{\mathbf{w}}_a$ and $\hat{\mathbf{w}}_b$, and activate different latent coordinates, $\tilde{z}_a$ and $\tilde{z}_b$. This explains why modality split can be reconstruction optimal rather than a failure of concept recovery. The SAE preserves both feature directions, but the corresponding image and text concepts appear at different latent indices.

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 [50, 18, 32], *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 has $m < 2n$ latent coordinates and is trained by minimizing $\mathcal{L}_{\mathrm{rec}}(\mathbf{W}; \mathbf{\Phi}) + \mathcal{L}_{\mathrm{rec}}(\mathbf{W}; \mathbf{\Psi})$. 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 of a matrix. As the sparse factor s in (2) approaches 1, a global minimizer $\hat{\mathbf{W}} := [\hat{\mathbf{W}}_1 \ \cdots \ \hat{\mathbf{W}}_m]$ is obtained by taking $\hat{\mathbf{W}}_j$ to be a top eigenvector with unit norm of $\sum_{i \in \mathbb{A}_j} \mathbf{M}_i$ for each $j \in [m]$, and $\mathcal{L}_{\mathrm{rec}}(\hat{\mathbf{W}}; \mathbf{\Phi}) + \mathcal{L}_{\mathrm{rec}}(\hat{\mathbf{W}}; \mathbf{\Psi}) = s^{n-1}(1-s)\Big(2\sum_{i \in [n]} \mathbb{E}\big[z_i^2 \mid z_i \neq 0\big] - \sum_{j \in [m]} \lambda_{\max}\Big(\sum_{i \in \mathbb{A}_j} \mathbf{M}_i\Big)\Big) + o(1-s)$.*

Theorem 1 shows that when an SAE lacks enough effective coordinates, it must reuse columns of the decoder weight matrix $\mathbf{W}$ to represent multiple feature directions. In this case, feature directions are grouped by their geometry in the embedding space rather than by semantic correspondence. Hence, close but distinct directions may share a column, even when they correspond to different concepts.

Under this limited capacity regime, when $\boldsymbol{\phi}_i$ and $\boldsymbol{\psi}_i$ are close but not identical, the two feature directions may be grouped together and represented by a single column. This can reduce modality split by assigning them to a shared coordinate. However, it comes at the cost of worse reconstruction and poorer feature recovery, because a single column must approximate two distinct directions.

Why alignment during training sacrifices reconstruction. The previous results show that reconstruction favors separate columns, while alignment favors shared coordinates. We examine this tension in existing methods that add auxiliary alignment losses during training [14, 25].

For clarity, we state a simplified result. In this result, $\lambda^\star(\rho)$ and $\lambda^\dagger(\rho)$ are critical alignment strengths at which the global minimizer changes. Their exact definitions, along with complete statements for the group-sparse loss [25] and the Iso-Energy alignment loss [14], are given in Propositions 3 and 4.

Proposition 1. *Consider $n = 1$ and define $\rho := \boldsymbol{\phi}^\top \boldsymbol{\psi} \in (0, 1)$. Suppose the SAE has $m = 2$ latent coordinates, and each column of $\mathbf{W}$ has unit norm. For the group-sparse loss [25], where λ is the weight of the auxiliary alignment term, a global minimizer is $[\boldsymbol{\phi} \ \boldsymbol{\psi}]$ when $\lambda \in \big[0, \lambda^\star(\rho)\big)$, $\big[\frac{\boldsymbol{\phi}+\boldsymbol{\psi}}{\|\boldsymbol{\phi}+\boldsymbol{\psi}\|} \ \ \mathbf{0}_d\big]$ when $\lambda \in \big(\lambda^\star(\rho), \lambda^\dagger(\rho)\big)$, and satisfies $\sigma(\mathbf{W}^\top \boldsymbol{\phi}) = \sigma(\mathbf{W}^\top \boldsymbol{\psi}) = \mathbf{0}$ when $\lambda \in \big(\lambda^\dagger(\rho), \infty\big)$. The reconstruction losses for each λ regime are 0, $(1-\rho)\mathbb{E}[z^2]$, and $2\mathbb{E}[z^2]$, respectively.*

Proposition 1 illustrates the trade-off between reconstruction and alignment. Small λ keeps two separate columns, $[\boldsymbol{\phi} \ \boldsymbol{\psi}]$, yielding zero reconstruction loss but preserving modality split. Larger λ collapses the two directions into a single shared column of the decoder weight matrix, $\frac{\boldsymbol{\phi}+\boldsymbol{\psi}}{\|\boldsymbol{\phi}+\boldsymbol{\psi}\|}$,

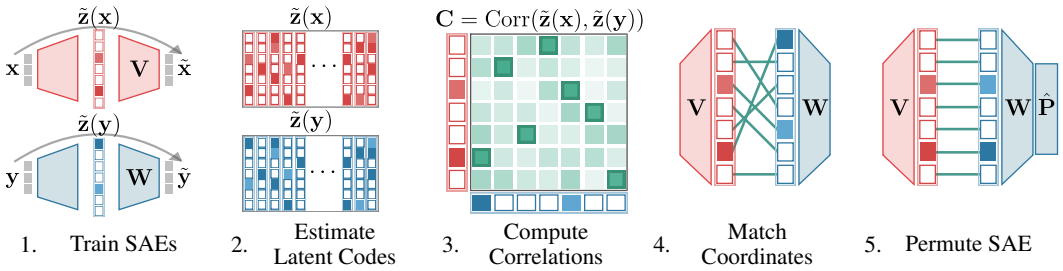


Figure 4: Overview of our approach. We first train modality-specific SAEs to reconstruct image and text embeddings, obtaining latent codes $\tilde{\mathbf{z}}(\mathbf{x})$ and $\tilde{\mathbf{z}}(\mathbf{y})$. We then compute their correlation matrix $\mathbf{C}$ in (5) over paired training data and apply the Hungarian algorithm to obtain the assignment $\hat{\mathbf{P}}$ in (6). 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.

improving alignment but incurring nonzero reconstruction loss. For the group-sparse loss, an even larger λ leads to a degenerate solution where both features become inactive.

5 Modality-Specific Sparse Autoencoders and Post-Hoc Alignment

Motivated by the analysis in Section 4, 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 (4). We minimize $\mathcal{L}_{\mathrm{rec}}(\mathbf{V}; \mathbf{\Phi})$ and $\mathcal{L}_{\mathrm{rec}}(\mathbf{W}; \mathbf{\Psi})$ 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 (5) to identify corresponding image and text latents. Specifically, we find a permutation matrix $\hat{\mathbf{P}}$ that maximizes the total correlation between aligned coordinates,

$$\hat{\mathbf{P}} \in \underset{\mathbf{P} \in \mathcal{P}_m}{\arg\max}\, \mathrm{tr}(\mathbf{C}\mathbf{P}), \tag{6}$$

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 [26][2]. We then apply the permutation only to the SAE for the text modality as

$$\tilde{\mathbf{z}}(\mathbf{y}; \mathbf{W}) := \sigma\big(\hat{\mathbf{P}}^\top \mathbf{W}^\top \mathbf{y}\big), \qquad \tilde{\mathbf{y}}(\mathbf{y}; \mathbf{W}) := \mathbf{W}\hat{\mathbf{P}}\tilde{\mathbf{z}}(\mathbf{y}; \mathbf{W}). \tag{7}$$

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.

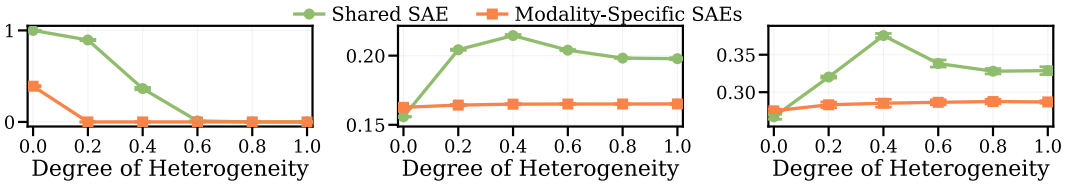
Inference. Embeddings are encoded into latent codes using modality-specific SAEs, and the permutation is applied to align their indices as in (7). This alignment enables cross-modal tasks such as retrieval in the aligned latent space, while encoding and decoding remain modality-specific.

6 Experiments

This section validates the theoretical findings in Section 4 and evaluates the approach proposed in Section 5. Section 6.1 uses synthetic embeddings where ground-truth feature directions and cross-modal correspondences are known, allowing us to directly test the theoretical predictions. Section 6.2 evaluates our method on real-world datasets to assess its effectiveness on cross-modal tasks.

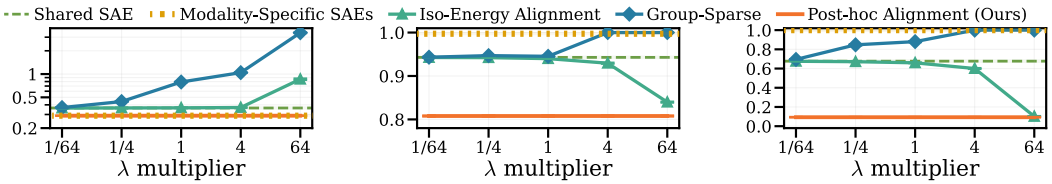
The codes used in our experiments are available in the anonymous repository.

[2]Dead latent coordinates are excluded in practice. Thus, (6) 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 (orange) as the cross-modal feature distance 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 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) Alignment error (↓) (c) Feature alignment error (↓)

Figure 6: Comparison between Post-hoc Alignment method (orange) and two auxiliary-loss baselines: Iso-Energy alignment [14] (teal) and group-sparse loss [25] (blue). The coefficient of auxiliary-loss λ 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.

6.1 Validation on Synthetic Embeddings

We validate two key design choices of the method proposed in Section 5. 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 4.

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 [14, 25]. We report the average over three independent runs. Details on synthetic embedding generation and training setup are provided in Appendix E.2.

Metrics. We evaluate each method along three axes: *Reconstruction Error* and *Feature Recovery Error* for reconstruction quality; *Alignment Error* and *Feature Alignment 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 E.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 between corresponding cross-modal features ϕ_i and ψ_i. As shown 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 4, where representing different feature directions with a single decoder column increases

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` [29], and by top-1 zero-shot classification accuracy on `ImageNet1K` [13]. The best and second-best results for each metric are shown in bold and underlined, respectively.

	MS-COCO [29]							ImageNet1K [13]	
	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	<u>7.1</u> (±0.1)	<u>16.7</u> (±0.4)	<u>23.8</u> (±0.3)	<u>4.3</u> (±0.1)	<u>12.2</u> (±0.1)	<u>18.3</u> (±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**	<u>25.1</u> (±0.3)

reconstruction loss. At zero cosine distance, the image and text feature directions coincide, so there is no cross-modal feature heterogeneity. The shared SAE can therefore represent the same concept with a single direction, which explains its lower reconstruction and feature recovery errors in this idealized case. Overall, these results show that modality-specific SAEs better preserve feature directions when semantically corresponding features differ across modalities, leading to better feature recovery. Appendix F.2 further shows that this phenomenon also appears in SAEs trained on VLM embeddings.

Post-hoc alignment improves alignment without sacrificing feature recovery. Figure 6 reports the results obtained by varying the auxiliary-loss strength 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 [14], increasing λ improves cross-modal alignment but degrades feature recovery, consistent with the reconstruction-alignment trade-off shown in Proposition 4. For the group-sparse loss [25], increasing λ eventually pushes the SAE into a degenerate regime where both feature recovery and alignment deteriorate, consistent with Proposition 3. 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.

6.2 Evaluation on Real-World Data

We evaluate our method on real-world image-text 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 [25], we evaluate whether estimated latent codes $\tilde{\mathbf{z}}$ in (5) provide useful representations for cross-modal retrieval, concept steering, and monosemanticity evaluation.

Experimental protocol. We use `CC-3M` [48] as a paired image-text dataset. From this dataset, we extract image and text embeddings using the `CLIP` model [42] with the ViT-B/32 architecture, and train SAEs on the resulting paired embeddings. We compare the method proposed in Section 5 with two auxiliary-loss baselines [14, 25]. 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 [38] and report averages over three runs. Details are provided in Appendix E.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` [13], 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 and families are provided in Appendix F.

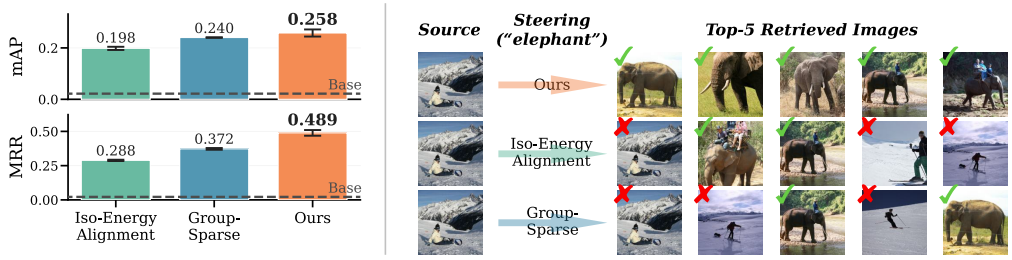


Figure 7: Image retrieval under concept latent steering on MS-COCO. We steer source image embeddings using target feature directions identified through aligned latent coordinates. Retrieval performance is measured by mean average precision (mAP) and mean reciprocal rank (MRR) over retrieved images containing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

Our method enables more effective cross-modal concept steering. We next 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, even though the source images do not.

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 latent codes are activated by semantically coherent inputs. We use the monosemanticity score [37] 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 [8] as the external encoder. Larger values indicate that the latent responds to a more semantically coherent concept.

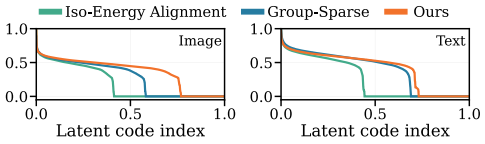


Figure 8: Monosemanticity scores for coordinates sorted in descending order, shown separately for image and text modalities.

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.

7 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. Overall, effective alignment in joint embedding spaces should preserve each modality's feature geometry.

Due to space constraints, a discussion of limitations and future directions is provided in Appendix B.

References

[1] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations*, 2017.

[2] Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Conference on Neural Information Processing Systems*, 2024.

[3] Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics*, 6:483–495, 07 2018.

[4] Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. Sega: Instructing text-to-image models using semantic guidance. *Conference on Neural Information Processing Systems*, 2023.

[5] Bart Bussmann, Patrick Leask, and Neel Nanda. Batchtopk sparse autoencoders. *arXiv preprint arXiv:2412.06410*, 2024.

[6] Zixiang Chen, Yihe Deng, Yuanzhi Li, and Quanquan Gu. Understanding transferable representation learning and zero-shot transfer in CLIP. In *International Conference on Learning Representations*, 2024.

[7] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Conference on Computer Vision and Pattern Recognition*, 2023.

[8] Yung-Sung Chuang, Yang Li, Dong Wang, Ching-Feng Yeh, Kehan Lyu, Ramya Raghavendra, James R. Glass, LIFEI HUANG, Jason E Weston, Luke Zettlemoyer, Xinlei Chen, Zhuang Liu, Saining Xie, Wen tau Yih, Shang-Wen Li, and Hu Xu. Meta CLIP 2: A worldwide scaling recipe. In *Conference on Neural Information Processing Systems*, 2026.

[9] Bilal Chughtai, Lawrence Chan, and Neel Nanda. A toy model of universality: Reverse engineering how networks learn group operations. In *International Conference on Machine Learning*, 2023.

[10] Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Conference on Neural Information Processing Systems*, 2023.

[11] Valérie Costa, Thomas Fel, Ekdeep Singh Lubana, Bahareh Tolooshams, and Demba E. Ba. From flat to hierarchical: Extracting sparse representations with matching pursuit. In *Conference on Neural Information Processing Systems*, 2026.

[12] Jingyi Cui, Qi Zhang, Yifei Wang, and Yisen Wang. On the limits of sparse autoencoders: A theoretical framework and reweighted remedy. In *International Conference on Learning Representations*, 2026.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition*, 2009.

[14] Grégoire DHIMOÏLA, Thomas Fel, Victor Boutin, and Agustin Martin Picard. Cross-modal redundancy and the geometry of vision–language embeddings. In *International Conference on Learning Representations*, 2026.

[15] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits, 2021. URL `https://transformer-circuits.pub/2021/framework/index.html`.

[16] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint*, 2022.

[17] Sedigheh Eslami and Gerard de Melo. Mitigate the gap: Improving cross-modal alignment in CLIP. In *International Conference on Learning Representations*, 2025.

[18] Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations*, 2025.

[19] Ian J Goodfellow, Aaron Courville, and Yoshua Bengio. Large-scale feature learning with spike-and-slab sparse coding. In *International Coference on International Conference on Machine Learning*, 2012.

[20] Eleonora Grassucci, Giordano Cicchetti, Emanuele Frasca, Aurelio Uncini, and Danilo Comminiello. Closing the modality gap aligns group-wise semantics. In *International Conference on Learning Representations*, 2026.

[21] Ruben Härle, Felix Friedrich, Manuel Brack, Stephan Wäldchen, Björn Deiseroth, Patrick Schramowski, and Kristian Kersting. Measuring and guiding monosemanticity. In *Conference on Neural Information Processing Systems*, 2025.

[22] John Hewitt and Christopher D Manning. A structural probe for finding syntax in word representations. In *North American Chapter of the Association for Computational Linguistics*, 2019.

[23] Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, 2023.

[24] Qian Jiang, Changyou Chen, Han Zhao, Liqun Chen, Qing Ping, Son Dinh Tran, Yi Xu, Belinda Zeng, and Trishul Chilimbi. Understanding and constructing latent modality structures in multi-modal representation learning. In *Conference on Computer Vision and Pattern Recognition*, 2023.

[25] Chiraag Kaushik, Davis Barch, and Andrea Fanelli. Decomposing multimodal embedding spaces with group-sparse autoencoders. In *International Conference on Learning Representations*, 2026.

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

[27] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Conference on Neural Information Processing Systems*, 2023.

[28] Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Conference on Neural Information Processing Systems*, 2022.

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014.

[30] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Conference on Neural Information Processing Systems*, 2023.

[31] Sheng Liu, Haotian Ye, and James Zou. Reducing hallucinations in large vision-language models via latent space steering. In *International Conference on Learning Representations*, 2025.

[32] Lu Lu, Yeonjong Shin, Yanhui Su, and George Em Karniadakis. Dying relu and initialization: Theory and numerical examples. *Communications in Computational Physics*, 28(5):1671–1706, 2020.

[33] Alireza Makhzani and Brendan Frey. K-sparse autoencoders. In *International Conference on Learning Representations*, 2014.

[34] Samuel Marks, Can Rager, Eric J Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *International Conference on Learning Representations*, 2025.

[35] Marco Mistretta, Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Andrew D. Bagdanov. Cross the gap: Exposing the intra-modal misalignment in CLIP via modality inversion. In *International Conference on Learning Representations*, 2025.

[36] Tuomas Oikarinen and Tsui-Wei Weng. CLIP-dissect: Automatic description of neuron representations in deep vision networks. In *International Conference on Learning Representations*, 2023.

[37] Mateusz Pach, Shyamgopal Karthik, Quentin Bouniot, Serge Belongie, and Zeynep Akata. Sparse autoencoders learn monosemantic features in vision-language models. In *Conference on Neural Information Processing Systems*, 2026.

[38] Isabel Papadimitriou, Huangyuan Su, Thomas Fel, Sham M. Kakade, and Stephanie Gil. Interpreting the linear structure of vision-language model embedding spaces. In *Conference on Language Modeling*, 2025.

[39] Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *International Conference on Machine Learning*, 2024.

[40] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *International Conference on Computer Vision*, 2021.

[41] Qi Qian, Yuanhong Xu, and Juhua Hu. Intra-modal proxy learning for zero-shot visual categorization with clip. In *Conference on Neural Information Processing Systems*, 2023.

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

[43] Sameera Ramasinghe, Violetta Shevchenko, Gil Avraham, and Ajanthan Thalaiyasingam. Accept the modality gap: An exploration in the hyperbolic space. In *Conference on Computer Vision and Pattern Recognition*, 2024.

[44] Sukrut Rao, Sweta Mahajan, Moritz Böhle, and Bernt Schiele. Discover-then-name: Task-agnostic concept bottlenecks via automated concept discovery. In *European Conference on Computer Vision*, 2024.

[45] Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *Annual Meeting of the Association for Computational Linguistics*, 2024.

[46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Conference on Computer Vision and Pattern Recognition*, 2022.

[47] Simon Schrodi, David T. Hoffmann, Max Argus, Volker Fischer, and Thomas Brox. 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.

[48] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Association for Computational Linguistics*, 2018.

[49] Abdul-Saboor Sheikh, Jacquelyn A Shelton, and Jörg Lücke. A truncated em approach for spike-and-slab sparse coding. *Journal of Machine Learning Research*, 15(1):2653–2687, 2014.

[50] Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L. Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet, 2024. URL `https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html`.

[51] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint*, 2025.

[52] Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations*, 2023.

[53] Shin'ya Yamaguchi, Dewei Feng, Sekitoshi Kanai, Kazuki Adachi, and Daiki Chijiwa. Post-pre-training for modality alignment in vision-language foundation models. In *Conference on Computer Vision and Pattern Recognition*, 2025.

[54] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022.

[55] Vladimir Zaigrajew, Hubert Baniecki, and Przemyslaw Biecek. Interpreting CLIP with hierarchical sparse autoencoders. In *International Conference on Machine Learning*, 2025.

[56] Kaichen Zhang, Yifei Shen, Bo Li, and Ziwei Liu. Large multi-modal models can interpret features in large multi-modal models. In *International Conference on Computer Vision*, 2025.

[57] Yuhui Zhang, Elaine Sui, and Serena Yeung. Connect, collapse, corrupt: Learning cross-modal tasks with uni-modal data. In *International Conference on Learning Representations*, 2024.

A 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 [10, 15, 52, 9]. A key assumption underlying this line of work is the *linear representation hypothesis* [16, 39], 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 [1, 22] and steering interventions [27, 2, 45]. Recent work has begun extending this hypothesis to multimodal representations [40, 4, 31], opening avenues for understanding feature structure across different modalities.

Sparse autoencoders for interpreting representations. SAEs have shown great promise for extracting monosemantic features from embeddings of language models [23, 50, 18, 34, 21]. Building on these advances, SAEs have been further applied to VLMs [44, 55, 37], especially joint embedding models trained with contrastive objectives [42]. Recent work has further examined how these features align or separate across modalities [11, 55, 56, 38, 37], 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 [28]. Prior work has analyzed this gap from multiple geometric, distributional, and optimization-level perspectives [28, 57, 47]. One line of work argues that the modality gap partly reflects modality-specific features in multimodal representations, rather than merely a failure of alignment [24, 43, 41]. Another line focuses on cross-modal misalignment of shared features and studies how reducing such mismatch improves downstream task performance [17, 53, 20, 35].

Recent SAE-based analyses further reveal *modality split* [38], where the same concept activates different latent codes across modalities. Existing remedies force shared latent activations [25, 14], implicitly assuming feature directions are aligned (*i.e.,* $\boldsymbol{\phi}_i = \boldsymbol{\psi}_i$) 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.

B 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.

C 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

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

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 2. *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 $\mathbf{\Phi} = \mathbf{\Psi}$. Hence, the embeddings satisfy $\mathbf{x} = \mathbf{\Phi}\mathbf{z} = \mathbf{\Psi}\mathbf{z} = \mathbf{y}$ 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.

D Proofs of Theoretical Results

We begin by clarifying notation. We use $[n] := \{1, 2, \cdots, n\}$. 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.

Theorem 1. *Suppose the SAE has $m < 2n$ latent coordinates and is trained by minimizing $\mathcal{L}_{\mathrm{rec}}(\mathbf{W}; \mathbf{\Phi}) + \mathcal{L}_{\mathrm{rec}}(\mathbf{W}; \mathbf{\Psi})$. 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 of a matrix. As the sparse factor s in (2) approaches 1, a global minimizer $\hat{\mathbf{W}} := [\hat{\mathbf{W}}_1 \ \cdots \ \hat{\mathbf{W}}_m]$ is obtained by taking $\hat{\mathbf{W}}_j$ to be a top eigenvector with unit norm of $\sum_{i\in\mathbb{A}_j} \mathbf{M}_i$ for each $j \in [m]$, and $\mathcal{L}_{\mathrm{rec}}(\hat{\mathbf{W}}; \mathbf{\Phi}) + \mathcal{L}_{\mathrm{rec}}(\hat{\mathbf{W}}; \mathbf{\Psi}) = s^{n-1}(1-s)\Big(2\sum_{i\in[n]} \mathbb{E}\big[z_i^2 \mid z_i \neq 0\big] - \sum_{j\in[m]} \lambda_{\max}\Big(\sum_{i\in\mathbb{A}_j} \mathbf{M}_i\Big)\Big) + o(1-s)$.*

Proof. Let $\theta_i := \boldsymbol{\phi}_i$ and $\theta_{n+i} := \boldsymbol{\psi}_i$ for $i \in [n]$, and $\mu_i' := \mathbb{E}[z_i^2 \mid z_i \neq 0] > 0$, with $\mu_{n+i}' := \mu_i'$. By the same expansion as in (9), in the sparsity regime $s \to 1$,

$$\mathcal{L}_{\mathrm{rec}}(\mathbf{W}; \mathbf{\Phi}) + \mathcal{L}_{\mathrm{rec}}(\mathbf{W}; \mathbf{\Psi}) = s^{n-1}(1-s) \sum_{i\in[2n]} \mu_i' \, \|\theta_i - \mathbf{W}\sigma(\mathbf{W}^\top \theta_i)\|_2^2 + o(1-s). \quad (8)$$

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

$$\sum_{i\in[2n]} \mu_i' \, \|\theta_i - \mathbf{W}\sigma(\mathbf{W}^\top \theta_i)\|_2^2 = \sum_{i\in[2n]} \mu_i' \, \|\theta_i - \hat{\mathbf{W}}_i \hat{\mathbf{W}}_i^\top \theta_i\|_2^2 = \sum_{i\in[2n]} \mu_i' \big(1 - (\theta_i^\top \hat{\mathbf{W}}_i)^2\big).$$

Since $\sum_{i\in[2n]} \mu_i'$ is a constant, minimizing $\mathcal{L}_{\mathrm{rec}}(\mathbf{W}; \mathbf{\Phi}, \mathbf{\Psi})$ is equivalent to maximizing

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

For each $j \in [m]$, define $\mathbb{A}_j := \{i \in [2n] : \hat{\mathbf{W}}_i = \mathbf{W}_j\}$. Then $(\mathbb{A}_1, \cdots, \mathbb{A}_m)$ is 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, the objective decouples over j, and for each $j \in [m]$,

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

attained by the top eigenvector. Since every choice of $(\mathbf{W}_1, \ldots, \mathbf{W}_m)$ induces a partition of $[2n]$, the global minimization of the loss is equivalent to maximizing $\sum_{j \in [m]} \lambda_{\max}\left(\sum_{i \in \mathbb{A}_j} \mathbf{M}_i \right)$ over all partitions of $[2n]$ into m subsets. For the maximizing partition $(\mathbb{A}_1, \cdots, \mathbb{A}_m)$, taking $\hat{\mathbf{W}}_j$ to be a unit-norm top eigenvector of $\sum_{i \in \mathbb{A}_j} \mathbf{M}_i$ yields a global minimizer $\hat{\mathbf{W}}$. Substituting back into (8), together with $\sum_{i \in [2n]} \mu_i' = 2 \sum_{i \in [n]} \mu_i'$, gives the stated reconstruction error:

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

□

Corollary 1. *Suppose an SAE has $m \geq 2n$ latent coordinates and is trained by minimizing $\mathcal{L}_{\mathrm{rec}}(\mathbf{W}; \mathbf{\Phi}) + \mathcal{L}_{\mathrm{rec}}(\mathbf{W}; \mathbf{\Psi})$. As the sparse factor s in (2) approaches 1, for any permutation matrix $\mathbf{P} \in \mathbb{R}^{m \times m}$, the weight matrix $\hat{\mathbf{W}} := [\mathbf{\Phi} \;\; \mathbf{\Psi} \;\; \mathbf{0}_{d \times (m-2n)}]\, \mathbf{P}$ is a global minimizer, and $\mathcal{L}_{\mathrm{rec}}(\hat{\mathbf{W}};\, \mathbf{\Phi}) + \mathcal{L}_{\mathrm{rec}}(\hat{\mathbf{W}};\, \mathbf{\Psi}) = 0$.*

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

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

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

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

Since $s^{n-1}(1-s) > 0$ and $\mu_i' > 0$, minimizing $\mathcal{L}_{\mathrm{rec}}(\mathbf{W};\, \mathbf{\Phi}, \mathbf{\Psi})$ in the regime $s \to 1$ is asymptotically equivalent to enforcing

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

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

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

which shows that $\hat{\mathbf{W}}$ is a global minimizer with $\mathcal{L}_{\mathrm{rec}}(\hat{\mathbf{W}};\, \mathbf{\Phi}, \mathbf{\Psi}) = 0$. □

Proposition 3 (Complete statement of 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 has $m \geq 2$ latent coordinates, and each column of $\mathbf{W}$ has unit norm. Consider the loss [25]*

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

Define $\lambda^\dagger := \dfrac{\mathbb{E}[z^2]}{\mathbb{E}[z]}$, $\lambda^\star(\rho) := \frac{(1-\rho)\lambda^\dagger}{2-\sqrt{1+\rho}}$, and $\lambda^\ddagger(\rho) := \sqrt{1+\rho}\,\lambda^\dagger$.

Up to any permutation matrix $\mathbf{P}$, a global minimizer $\hat{\mathbf{W}}$ of the loss is:

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

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

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

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

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 $j_a \neq j_b$, setting $[\mathbf{W}]_{[:,j_a]} = \boldsymbol{\phi}$ and $[\mathbf{W}]_{[:,j_b]} = \boldsymbol{\psi}$ gives zero reconstruction error and group penalty 2; hence

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

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 $a^2 + b^2 = \mathbf{w}^\top \mathbf{M}\mathbf{w}$. Under $\rho \in (0, 1)$ 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

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

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 4. *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 has $m \geq 2$ latent coordinates, and each column of $\mathbf{W}$ has unit norm. Consider the loss [14]*

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

Define $\lambda^\star(\rho) := \frac{2(1-\rho)}{1+\rho}$.

Up to any permutation matrix $\mathbf{P}$, a global minimizer $\hat{\mathbf{W}}$ of the loss is:

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

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

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

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

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

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,

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

We now compare the three values:

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

Since $\lambda \ge 0$, we have $L_{\mathrm{dead}} = 2 > 0 = L_{\mathrm{sep}}$, so the dead solution is never optimal. Comparing the remaining two, $L_{\mathrm{sh}} < L_{\mathrm{sep}}$ 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

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

□

E Experimental Details

This section provides experimental details for the empirical analyses in Sections 3 and 6.

All experiments are conducted on a single NVIDIA A100 GPU.

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

This section provides experimental details for Figure 2 in Section 3, 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 (5) 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 [33] activation function with $K = 8$. We follow the training protocol of Papadimitriou et al. [38]. Specifically, we use the AdamW optimizer with learning rate $5{\times}10^{-4}$ and weight decay 10^{-5}, batch size 1024, and a cosine schedule with 5% linear warmup over 30 training epochs.

E.2 Implementation Details for Synthetic Data Experiments

This section provides the data-generation specification, SAE architectures, and optimization hyperparameters used in Section 6.1.

Dataset. Following prior work [19, 49], we generate each pair of synthetic embeddings in two steps. 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: $n_S = 1024$ shared concepts present in both modalities, $n_I = 512$ image-only concepts (with $\boldsymbol{\psi}_i = \mathbf{0}$), and $n_T = 512$ 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, giving us direct control over the latent structure of the generative process.

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

Modality-specific feature directions are sampled as unit-norm vectors with maximum pairwise interference bounded by $\epsilon_{\max} = 0.30$. 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 $\mathbf{x} = \mathbf{\Phi}\mathbf{z} + \boldsymbol{\epsilon}_I$ and $\mathbf{y} = \mathbf{\Psi}\mathbf{z} + \boldsymbol{\epsilon}_T$ via (1), with independent observation noise $\boldsymbol{\epsilon}_I, \boldsymbol{\epsilon}_T \sim \mathcal{N}(\mathbf{0}, \sigma_{\mathrm{obs}}^2 \mathbf{I})$ of standard deviation $\sigma_{\mathrm{obs}} = 0.05$. We generate $50{,}000$ paired embeddings for training and $10{,}000$ for evaluation, and repeat each experiment over three independent runs, reporting the average.

Training. The shared SAE has latent dimension $m = 8192$ and uses the Top-K activation function [33] with $K = 16$. Each modality-specific SAE uses latent dimension $m = 4096$, so that the total number of learnable parameters matches that of the shared SAE. The two baselines [14, 25] augment the shared SAE by training with an auxiliary loss.

Optimization. All SAEs are trained with AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, weight decay 0) at a constant learning rate of 5×10^{-4} with no warmup. We use a batch size of 256 and train for 10 epochs, corresponding to approximately $1{,}950$ optimization steps.

E.3 Evaluation Metrics for Synthetic Data Experiments

We provide formal definitions of the five metrics introduced in Section 6.1.

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 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) *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) require access to the ground-truth features $(\mathbf{\Phi}, \mathbf{\Psi})$ 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) *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:*

$$\frac{1}{n_{\mathrm{S}}} \sum_{i \in [n_{\mathrm{S}}]} \mathbf{1}\left[\operatorname*{arg\,max}_{j \in [m]} \cos\big([\mathbf{W}]_{:,j},\ \boldsymbol{\phi}_i\big) \;=\; \operatorname*{arg\,max}_{j \in [m]} \cos\big([\mathbf{W}]_{:,j},\ \boldsymbol{\psi}_i\big) \right] \;\in\; [0, 1].$$

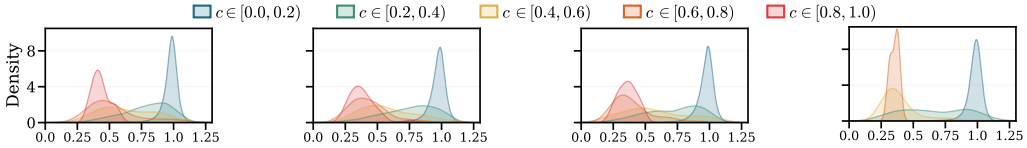
E.4 Implementation Details for Real-World Experiments

Implementation Details. We use the `CC-3M` [48] dataset with 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 [33] with $K = 32$, and we adopt the training protocol of Papadimitriou et al. [38].

Licenses of datasets. We use `CC-3M` [48] and `MS-COCO` [29] for our experiments. `CC-3M` is released by Google for free use under the terms of its repository license. `MS-COCO` annotations are released under `CC BY 4.0`.

F Additional Experimental Results

F.1 Evidence for Cross-Modal Feature Heterogeneity on Larger Backbones



(a) `CLIP` [42] VIT-L/14 (b) `MetaCLIP` [8] L/14 (c) `OpenCLIP` [7] L/14 (d) `SigLIP` [51] Large

Figure 9: Distribution of cosine distances between image–text feature pairs estimated from embeddings of four VLMs. We group image–text feature pairs by their co-activation correlation $c_{i,j}$ in (5), shown in different colors. The value $c_{i,j}$ measures how strongly the i-th image feature and the j-th text feature co-activate on paired image–text embeddings, so pairs with larger correlations are more likely to represent the same shared concept. Across all models, the distribution does not concentrate near 0 but 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 2 in the main text uses the base variants (ViT-B/32) of the four VLMs. To check whether the observed cross-modal feature heterogeneity persists at larger scales, we repeat the same analysis on the corresponding large variants (ViT-L/14) of `CLIP`, `MetaCLIP`, `OpenCLIP`, and `SigLIP2`, and report the results in Figure 9. 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 is not an artifact of model capacity and persists across backbone sizes.

F.2 Evidence of Feature Collapse on Real-World Data

We complement the synthetic results in Figure 5a with a real-world experiment that checks whether the feature collapse predicted by Theorem 1 actually occurs in SAEs trained on VLM embeddings.

In principle, SAEs use a large latent dimension m, so one might expect to be in the safe regime of Corollary 1 ($m \geq 2n$). Practically, however, a substantial fraction of latent coordinates are dead [50, 18], so the effective number of usable coordinates can be much smaller than m. This pushes the SAE toward the regime of Theorem 1, where directionally close cross-modal features are collapsed into a single feature. We conduct a simple diagnostic to detect such collapses on real data.

Setup. We follow the same protocol as Appendix E.1 on `CLIP` ViT-B/32 embeddings of `MS-COCO` [29], except that we now train a single *shared* SAE on the union of image and text embeddings, with $m = 8192$ and Top-K activation [33] ($K = 8$).

Diagnostic. We compute the coactivation correlation matrix $\mathbf{C}$ in (5) from the shared encoder, and obtain the Hungarian matching $\hat{\mathbf{P}}$ in (6). The two regimes in Section 4 make opposite predictions about $\hat{\mathbf{P}}$:

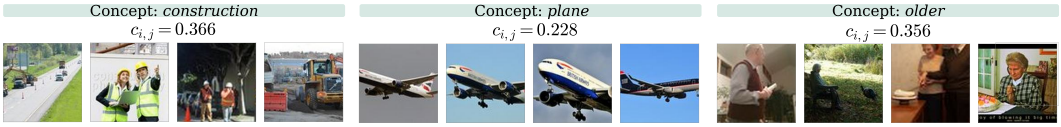


Figure 10: Examples of images that activates collapsed image–text feature pairs identified by a single shared SAE on `CLIP` embeddings of `MS-COCO` [29]. We define a coordinate is *collapsed* when $\hat{\mathbf{P}}(i) = i$ in the Hungarian matching of (6) applied to the correlation matrix $\mathbf{C}$ in (5), indicating that the same latent coordinate represents a concept in both modalities. Each row shows the top-activating images for one collapsed coordinate, along with its coactivation correlation $c_{i,j}$. We show the three collapsed pairs with the *lowest* $c_{i,j}$ rather than the highest. Even in this worst case, the coordinates correspond to coherent concepts (*plane*, *older*, *construction*), but $c_{i,j}$ remains low (0.23–0.37).

- Under modality split (Corollary 1), the same concept activates two *different* coordinates across modalities, so a Hungarian match aligns the coordinate with index i to a different coordinate $\hat{\mathbf{P}}(i) \neq i$.
- Under collapse (Theorem 1), the same concept activates the *same* coordinate across modalities, so the match is to itself, $\hat{\mathbf{P}}(i) = i$.

This leads to a natural diagnostic for collapse:

$$\text{collapse rate} := \frac{\left|\{\, i \in [\tilde{m}] : \hat{\mathbf{P}}(i) = i \,\}\right|}{\tilde{m}},$$

where $\tilde{m}$ is the number of live coordinates after excluding dead neurons.

Result. On the embeddings above, we observe a collapse rate of 4.64%. While small in absolute terms, this confirms that the collapse regime characterized by Theorem 1 does occur in real-world data. To verify that the diagonal matches reflect genuine semantic correspondences rather than coincidences, Figure 10 visualizes top-activating images for three collapsed coordinates. We deliberately select the *three collapsed pairs with the lowest* $c_{i,j}$, presenting a worst case rather than a favorable one. Even in this worst case, each coordinate corresponds to a coherent concept (*plane*, *older*, *construction*).

A further observation is that even at these collapsed coordinates, the coactivation correlation $c_{i,j}$ stays modest (0.23–0.37). This may seem counterintuitive, since collapsed pairs share a single column and one might expect strong coactivation. But it is consistent with the analysis in Proposition 1. When image and text features for the same concept are directionally distinct yet forced to share one column, the optimal shared direction is the compromise $\frac{\phi_i + \psi_i}{\|\phi_i + \psi_i\|}$ (see Propositions 4 and 3 for details), which fits neither modality well and produces only weak coactivation. In other words, collapse hurts not just reconstruction but cross-modal alignment as well.

F.3 Real-World Results Across Experimental Configurations

In the main experiment in Section 6.2, Table 1 reports results for a single configuration (`CLIP` ViT-B/32 embeddings with Top-K activation at $K = 32$). To verify that the conclusions drawn there are not artifacts of this particular setup, we conduct extended experiments along three orthogonal axes: (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 6.2, and we report averages over three independent runs. Tables 2, 3, and 4 report the per-configuration results. Across every configuration, our method matches the lowest reconstruction error of modality-specific SAEs while delivering the strongest cross-modal retrieval performance, mirroring the conclusions of Table 1.

Robustness to sparsity level K. We first vary the Top-K sparsity level on `CLIP` ViT-B/32 embeddings (Table 2). Together with the $K = 32$ result in Table 1, this covers $K \in \{16, 32, 64\}$. Across all three sparsity levels, our method preserves the lowest reconstruction error inherited from modality-specific SAEs. On cross-modal retrieval, $K = 32$ stands out as the best-performing sparsity level: our method peaks there with I→T Recall@1 of 16.0, a level not matched by any method at either $K = 16$ or $K = 64$. The retrieval gain over the strongest baseline is most pronounced in the sparser regime: at $K = 16$, our method more than doubles the performance of the strongest baseline

Table 2: Effect of Top-K sparsity level K on CLIP ViT-B/32 embeddings ($L = 8192$). Each section corresponds to a different value of K, and the $K = 32$ setting is reported in Table 1. Best and second-best results within each section are shown in bold and underlined, respectively.

	MS-COCO [29]							ImageNet1K [13]	
	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 pre-trained VLM backbone with $K = 32$ and Top-K activation. Each section corresponds to a different backbone, and the `CLIP` ViT-B/32 setting is reported in Table 1. Best and second-best results within each section are shown in bold and underlined, respectively.

	MS-COCO [29]							ImageNet1K [13]	
	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 ($L = 12288$)									
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 ($L = 8192$)									
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 ($L = 12288$)									
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)

on both retrieval directions. At $K = 64$, the auxiliary-loss baselines benefit from the larger active latent budget and the group-sparse loss becomes competitive on I→T Recall@1; however, retrieval scores of our method also drop relative to $K = 32$, so the narrowed gap reflects a regime where no method attains the peak achievable at $K = 32$, rather than one where the baselines genuinely surpass us. Even at $K = 64$, our method still leads on T→I Recall and for top-10 retrieval in both directions, while maintaining a substantially lower reconstruction error. Overall, $K = 32$ is the optimal operating point for cross-modal retrieval, and at this K our method generally outperforms all baselines.

Robustness across vision-language model backbones. We next evaluate three additional backbones at $K = 32$ with Top-K activation, as shown in Table 3. We span two architecture scales (ViT-B/32 vs. ViT-L/14) and two pre-training corpora (OpenAI `CLIP` [42] vs. `OpenCLIP` [7]). For each backbone, we scale the SAE width m with the embedding dimension d so that the expansion ratio m/d matches the setting of Table 1, ensuring a fair comparison across architectures of different sizes. On all three backbones, our method matches the lowest reconstruction error and obtains the strongest cross-modal retrieval performance, with the same qualitative behavior as the `CLIP` ViT-B/32 result in Table 1. The retrieval gain over the strongest baseline is largest on the ViT-B/32 backbones and smaller on ViT-L/14, reflecting that the L/14 embedding spaces already exhibit stronger cross-modal alignment and thus leave less room for improvement at Recall@1. The advantage on T→I Recall@1

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 [5] variant. Best and second-best results are shown in bold and underlined, respectively.

	MS-COCO [29]							ImageNet1K [13]	
	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)	<u>0.115</u>	16.4 (±3.1)
+ Iso-Energy alignment loss	<u>0.089</u>	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	<u>7.8</u> (±0.3)	<u>19.0</u> (±0.4)	<u>26.2</u> (±0.4)	<u>4.9</u> (±0.0)	<u>13.7</u> (±0.1)	<u>20.6</u> (±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**	<u>22.8</u> (±1.0)

is nevertheless consistently positive across all three backbones. As in the main experiment, the only metric where the group-sparse baseline occasionally surpasses our method is ImageNet zero-shot accuracy, but this comes at the cost of substantially degraded reconstruction.

Robustness to activation function. Finally, we replace Top-K with BatchTop-K [5], which selects the top-K activations across the entire batch rather than per token, on CLIP ViT-B/32 with $K = 32$, as shown in Table 4. Our method again attains the lowest reconstruction error (0.0878) and the strongest cross-modal retrieval, improving I→T Recall@1 from 7.81 to 16.19 and T→I Recall@1 from 4.88 to 10.46 over the strongest baseline. The magnitude of these gains closely matches the corresponding Top-K setting in Table 1, confirming that the benefits of post-hoc alignment are not specific to vanilla Top-K and transfer cleanly to alternative sparsifying activations.