DON'T MIND THE GAP: INTERPRETABLE VISION-LANGUAGE MODELS VIA INDEPENDENT SPARSE AUTOENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Contrastive Vision-Language Models (VLMs) produce general-purpose, multi-modal representations, but their black-box nature hinders safe and ethical deployment. Sparse Autoencoders (SAEs) can disentangle these dense representations into interpretable features, but the application of a single SAE across modalities creates a *split-dictionary* phenomenon, where learned features activate for a single modality, preventing the discovery of shared features. Current approaches attempt to bridge this gap using a co-activation loss, but we find that forcing a single shared SAE inherently compromises reconstruction quality, individual modality interpretability, and zero-shot performance. Instead, we propose decoupling the SAEs by modality while retaining the co-activation guidance. This dual architecture discovers features that retain their modality-specific nature while being *synchronized* to enable consistent cross-modal mapping. Our approach achieves superior interpretability and downstream zero-shot performance, especially in challenging high-to-medium sparsity regimes. Our findings indicate that exploiting the modality gap, rather than collapsing it, is the key to interpreting contrastive VLMs.

1 INTRODUCTION

Contrastive Vision-Language Models (VLMs) are multi-modal models that synchronize visual and textual data with similar semantics in a shared embedding space. Their general-purpose knowledge and cross-modal understanding have made them central to numerous applications, such as zero-shot classification and retrieval (Radford et al., 2021; Zhai et al., 2023; Tschannen et al., 2025), composed retrieval (Psomas et al., 2025), image generation (Ramesh et al., 2022; Blattmann et al., 2022), and text generation (Liu et al., 2023; Li et al., 2023). However, our understanding of their internal workings and how concepts are organized in their representational space remains limited, thus hindering our ability to improve model design (Lou et al., 2025), diagnose failure modes (Park et al., 2026), and ensure safe and ethical deployment (Sharkey et al., 2025).

Few recent works have addressed this lack of interpretability in contrastive VLMs by resorting to Sparse AutoEncoders (SAEs) to decompose dense representations as sparse linear combinations of an over-complete set of *dictionary vectors*, also called *dictionary words* or *atoms* (Joseph et al., 2025b; Papadimitriou et al., 2025; Dhimoïla et al., 2026; Kaushik et al., 2026). Intuitively, using bases composed of small subsets from these high-dimensional vectors allows to represent distinct concepts with minimal interference (see Fig. 1, left) (Gao et al., 2025; Cunningham et al., 2024; Bussmann et al., 2024; 2025). Yet, despite this intuition was proven correct with single-modality neural networks (Härle et al., 2025; Bhalla et al., 2026; Cywiński & Deja, 2025; Rao et al., 2024), the application of SAEs to contrastive VLMs faces the challenge of reconciling concepts across modalities. Somewhat surprisingly, it was observed that training a single SAE on VLM activations causes an effect called *split-dictionary* (Papadimitriou et al., 2025; Costa et al., 2025), where the latent atoms tend to be active for only one modality. Namely, even though textual and visual embeddings in VLMs are aligned by training, the SAEs' learned dictionary vectors may only support one modality. This failure in learning shared atoms limits the practical utility of the decomposition, because tasks like concept-based retrieval (Wen et al., 2025; Kaushik et al., 2026) or cross-modal steering (Dhimoïla et al., 2026) crucially hinge on the ability to identify a common concept among data of the different modalities.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

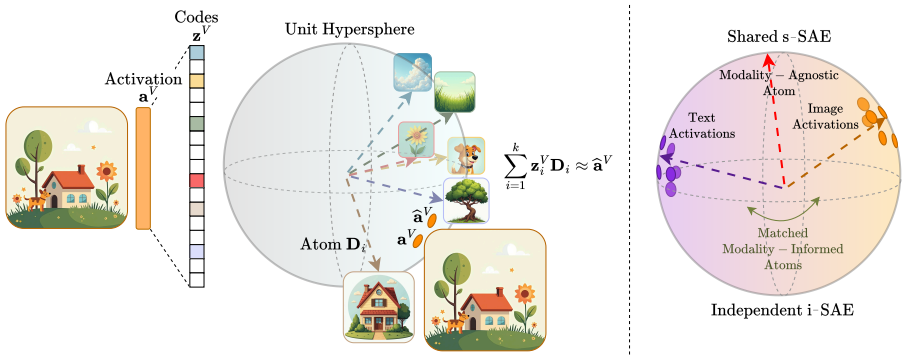


Figure 1: **SAE decomposition and atoms in the unit hypersphere.** On the left, the SAE learns to decompose the original activations into a sparse combination of learned dictionary atoms, $\mathbf{D}_i$, residing in the same hypersphere as the original activations. The weights of the sparse combination are given by the codes $\mathbf{z}_i$ (cartoon images generated with FLUX.1-dev (Labs, 2024)). On the right, we visualize the textual activations (purple circles) and the visual activations (orange circles), together with the learned dictionary atoms for modality-agnostic shared SAE (red arrow) and for modality-specific SAEs (dark purple and dark orange arrows). Each highlighted atom resides in a different subspace.

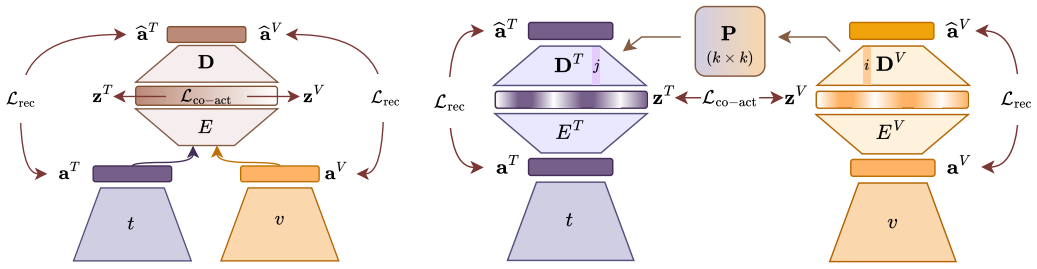
To mitigate the split-dictionary phenomenon, recent studies (Dhimoïla et al., 2026; Kaushik et al., 2026) introduce an additional co-activation loss that encourages dictionary words to be active across modalities. These *bimodal atoms* act as a conceptual bridge between modalities (Liang et al., 2022; Bangachev et al., 2025). Furthermore, Dhimoïla et al. (2026) show that the co-activation loss shapes the geometry of the learned dictionary to reflect the modality gap in the VLM: there is a small number of high-energy[1] *unimodal atoms* that are aligned with the modality-specific embeddings, while the vast majority of atoms are *bimodal* and approximately orthogonal to these modality-specific directions (see Fig. 1, right). Building on this insight, they show that removing the unimodal atoms can effectively narrow the gap and improve downstream performance by restricting vector arithmetic to words in the shared dictionary. However, this strategy comes at the cost of discarding a relevant portion of the information encoded in VLM representations. In particular, the geometric arrangement of bimodal atoms makes them poorly suited to accurately reconstruct embeddings corresponding to unimodal concepts, as evidenced by their inability to account for a large fraction of the activation magnitude when used for single-modality activation reconstruction (c.f. Fig. 3).

This limitation calls for rethinking how bimodality is defined. Rather than enforcing strict *alignment* through a single shared SAE, we propose to view bimodality through the lens of *synchronization* (Bangachev et al., 2025): modality-specific representations are preserved in their native modality subspaces while being linked through consistent cross-modal correspondences. This perspective captures the shared structure between modalities without conflating their distinct information, enabling interpretable decompositions that faithfully preserve the underlying nature of the VLM model.

We empirically show that independent, modality-specific SAEs (i-SAEs) offer several advantages over a shared SAE (s-SAE). Specifically, **(i)** i-SAEs achieve substantially better reconstruction quality, as expected from their ability to preserve modality-specific information rather than discard it through a shared dictionary. More importantly, **(ii)** this improvement in reconstruction does not come at the expense of interpretability or downstream utility: i-SAEs also yield more interpretable features and superior zero-shot downstream performance, particularly in the medium-to-high sparsity regime, where the trade-off between reconstruction fidelity and interpretability is typically most pronounced. Finally, **(iii)** we find that the latent dictionaries learned independently for each modality exhibit strong cross-modal alignment even without explicit alignment objectives. This correspondence can be further strengthened through a co-activation loss applied directly to the latent codes of the two SAEs, in contrast to Dhimoïla et al. (2026); Kaushik et al. (2026), where such a loss is applied within the shared latent space. This additional guidance becomes particularly beneficial under degraded data quality, improving interpretability on datasets such as YFCC15M-V2.

These findings indicate that exploiting the modality gap, rather than collapsing it, is key for interpreting contrastive VLMs.

[1]Energy is intended as the average activation of the atom.



(a) Shared (s-SAE). (b) Independent (i-SAE).

Figure 2: **SAE architectures.** (a) trains a single shared SAE for both input modalities, while (b) maintains separate SAEs for each modality. Since a certain concept may be represented by the j-th atom in one dictionary and by the i-th atom in the other, this requires finding a permutation matrix $\mathbf{P}$ that provides this correspondence. $\mathbf{P}$ is found with a single post-training step. Symmetric arrows indicate the arguments of the loss functions $\mathcal{L}_{\text{rec}}$ and $\mathcal{L}_{\text{co-act}}$.

2 Sparse Autoencoders in Contrastive VLM Spaces

2.1 Contrastive Vision-Language Model

A contrastive VLM (Radford et al., 2021; Zhai et al., 2023; Tschannen et al., 2025) uses a dual-encoder architecture and a contrastive objective to synchronize multi-modal data in a shared embedding space. Let $\mathcal{D} = \{(\mathbf{x}^V_{(i)}, \mathbf{x}^T_{(i)})\}_{i=1}^N$ denote a dataset of N paired samples, where $\mathbf{x}^V_{(i)} \in \mathbb{V}$ and $\mathbf{x}^T_{(i)} \in \mathbb{T}$ are the visual and textual inputs, respectively. The VLM consists of a visual encoder $v : \mathbb{V} \to \mathbb{R}^d$ and a text encoder $t : \mathbb{T} \to \mathbb{R}^d$. The two encoders project the inputs to a common d-dimensional space, yielding activations $\mathbf{a}^V_{(i)} \coloneqq v(\mathbf{x}^V_{(i)})$ and $\mathbf{a}^T_{(i)} \coloneqq t(\mathbf{x}^T_{(i)})$. The training objective encourages representations of paired inputs $(\mathbf{a}^V_{(i)}, \mathbf{a}^T_{(i)})$ to be more similar, according to the cosine similarity, than mismatched samples $(\mathbf{a}^V_{(i)}, \mathbf{a}^T_{(j)})$ for $j \neq i$. The activations of the N samples can be stacked to form the visual and textual activation matrices, $\mathbf{A}^V, \mathbf{A}^T \in \mathbb{R}^{N \times d}$, respectively.

2.2 Sparse Autoencoder

Given a generic activation matrix $\mathbf{A} \in \mathbb{R}^{N \times d}$, a SAE uses an encoder-decoder architecture to provide a differentiable solution to the *Dictionary Learning problem* Dumitrescu & Irofti (2018) and represent complex, high-dimensional data as a sparse linear combination of fundamental elements called *atoms*.

The encoder implements a map $E : \mathbb{R}^d \to \mathbb{R}^k$ that projects the activations in $\mathbf{A}$ to a set of sparse *codes* (or features) $\mathbf{Z} \in \mathbb{R}^{N \times k}$. Formally we have $\mathbf{Z} = \sigma\big(E(\mathbf{A})\big)$, where E is commonly realized as a single-layer MLP while $\sigma : \mathbb{R}^k \to \mathbb{R}^k_+$ denotes a nonlinearity that promotes sparsity and/or enforces non-negativity. In this work, we choose σ to be the TopK (Gao et al., 2025) function, which maps the pre-activation vector to the closest point in the set of K-sparse non-negative vectors:

$$\text{TopK}(\mathbf{z}, K) = \arg\min_{\mathbf{z}^\star \in \mathcal{S}^+_K} \|\mathbf{z}^\star - \mathbf{z}\|_2^2, \quad \text{where} \quad \mathcal{S}^+_K = \{\mathbf{u} \in \mathbb{R}^k_+ : \|\mathbf{u}\|_0 \leq K\}. \tag{1}$$

Namely, this operator retains only the highest K positive entries and zeroes out the rest.

The decoder is implemented via a learned dictionary $\mathbf{D} \in \mathbb{R}^{d \times k}$, with each of its columns $\mathbf{D}_j \; \forall j \in [k]$ defining an atom, or word. Using this dictionary, the activations are reconstructed as $\widehat{\mathbf{A}} = \mathbf{Z}\mathbf{D}^\intercal$. The SAE is trained by minimizing the reconstruction loss $\|\mathbf{A} - \widehat{\mathbf{A}}\|_F^2$, while also enforcing that all atoms have unit norm, *i.e.*, $\|\mathbf{D}_j\|_2 = 1 \; \forall j \in [k]$.

By setting $k \gg d$ and $K \ll k$, the SAE learns an over-complete dictionary while activating at most K atoms per sample. This encourages salient and disentangled features that are easier to interpret as meaningful concepts or attributes in the original activation space, as each feature is activated selectively and for consistent inputs.

2.3 Interpreting VLM Spaces using Shared SAEs (s-SAEs)

As contrastive VLMs are trained to synchronize the visual and textual modalities in a shared embedding space, a natural approach for learning cross-modal features is to train a single SAE on the combined activations of both modalities, as illustrated in Fig. 2a. This architecture uses the same encoder and dictionary parameters for both modalities, with the reconstruction loss defined as

$$\mathcal{L}_{\text{rec}} = \|\mathbf{A}^V - \widehat{\mathbf{A}}^V\|_F^2 + \|\mathbf{A}^T - \widehat{\mathbf{A}}^T\|_F^2 \,, \tag{2}$$

where $\widehat{\mathbf{A}}^V = \mathbf{Z}^V\mathbf{D}^\intercal$ and $\widehat{\mathbf{A}}^T = \mathbf{Z}^T\mathbf{D}^\intercal$ are the reconstructions of the visual and textual activations, respectively. Prior works (Papadimitriou et al., 2025; Costa et al., 2025) have shown that this training procedure often induces a split-dictionary effect, where the SAE learns to resort to modality-specific atoms to explain the activations. This result may be attributed to the modality gap typically found in contrastive VLMs, where the activations of the two modalities occupy partially distinct portions of the shared embedding space (Liang et al., 2022; Bangachev et al., 2025), and the SAE learns a partitioned dictionary (Papadimitriou et al., 2025; Dhimoïla et al., 2026) that reflects this geometric organization. This phenomenon hinders the ability of the SAE to learn features that are active across modalities and, as a consequence, it restricts the inspectability of the learned representations as it requires laborious manual effort to identify the modality-specific features representing the same concepts. Additionally, it prevents the usage of the SAE for tasks that require a unified cross-modal representation space, *e.g.*, filtering highly-aligned multi-modal data (Lou et al., 2025). Recent works (Dhimoïla et al., 2026; Kaushik et al., 2026) counter this effect by using an additional loss term $\mathcal{L}_{\text{co}-\text{act}}$ on $\mathbf{Z}^V$ and $\mathbf{Z}^T$ that encourages co-activation of the same features across modalities (loss expressions are reported in Sec. D). This strategy affects the SAE dictionary, promoting the emergence of bimodal atoms to effectively bridge the modality gap.

While restricting the dictionary to bimodal atoms was shown to improve performance on some downstream tasks w.r.t. using all atoms, this comparison does not establish that unimodal atoms are uniformly irrelevant for downstream behavior. Moreover, the effect of such an approximation on the quality of reconstructions has not been thoroughly explored, despite its direct implications for the faithfulness of the resulting conceptual decomposition to the original model.

Let us denote with $\mathbf{D}^B \in \mathbb{R}^{d\times k'}$ the sub-dictionary obtained from $\mathbf{D}$ by retaining only the k' bimodal atoms, as done in (Dhimoïla et al., 2026). By restricting the dictionary to these bimodal atoms, the reconstructed activations become $\widehat{\mathbf{A}}^B = \mathbf{Z}^B\mathbf{D}^{B\intercal}$, where $\mathbf{Z}^B \in \mathbb{R}^{N\times k'}$ denotes the components of the latent codes corresponding to the bimodal atoms. Intuitively, the fact that the dictionary vectors in $\mathbf{D}^B$ are approximately orthogonal w.r.t. the mean visual and textual embeddings Dhimoïla et al. (2026) implies that a base constituted solely by directions in $\mathbf{D}^B$ cannot support a precise reconstruction of those embeddings that do not reside in a shared (multi-modal) portion of the representational space. To quantify the magnitude of this reconstruction error, we evaluate the unexplained activation percentage for each modality when using only the bimodal sub-dictionary $\mathbf{D}^B$ as follows:

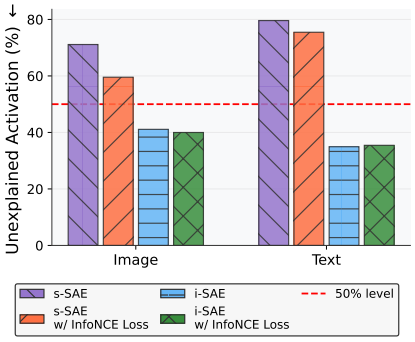


Figure 3: **Percentage of unexplained activation** by high-energy multi-modal atoms for CLIP ViT-L/14 ($K = 32$) on CC3M. Lower values are better.

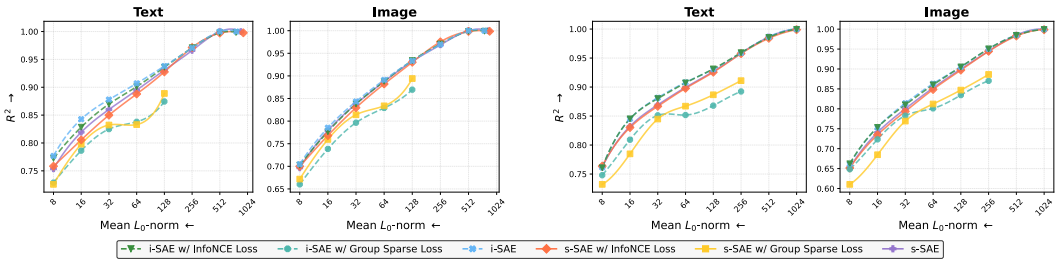
$$e_{\text{rec}} = \frac{\|\mathbf{A} - \mathbf{Z}^\star\mathbf{D}^{B\intercal}\|_F}{\|\mathbf{A}\|_F} \qquad \text{with} \quad \mathbf{Z}^\star = \arg\min_{\mathbf{W}\in\mathbb{R}_+^{N\times k'}} \|\mathbf{A} - \mathbf{W}\mathbf{D}^{B\intercal}\|_F, \tag{3}$$

where $\mathbf{Z}^\star$ is solved using projected gradient descent. More details about the computation of Eq. (3) are reported in Sec. F.1. With this setting, we find that a single SAE shared across modalities fails to describe more than 50% of the modality activation magnitudes when using only bimodal atoms (c.f. Fig. 3). Note that e_{rec} is a lower bound on the actual reconstruction error that would be obtained by using $\mathbf{Z}^B$. Thus, a single SAE cannot simultaneously maximize bimodal feature discovery and faithfully reconstruct the original VLM activations, revealing a structural limitation of shared cross-modal decomposition (Karvonen et al., 2025).

2.4 Leveraging the Modality Gap using Independent SAEs (i-SAEs)

To remedy the aforementioned limitation of the s-SAE paradigm, we propose using two distinct SAEs, one for each modality, as illustrated in Fig. 2b. Each SAE is trained independently to minimize the visual or textual term of $\mathcal{L}_{\text{rec}}$ (see Eq. (2)), yielding modality-specific dictionaries $\mathbf{D}^V$, $\mathbf{D}^T$. This separation allows each dictionary to be composed of atoms that remain in their respective modality-specific subspace, thereby improving reconstruction fidelity. We denote the corresponding encoder mappings by $E^V(\cdot)$ and $E^T(\cdot)$, and the resulting latent codes by $\mathbf{z}^V$ and $\mathbf{z}^T$.

Together with the reconstruction objective, we can add a co-activation loss $\mathcal{L}_{\text{co}-\text{act}}$ to encourage modality-specific features to be active simultaneously ($\mathbf{z}^T \sim \mathbf{z}^V$). This helps to regularize the model



(a) CLIP ViT-B/32 (b) SigLIP ViT-B/16

Figure 4: **Reconstruction performance** on CC3M. Each plot shows the R^2 score of the reconstructed activations against the original activations for both modalities across different sparsity levels. The x-axis represents the mean L_0-norm of the codes (log scale).

and promotes code alignment between the two modalities, but does not imply that a given code element refers to the same concept across modalities. Indeed, a concept may be mapped to the j-th atom of $\mathbf{D}^T$ and to the i-th atom of $\mathbf{D}^V$ (see Fig. 2b). Since the dictionary atoms lie in the joint output space of a contrastive VLM, their cosine similarities provide a natural measure of cross-modal semantic correspondence. We can therefore map atoms across modalities by finding the permutation matrix that maximizes the total pairwise similarity:

$$\mathbf{P} = \arg\max_{\mathbf{P}\in\mathcal{P}(k)} \frac{1}{k} \operatorname{Tr}\left((\mathbf{D}^V\mathbf{P})^\intercal \mathbf{D}^T\right), \tag{4}$$

where $\mathcal{P}(k)$ denotes the set of $k \times k$ permutation matrices. Hungarian Matching (Kuhn, 1955) yields the globally optimal permutation matrix for Eq. (4). As shown in Fig. 1 (right), this approach enables independently trained SAEs to acquire *synchronized* cross-modal semantics.

In the next section, we systematically compare s-SAE and i-SAE considering interpretability metrics, geometric structure of the learned decompositions, and zero-shot downstream performance using different instantiations of $\mathcal{L}_{\text{co-act}}$.

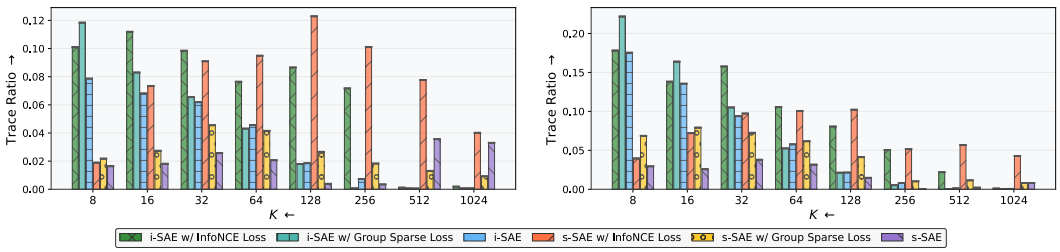
3 Experiments

Implementation Details. We train SAEs on the output activations of four contrastive VLMs: CLIP ViT-B/32 and CLIP ViT-L/14 (Dosovitskiy et al., 2021; Radford et al., 2021), and SigLIP ViT-B/16 (Zhai et al., 2023) and SigLIP 2 ViT-B/16 (Tschannen et al., 2025), as representative of the softmax-based and sigmoid-based contrastive learning paradigms, respectively. We experiment with two datasets: CC3M (Sharma et al., 2018) and YFCC15M-V2 (Li et al., 2022; Cui et al., 2022). CC3M is a highly curated dataset consisting of approximately (originally) 3.3 million image-text pairs, whose data has undergone a rigorous filtering process. In contrast, YFCC15M-V2 is a 15 million image-text subset of YFCC100M (Thomee et al., 2016), constituting a mid-scale dataset used for assessing the scalability of SAEs training as data-quality degrades and dataset size increases. We vary the sparsity level of the SAEs by setting $K \in \{2^x : x = 3, 4, \ldots, 10\}$ for the TopK activation function and we set the $\mathcal{L}_{\text{co-act}}$ to be the group-sparse loss (Kaushik et al., 2026), the InfoNCE loss (Oord et al., 2018; Wen et al., 2025), or no co-activation loss at all. We set $k = 16 \times d$ for all trained SAEs (see Sec. F.2.1 for additional experiments varying the dictionary size). Using the same dictionary size ensures that all SAEs have the same number of features per modality, as well as the same potential upper bound for monosemantic bimodal atoms. See Appendix D for further details on the implementation and evaluation protocols.

3.1 Quantitative Results

Reconstruction Performance. To evaluate the faithfulness of SAEs to the original VLM (Karvonen et al., 2025), we use a held-out subset from the dataset on which SAEs were trained to measure the R^2 score (see Eq. (10)) between the original and the reconstructed activations per-modality, as well as the average L_0-norm of the codes, which is constrained to be at most K by the TopK function.

Although the quality of the reconstructions is dependent on both the VLM and the modality, we find that variants trained on both CC3M and YFCC15M-V2 exhibit a clear trend of better reconstruction performance for i-SAEs, as shown in Figs. 4, 11 and 24. For sparsity levels with $K \leq 128$, the performance gap between i-SAEs and s-SAEs is generally wider, while performances converge as



(a) CLIP ViT-B/32 (b) SigLIP ViT-B/16

Figure 5: **Trace Ratio comparison** on CC3M. Metric values, represented as bars, are grouped by the sparsity level K in TopK activation function.

K increases. Notably, SAEs trained with the group-sparse loss perform significantly worse than the other variants, as the loss causes code magnitudes to shrink, producing activations that conflict with the reconstruction objective.

Cross-Modal Mapping Purity. To investigate the cross-modal mapping purity of the learned atoms across modalities, we evaluate the trace ratio of the Bridge Matrix $\mathbf{B} \in \mathbb{R}^{k \times k}$ (Papadimitriou et al., 2025) defined as:

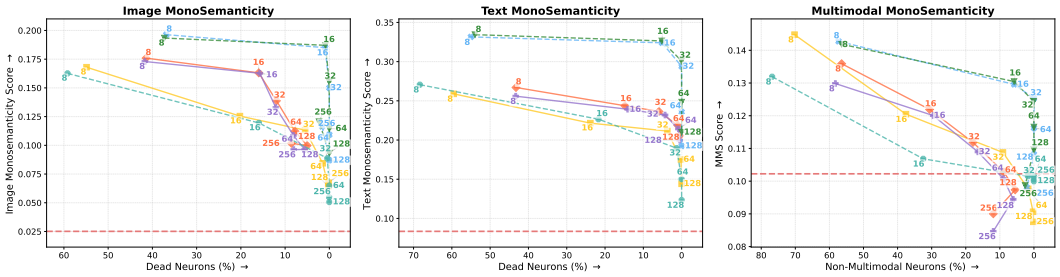
$$\text{Trace Ratio} = \frac{\sum_{i=1}^{k} \mathbf{B}_{ii}}{\sum_{i,j=1}^{k} |\mathbf{B}_{ij}|}, \quad \text{where} \quad \mathbf{B} = \underbrace{\mathbb{E}_{(\mathbf{z}^V, \mathbf{z}^T)} \left[\mathbf{P}^\intercal \mathbf{z}^V (\mathbf{z}^T)^\intercal \right]}_{\text{co-activation}} \odot \underbrace{\left((\mathbf{D}^V \mathbf{P})^\intercal \mathbf{D}^T \right)}_{\text{alignment}} \tag{5}$$

where $\odot$ denotes the Hadamard product. The first term in $\mathbf{B}$ captures the statistical co-activation between concepts: how frequently concept i in the image code and j in the text code are simultaneously active for semantically aligned inputs. The second term reflects the geometric alignment between corresponding dictionary atoms computed as their cosine similarity. The trace ratio provides the fraction of the total cross-modal interaction energy that is successfully concentrated onto uniquely paired features. A higher value is desirable, since it indicates that cross-modal capabilities emerge from a clean, one-to-one mapping of features across modalities. In the case of s-SAE, the permutation matrix $\mathbf{P}$ is the identity matrix, since the features are already aligned ($\mathbf{D}^V = \mathbf{D}^T$).

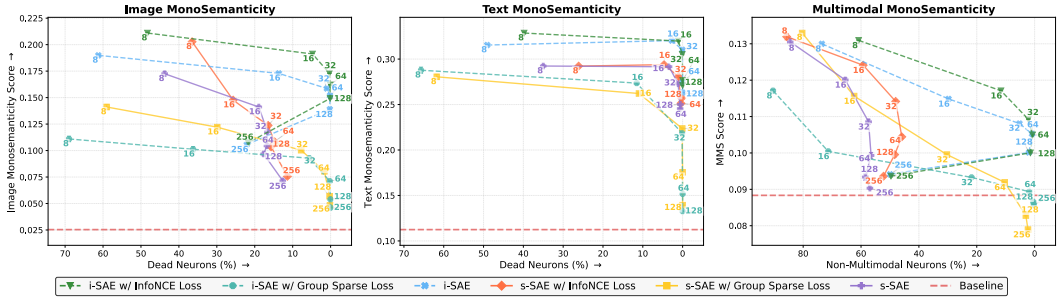
As shown in Figs. 5, 12 and 25, across both CLIP and SigLIP models, i-SAEs dominate the high-to-mid sparsity regime ($K \leq 64$), while s-SAEs show better cross-modal mapping at lower sparsity levels ($K \geq 128$). The high-sparsity regime is particularly important for interpretability, since it is where the most disentangled and monosemantic features typically emerge (Pach et al., 2025; Karvonen et al., 2025), making individual concepts easier to identify and analyze in isolation.

Monosemanticity Scores. To evaluate the interpretability of the learned features, we employ the same held-out subset of the reconstruction evaluation. We use the same evaluation protocol as (Pach et al., 2025), where a MonoSemanticity score (MS) is computed for each feature $j \in [k]$ as the weighted average of pairwise cosine similarities between the activating samples of that feature, using co-activation code magnitudes of feature j for weighting. We use DINOv2 (Oquab et al., 2024) for computing cosine similarities for the visual modality, and Sentence BERT embedding similarity (Reimers & Gurevych, 2019) through the "all-MiniLM-L6-v2" model for the textual modality (Minder et al., 2025). We further inspect the multi-modality semantics of the features using the Multimodal MonoSemanticity score (MMS) (Kaushik et al., 2026), which uses a different weighting scheme than (Pach et al., 2025) and computes the cosine similarities using samples from both modalities. This metric specifically assesses whether features across modalities correspond to the same monosemantic concept. As in the original work, we use CLIP RN-50 (Radford et al., 2021) for encoding samples.

The distribution of Image MS, Text MS, and MMS scores over different sparsity regimes are reported in Figs. 6, 13 and 26. We find that i-SAEs consistently outperform or match s-SAEs for Image and Text MS scores, while also generally reducing the percentage of dead neurons for the visual domain. Importantly, the results on YFCC15M-V2 show that the additional co-activation loss plays a critical role in enhancing the performance of i-SAEs, with the feature synchronization leading to fewer dead neurons and a higher average monosemantic score across all the sparsity spectrum. This may be attributed to mitigation of uninterpretable noise due to cooperation among modalities. For the MMS scores, the performance between the two architectures is comparable in the mid-to-low sparsity



(a) CC3M - CLIP ViT-L/14



(b) YFCC15M-V2 - SigLIP ViT-B/16

Figure 6: **Interpretability evaluation**. Each point corresponds to the mean score across all features for a given sparsity level, highlighted in correspondence of each point. Points closer to the top-right corner are better. The x-axis indicates the percentage of dead neurons, *i.e.*, neurons that are never activated across the evaluation set. For MMS scores, we indicate the percentage of non-multimodal neurons, *i.e.*, neurons that are never activated across both modalities in the evaluation set. The dashed horizontal line indicates the mean score across all features for the dense VLM model.

regime ($K \geq 32$), while i-SAEs contribute to a consistently higher percentage of multimodal neurons. Remarkably, while the group-sparse loss is an effective solution for extracting cross-modal features in s-SAE, it significantly reduces monosemanticity scores for i-SAEs. The distribution of code magnitudes tends towards zero for all examples of both modalities, which effectively discourages features from focusing on a single, clear concept.

Downstream Performance. While SAEs notoriously increase the interpretability of representations, this often comes at the cost of reduced downstream performance. We focus on zero-shot classification and retrieval to evaluate SAE variants using their reconstructions. This standard evaluation protocol additionally allows us to assess whether the features useful for reconstruction are also functionally relevant for explaining downstream behavior beyond the training distribution (Braun et al., 2024). By expanding the cosine similarity for zero-shot predictions for samples $\widehat{\mathbf{a}}^V$ and $\widehat{\mathbf{a}}^T$ we get:

$$s(\widehat{\mathbf{a}}^V, \widehat{\mathbf{a}}^T) = \frac{\sum_{i,j=1}^{k} \overbrace{\mathbf{z}_i^V \mathbf{z}_j^T}^{\text{co-activation}} \overbrace{\langle \mathbf{D}_i^V, \mathbf{D}_j^T \rangle}^{\text{alignment}}}{\left(\sum_{i=1}^{k} (\mathbf{z}_i^V)^2 + \underbrace{\sum_{\substack{i,j=1 \\ i \neq j}}^{k} \mathbf{z}_i^V \mathbf{z}_j^V \langle \mathbf{D}_i^V, \mathbf{D}_j^V \rangle}_{\text{intra-modal interference}} \right)^{\frac{1}{2}} \left(\sum_{i=1}^{k} (\mathbf{z}_i^T)^2 + \underbrace{\sum_{\substack{i,j=1 \\ i \neq j}}^{k} \mathbf{z}_i^T \mathbf{z}_j^T \langle \mathbf{D}_i^T, \mathbf{D}_j^T \rangle}_{\text{intra-modal interference}} \right)^{\frac{1}{2}}}. \quad (6)$$

The SAE-decomposed similarity balances cross-modal alignment (numerator) against intra-modal interference (denominator). Effective zero-shot performance with sparse representations requires dictionaries that minimize feature interference (Elhage et al., 2022), ensuring that similarity is driven by semantic cross-modal feature matches. In the case of s-SAE, $\mathbf{D}_i^V = \mathbf{D}_i^T, \forall i \in [k]$.

The results for CLIP ViT-L/14 on CC3M are reported in Fig. 7, while the remaining results are in Secs. F.2 and F.3. We find that the performance gap between the two architectures narrows as the sparsity level decreases (*i.e.*, as K increases). Importantly, i-SAEs substantially and consistently outperform s-SAEs in the very-high-to-mid sparsity regimes ($K \leq 64$), where the conceptual decomposition becomes increasingly intelligible. The impact of missing semantic information is

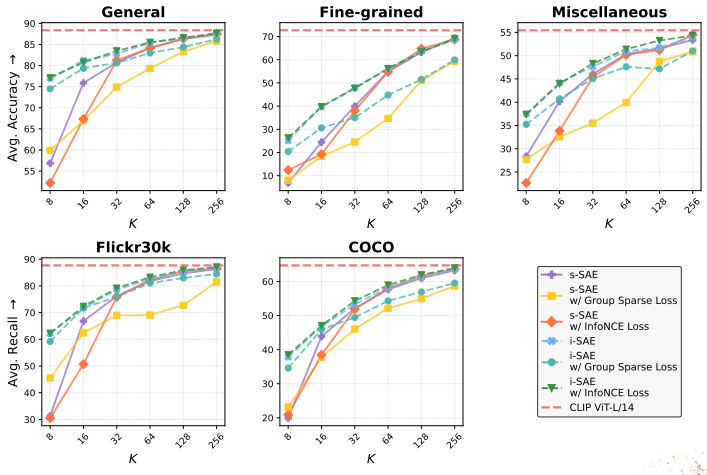
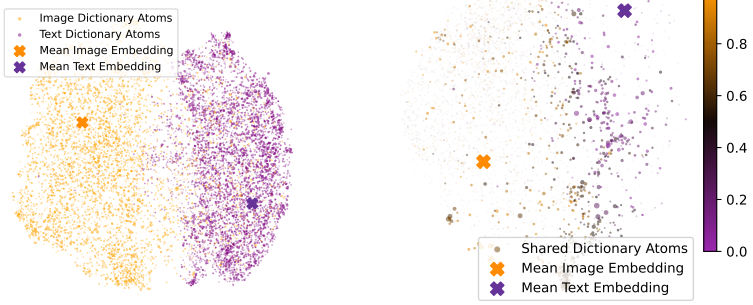


Figure 7: **Zero-shot classification accuracy** (%) (top) **and retrieval** (bottom) across datasets, grouped by sparsity level K, for CLIP ViT-L/14 on CC3M. The classification results are averaged across datasets grouped in General, Fine-grained, Miscellaneous, while the retrieval results are the average of $R@\{1, 5, 10\}$ for text and image retrieval in the specified dataset (see Tabs. 5 and 6 for full results).



(a) i-SAE w/ InfoNCE Loss (b) s-SAE w/ Group-Sparse Loss

Figure 8: **UMAP visualization of learned dictionary atoms** with modality scores in $[0.05, 0.95]$ for SAEs with $K = 8$ for CLIP ViT-L/14 on CC3M. Orange and purple points represent image and text dictionary atoms, respectively, while dark crosses represent the mean modality-specific embeddings. For the shared SAE in (b), the point color represents the modality score. Each point size is proportional to the energy score (Papadimitriou et al., 2025) of the corresponding atom.

greater for retrieval than for classification, as reflected by the more pronounced gains of i-SAE over s-SAE on the former task.

Taken together with the reconstruction results, our downstream evaluations indicate that reconstruction quality and downstream performance capture complementary aspects of the learned representations. Models trained with the Group Sparse Loss consistently underperform in reconstruction quality and interpretability, yet largely recover the downstream performance gap, especially for i-SAEs, whose atoms remain in their modality-specific subspaces. This shows that reconstruction-favorable features are not necessarily those most relevant for downstream behavior, while preserving modality-specific structure can remain beneficial even with lower reconstruction fidelity. More broadly, the consistent downstream gains of i-SAEs suggest that their advantage extends beyond reconstruction, reflecting their ability to retain and expose a richer set of semantically meaningful concepts, consistent with the monosemanticity results. Importantly, this advantage persists even when controlling for the number of atoms, as discussed in the following paragraph.

Comparison with Equal Number of Atoms. All the experiments so far use the same number k of features to encode each modality, both in the s-SAE and in the i-SAEs. Although this effectively means that the i-SAE architecture has double the budget of atoms, all SAEs are theoretically allowed to encode each sample using the same number of latent features. Nevertheless, we experimentally verify that the performance gap still remains even when halving the number of features for each modality for i-SAEs w.r.t. s-SAEs (see Sec. F.2.2). This suggests that the advantage of the i-SAE variants lies in the modality-specific nature of their dictionaries.

Cross-Modal Steering. We finally perform cross-modal steering using only multi-modal features on the FashionIQ (Wu et al., 2021) dataset as in Dhimoïla et al. (2026). The goal is to compose a candidate image and a relative caption by transferring only multi-modal information from one modality to another to retrieve the correct target

from a pool of images (see Sec. F.2.3 for full details). Table 1 show that i-SAEs consistently outperform s-SAEs.

Table 1: **Cross-modal steering** towards text distribution, SAEs trained on CC3M with CLIP ViT-L/14 ($K = 32$) on FashionIQ.

SAE	$0.5(R@10 + R@50)$	
	$\hat{\mathbf{a}}_{\text{original}}$	$\hat{\mathbf{a}}_{\text{steered}}$
s-SAE	69.34	68.78
w/ Group Sparse Loss	69.12	68.06
w/ InfoNCE Loss	69.05	71.81
i-SAE	69.47	**72.72**
w/ Group Sparse Loss	67.81	**72.94**
w/ InfoNCE Loss	69.89	**73.96**

3.2 Qualitative Analysis of Learned Features

Atoms living in modality-specific subspaces. To confirm that cross-modal features using i-SAEs emerge from atoms residing in the modality subspaces, we show in Fig. 8 the UMAP (McInnes et al., 2018) visualization of atoms with modality scores (Papadimitriou et al., 2025; Dhimoïla et al., 2026) in $[0.05, 0.95]$, which (Dhimoïla et al., 2026) uses to identify multi-modal features (see Sec. D).

As expected, while the s-SAE with the group-sparse loss discovers higher-energy cross-modal features mainly in the intermediate region between the two modality subspaces, the i-SAEs with the InfoNCE loss allow those features to live inside the modality subspaces. This validates that cross-modal understanding is achieved using synchronized features (Bangachev et al., 2025).

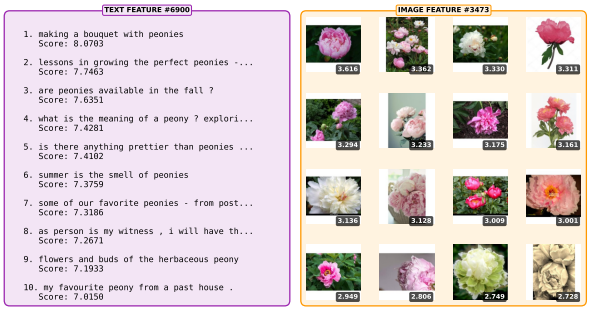


Figure 9: **Top-activating examples** for a cross-modal feature from i-SAE ($K = 32$) with InfoNCE Loss for SigLIP 2 ViT-B/16 on CC3M. The feature describes the concept of "peonies", present in the examples for both modalities. For each sample, we report the activation value for the feature.

Modality-specific features activating for the same concept. Figure 9 shows the top-activating examples for a cross-modal feature from i-SAEs with the InfoNCE loss. By analyzing these examples for both modalities, we can see that they activate for the same underlying concept of "peonies", highlighting that cross-modal semantics and interpretability can indeed be achieved while remaining in the modality subspaces. See more examples in Sec. F.2 and Sec. F.3.

4 Discussion and Findings

This work investigates the capabilities and limitations of a SAE shared across modalities (s-SAE) to interpret features in contrastive VLM spaces, and it introduces a new paradigm based on modality-specific SAEs (i-SAEs). Our experiments provide valuable insights which shed a light on:

Reconstruction: A s-SAE architecture cannot faithfully reconstruct contrastive VLM activations, due to the representational collapse of bimodal concepts into the same atoms. The reconstruction gap between shared and independent SAEs tends to decrease as the sparsity constraint gets relaxed, as the code features becomes less monosemantic and interpretable.

Monosemanticity and Dead Neurons: i-SAEs contribute to decrease the number of dead neurons and increasing the number of multimodal features, accompanied by a general increase in the interpretability of the learned features, especially in a high-to-medium sparsity regime.

Semantic correspondence across modalities: As the sparsity increases, features are better mapped across modalities by using i-SAEs, which suggests that monosemantic features requires atoms living in different subspaces for effective interpretability of shared semantics across modalities.

Downstream performance: i-SAEs exhibit superior zero-shot downstream performance compared to s-SAEs, with the performance gap widening as the sparsity constraint gets stronger, when the trade-off between interpretability and downstream performance is more pronounced.

Performance gap comes from modality-specific dictionaries: The performance gap still remains even when halving the number of features for each modality for i-SAEs w.r.t. s-SAEs (see Sec. F.2.2).

Co-activation loss effect: In s-SAEs, the co-activation loss is indispensable for learning bimodal features, while i-SAEs can naturally learn shared semantics without it; however, it helps in increasing the interpretability of i-SAEs' features when the training data quality degrades.

All the above insights indicate that exploiting the modality gap naturally present in contrastive VLMs is key to interpreting their representations without sacrificing reconstruction accuracy and downstream performance.

AI USE STATEMENT

In this work, generative AI tools were used to implement some of the experiments, after the authors had coded the methods and metrics. Each AI-assisted contribution was carefully reviewed and verified by the authors to ensure the experiments were correctly implemented, with the outputs refined and validated in a human-in-the-loop manner. AI assistance was sparingly employed to double-check some results to further ensure their correctness. Finally, AI assistance was used to create templates for tables and figures and to manage the space within the manuscript effectively. We take full responsibility for the final content of this work, including the formatting, text and code produced with the aid of generative AI.

ETHICS STATEMENT

Our study builds on existing works on Mechanistic Interpretability, *i.e.*, the interpretability field of understanding the inner workings of deep learning models by analyzing their components and their interactions. As large-scale models are increasingly deployed in high-stakes domains like healthcare and law, mitigating opaque "black-box" risks is crucial for safe deployment. By using Sparse Autoencoders, practitioners can examine internal representations to proactively identify and resolve model issues. While interpretability techniques carry inherent risks, such as misinterpreting simplified explanations or enabling malicious output modifications like targeted hate towards certain groups of people or communities, our approach neither introduces novel misuse vectors nor exacerbates existing risks. Furthermore, in our human study (see Sec. E), performed within our institution, we did not collect any user information beyond a unique identifier to distinguish participants.

REPRODUCIBILITY STATEMENT

We have made significant efforts to ensure the reproducibility of our results. Implementation details, datasets, and evaluation metrics are presented in the main text (Sec. 3), while an extended description of the training procedures, hyperparameters, and additional details on datasets and metrics are provided in the Sec. D in the appendix. Upon acceptance, we will publicly release the codebase to reproduce our experimental results.

REFERENCES

Kumar Ashutosh, Yossi Gandelsman, Xinlei Chen, Ishan Misra, and Rohit Girdhar. LLMs can see and hear without any training. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=cJeFULIiot.

Kiril Bangachev, Guy Bresler, Iliyas Noman, and Yury Polyanskiy. Global minimizers of sigmoid contrastive loss. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=IeM6Io4Rsh.

Dor Bank, Noam Koenigstein, and Raja Giryes. Autoencoders. *Machine learning for data science handbook: data mining and knowledge discovery handbook*, pp. 353–374, 2023.

Usha Bhalla, Alex Oesterling, Suraj Srinivas, Flavio Calmon, and Himabindu Lakkaraju. Interpreting CLIP with sparse linear concept embeddings (spliCE). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=7UyBKTFrtd.

Usha Bhalla, Alex Oesterling, Claudio Mayrink Verdun, Himabindu Lakkaraju, and Flavio Calmon. Temporal sparse autoencoders: Leveraging the sequential nature of language for interpretability. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=bojVI4l9Kn.

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models. https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html, 2023.

Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. Semi-parametric neural image synthesis. *arXiv preprint arXiv:2204.11824*, 2022.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pp. 446–461. Springer, 2014.

Dan Braun, Jordan Taylor, Nicholas Goldowsky-Dill, and Lee Sharkey. Identifying functionally important features with end-to-end sparse dictionary learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=7txPaUpUnc.

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

Bart Bussmann, Patrick Leask, and Neel Nanda. Batchtopk sparse autoencoders. In *NeurIPS 2024 Workshop on Scientific Methods for Understanding Deep Learning*, 2024. URL https://openreview.net/forum?id=d4dpOCqybL.

Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning multi-level features with matryoshka sparse autoencoders. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=m25T5rAy43.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

Valérie Costa, Thomas Fel, Ekdeep Singh Lubana, Bahareh Tolooshams, and Demba E. Ba. From flat to hierarchical: Extracting sparse representations with matching pursuit. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=Ll5miDx8KB.

Yufeng Cui, Lichen Zhao, Feng Liang, Yangguang Li, and Jing Shao. Democratizing contrastive language-image pre-training: A CLIP benchmark of data, model, and supervision. In *First Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022*, 2022. URL https://openreview.net/forum?id=wAEizNH9jJ6.

Hoagy Cunningham, Aidan Ewart, Logan Riggs Smith, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=F76bwRSLeK.

Bartosz Cywiński and Kamil Deja. SAeUron: Interpretable concept unlearning in diffusion models with sparse autoencoders. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=6N0GxaKdX9.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. In *International Conference on Machine Learning*, pp. 7694–7731. PMLR, 2023.

Grégoire Dhimoïla, Thomas Fel, Victor Boutin, and Agustin Martin Picard. Cross-modal redundancy and the geometry of vision–language embeddings. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=VYQuICALXj.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Bogdan Dumitrescu and Paul Irofti. *Dictionary Learning Algorithms and Applications*. Springer Publishing Company, Incorporated, 1st edition, 2018. ISBN 3319786733.

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders find interpretable LLM feature circuits. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=J6zHcScAo0.

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 arXiv:2209.10652*, 2022.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pp. 178–178. IEEE, 2004.

Thomas Fel, Ekdeep Singh Lubana, Jacob S. Prince, Matthew Kowal, Victor Boutin, Isabel Papadimitriou, Binxu Wang, Martin Wattenberg, Demba E. Ba, and Talia Konkle. Archetypal SAE: Adaptive and stable dictionary learning for concept extraction in large vision models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=9v1eW8HgMU.

Yossi Gandelsman, Alexei A Efros, and Jacob Steinhardt. Interpreting CLIP's image representation via text-based decomposition. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=5Ca9sSzuDp.

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 *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=tcsZt9ZNKD.

Leander Girrbach and Zeynep Akata. Sparse autoencoders are topic models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=gXXrubolmR.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Ruben Härle, Felix Friedrich, Manuel Brack, Björn Deiseroth, Stephan Waeldchen, Patrick Schramowski, and Kristian Kersting. Measuring and guiding monosemanticity. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=REHjkmWdQL.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

Sonia Joseph, Praneet Suresh, Ethan Goldfarb, Lorenz Hufe, Yossi Gandelsman, Robert Graham, Danilo Bzdok, Wojciech Samek, and Blake Aaron Richards. Steering CLIP's vision transformer with sparse autoencoders. In *Mechanistic Interpretability for Vision at CVPR 2025 (Non-proceedings Track)*, 2025a. URL https://openreview.net/forum?id=pKbhPioHKe.

Sonia Joseph, Praneet Suresh, Lorenz Hufe, Edward Stevinson, Robert Graham, Yash Vadi, Danilo Bzdok, Sebastian Lapuschkin, Lee Sharkey, and Blake Aaron Richards. Prisma: An open source toolkit for mechanistic interpretability in vision and video. *arXiv preprint arXiv:2504.19475*, 2025b.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3128–3137, 2015.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Isaac Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum Stuart McDougall, Kola Ayonrinde, Demian Till, Matthew Wearden, Arthur Conmy, Samuel Marks, and Neel Nanda. SAEBench: A comprehensive benchmark for sparse autoencoders in language model interpretability. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=qrU3yNfX0d.

Chiraag Kaushik, Davis Barch, and Andrea Fanelli. Learning multimodal dictionary decompositions with group-sparse autoencoders. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ZJlVXZ5dmK.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561, 2013.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

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

Black Forest Labs. Announcing black forest labs. *Black Forest Labs Blog*, 2024. URL https://bfl.ai/announcing-black-forest-labs/.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=zq1iJkNk3uN.

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 Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=S7Evzt9uit3.

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*, pp. 740–755. Springer, 2014.

Zihao Lin, Samyadeep Basu, Mohammad Beigi, Varun Manjunatha, Ryan A Rossi, Zichao Wang, Yufan Zhou, Sriram Balasubramanian, Arman Zarei, Keivan Rezaei, et al. A survey on mechanistic interpretability for multi-modal foundation models. *arXiv preprint arXiv:2502.17516*, 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=w0H2xGHlkw.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Skq89Scxx.

Hantao Lou, Changye Li, Jiaming Ji, and Yaodong Yang. SAE-v: Interpreting multimodal models for enhanced alignment. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=S4HPn5Bo6k.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Julian Minder, Clément Dumas, Caden Juang, Bilal Chughtai, and Neel Nanda. Overcoming sparsity artifacts in crosscoders to interpret chat-tuning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=yFdNygEryH.

Andrew Ng et al. Sparse autoencoder. *CS294A Lecture notes*, 72(2011):1–19, 2011.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pp. 722–729. IEEE, 2008.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

Mateusz Pach, Shyamgopal Karthik, Quentin Bouniot, Serge Belongie, and Zeynep Akata. Sparse autoencoders learn monosemantic features in vision-language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=DaNnkQJSQf.

Avik Pal, Max van Spengler, Guido Maria D'Amely di Melendugno, Alessandro Flaborea, Fabio Galasso, and Pascal Mettes. Compositional entailment learning for hyperbolic vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=3i13Gev2hV.

Isabel Papadimitriou, Huangyuan Su, Thomas Fel, Sham M. Kakade, and Stephanie Gil. Interpreting the linear structure of vision-language model embedding spaces. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=qPsmGjpq1j.

Sangha Park, Seungryong Yoo, Jisoo Mok, and Sungroh Yoon. Save: Sparse autoencoder-driven visual information enhancement for mitigating object hallucination. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7935–7944, March 2026.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Gonçalo Santos Paulo, Alex Troy Mallen, Caden Juang, and Nora Belrose. Automatically interpreting millions of features in large language models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=EemtbhJOXc.

Kenny Peng, Rajiv Movva, Jon Kleinberg, Emma Pierson, and Nikhil Garg. Position: Use sparse autoencoders to discover unknowns. In *Forty-third International Conference on Machine Learning Position Paper Track*, 2026. URL https://openreview.net/forum?id=x1Px74tbvs.

Bill Psomas, George Retsinas, Nikos Efthymiadis, Panagiotis Filntisis, Yannis Avrithis, Petros Maragos, Ondrej Chum, and Giorgos Tolias. Instance-level composed image retrieval. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders. *arXiv preprint arXiv:2407.14435*, 2024.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

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*, pp. 444–461. Springer, 2024.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeffrey Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Isaac Bloom, Stella Biderman, Adrià Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Mary Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, William Saunders, Eric J Michaud, Stephen Casper, Max Tegmark, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Thomas McGrath. Open problems in mechanistic interpretability. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=91H76m9Z94. Survey Certification.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, 2018.

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

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 arXiv:2502.14786*, 2025.

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Tiansheng Wen, Yifei Wang, Zequn Zeng, Zhong Peng, Yudi Su, Xinyang Liu, Bo Chen, Hongwei Liu, Stefanie Jegelka, and Chenyu You. Beyond matryoshka: Revisiting sparse coding for adaptive representation. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=z19u9B2fCZ.

Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11307–11317, June 2021.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D Manning, and Christopher Potts. Axbench: Steering LLMs? even simple baselines outperform sparse autoencoders. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=K2CckZjNy0.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics*, 2:67–78, 2014.

Vladimir Zaigrajew, Hubert Baniecki, and Przemyslaw Biecek. Interpreting CLIP with hierarchical sparse autoencoders. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=5MQQsenQBm.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

Kaichen Zhang, Yifei Shen, Bo Li, and Ziwei Liu. Large multi-modal models can interpret features in large multi-modal models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3650–3661, 2025.

APPENDIX

A Related Works **17**

B Limitations and Future Works **17**

C Notation **18**

D Implementation Details **18**

E Human Study **20**

F Additional Results **20**

F.1 Unexplained activation percentage using multi-modal atoms 22

F.2 CC3M 22

F.2.1 Varying the dictionary size 35

F.2.2 Halving the number of features for each modality in the independent SAE variants 35

F.2.3 Cross-modal steering 38

F.3 YFCC15M-V2 40

G Autointerpretability Scores **48**

A Related Works

Sparse Autoencoders (SAEs). Neural networks are hypothesized to operate in a superposition regimen (Elhage et al., 2022), encoding more features than the number of neurons they have. SAEs (Ng et al., 2011; Bricken et al., 2023) represent an approach to recover monosemantic features from this superposition, by leveraging over-complete autoencoders (Bank et al., 2023) with sparsity constraints (Cunningham et al., 2024). Recent advancements (Gao et al., 2025; Bussmann et al., 2024; Braun et al., 2024; Rajamanoharan et al., 2024; Fel et al., 2025; Bhalla et al., 2026) in architectural design and training methodologies have enhanced the capability of SAEs to reconstruct neural network activations, while enabling the interpretation of the learned features, and the discovery of circuits (Dunefsky et al., 2024) and steerable features (Härle et al., 2025; Joseph et al., 2025a) in single modality models.

Vision-Language Models. Vision-Language Models (VLMs) (Radford et al., 2021; Zhai et al., 2023; Tschannen et al., 2025; Liu et al., 2023; Li et al., 2023) are multi-modal models trained on large-scale datasets containing both visual and textual data, enabling them to solve tasks across both modalities. While their cross-modal understanding has been applied to a wide range of applications, such as image captioning (Ashutosh et al., 2025), image generation (Rombach et al., 2022), and composed retrieval (Psomas et al., 2025), the interpretation of their internal representations and work remains under explored (Sharkey et al., 2025; Lin et al., 2025).

Sparse Autoencoders (SAEs) applied to Vision-Language Models (VLMs). Recently, some works have tried to decompose the activations of VLMs using human-intepretable concepts (Bhalla et al., 2024; Gandelsman et al., 2024) or monosemantic features learned by SAEs (Pach et al., 2025; Zhang et al., 2025; Lou et al., 2025), enabling applications such as steered text generation (Pach et al., 2025), enhanced multi-modal alignment (Lou et al., 2025) and object hallucination mitigation (Park et al., 2026). However, the training recipe for SAEs applied to contrastive VLM output spaces, where the SAE reconstructs activations from both modalities, suffers from the split-dictionary phenomenon (Bhalla et al., 2024; Costa et al., 2025), where the learned features are active for only one modality, which limits the usefulness of the learned decomposition for cross-modal understanding and applications. Recent works (Dhimoïla et al., 2026; Kaushik et al., 2026) mitigate this issue with an additional co-activation loss term that encourages the same features to be active across modalities. However, the impact of this training strategy on the overall interpretability and downstream performance of the learned decompositions have not been thoroughly investigated. In this work, we propose to use separate, modality-specific SAEs guided by the co-activation objective, producing synchronized features between modalities that achieve both superior interpretability and downstream performance, and more faithful reconstruction of the activations.

B Limitations and Future Works

While our work provides a comprehensive analysis of the downstream performance, interpretability and quality of SAE features across different sparsity levels, VLM models, and datasets, the results are limited to the TopK SAE architecture (Gao et al., 2025). Notably, this specific implementation doubles the number of parameters for our dual architecture relative to the shared one, whose reduction in parameter counts constitutes a key direction for future research. Although this architecture allows to isolate the effect of the sparsity, training loss and dataset on the properties of the learned features, it would be interesting to extend our analysis to other SAE architectures, such as the ones introduced by (Bussmann et al., 2024; 2025; Zaigrajew et al., 2025), to dinamically allocate the active features depending on the sample or to learn hierarchical features. Moreover, as discussed in Sec. F.3, the optimal sparsity level appears to depend on both the training dataset and the underlying VLM model. This suggests that enabling different sparsity levels for each modality within a dual architecture could be beneficial, and examining how such flexibility impacts the properties of the learned features would be an interesting direction for future work. Finally, since our analysis is restricted to VLM embedding spaces defined on the unit sphere $\mathbb{S}^{d-1}$, extending it to alternative spaces, such as hyperbolic spaces (Desai et al., 2023; Pal et al., 2025), represents another promising avenue for future research.

C NOTATION

We indicate the set of the first n natural numbers as $[n] = \{1, 2, \dots, n\}$. We denote vectors in bold lowercase (*e.g.*, $\mathbf{a}$), matrices in bold uppercase (*e.g.*, $\mathbf{A}$), and scalars in regular lowercase (*e.g.*, a). Indexing of vectors and matrices is done using subscripts, where $\mathbf{a}_i$ denotes the i-th element of vector $\mathbf{a}$, and $\mathbf{A}_{ij}$ denotes the element in the i-th row and j-th column of matrix $\mathbf{A}$. If we use a single subscript on a matrix, such as $\mathbf{A}_i$, it denotes the i-th column of $\mathbf{A}$. If subscripts are in parentheses, they represent the indices of the elements of a set (*e.g.*, $\mathbf{x}_{(i)}$ is the i-th element of the set). Superscripts (*e.g.*, in $\mathbf{a}^V$ or $\mathbf{A}^V$) are used to denote different modalities, with V and T representing the visual and textual modalities, respectively. We use $\mathbf{A}^\intercal$ to denote the transpose of matrix $\mathbf{A}$. We denote the L_p-norm as $\|\cdot\|_p$, and $\|\cdot\|_F$ for the Frobenius norm. The cosine similarity between two vectors $\mathbf{a}$ and $\mathbf{b}$ is denoted as $s(\mathbf{a}, \mathbf{b}) = \frac{\langle \mathbf{a}, \mathbf{b} \rangle}{\|\mathbf{a}\|_2 \|\mathbf{b}\|_2}$. We use $\odot$ to denote the Hadamard product. The trace of a square matrix $\mathbf{A}$ is denoted as $\mathrm{Tr}(\mathbf{A})$.

D IMPLEMENTATION DETAILS

Dataset Details. We use the CC3M (Sharma et al., 2018) and YFCC15M-V2 (Li et al., 2022; Cui et al., 2022) datasets for training SAEs. At the time of project development, the number of available samples for CC3M was 1.7 million, and for YFCC15M-V2 was 13.3 million. We use a held-out subset of approximately 169k samples and 1 million samples for evaluation for CC3M and YFCC15M-V2, respectively. The remaining samples are used for training. We extract and cache the activations of the visual and textual encoders of the contrastive VLMs for all samples to speed-up trainings and quality evaluations.

For zero-shot downstream performance evaluation, we use ImageNet (Deng et al., 2009; Russakovsky et al., 2015), CIFAR-10 and CIFAR-100 (Krizhevsky & Hinton, 2009), Caltech-101 (Fei-Fei et al., 2004), STL-10 (Coates et al., 2011), Food-101 (Bossard et al., 2014), CUB-2011 (Wah et al., 2011), Standford Cars (Krause et al., 2013), FGVC Aircraft (Maji et al., 2013), Oxford-IIIT Pet (Parkhi et al., 2012), Oxford Flowers (Nilsback & Zisserman, 2008), DTD (Cimpoi et al., 2014), EuroSAT (Helber et al., 2019), RESISC (Cheng et al., 2017) and Country211 (Radford et al., 2021) for classification, and the COCO validation set (Lin et al., 2014) and the Flickr30K test set (Young et al., 2014; Karpathy & Fei-Fei, 2015) for retrieval.

Training Details. We train all SAEs using the Adam optimizer (Kingma & Ba, 2014) for 300 epochs with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and without weight decay. We use a learning rate of 10^{-3} and a Cosine Annealing schedule (Loshchilov & Hutter, 2017) with a linear warmup of 10 epochs. We fix the batch size to $16,384$ and train with mixed precision using the PyTorch framework (Paszke et al., 2019) on NVIDIA A40 and RTX 3090 GPUs. During optimization, we ensure that the dictionary atoms $\mathbf{D}_j$ remain unit-norm. Before each optimization step, we remove the component of the gradient that is parallel to the current atom direction, and we clip the gradient at a maximum norm of 1. We instantiate the co-activation loss $\mathcal{L}_{\mathrm{co-act}}$ as the group-sparse loss introduced by (Kaushik et al., 2026)

$$\mathcal{L}_{\mathrm{gs}} = \mathbb{E}_{(\mathbf{z}^V, \mathbf{z}^T)} \left[\sum_{i=1}^{k} \sqrt{(\mathbf{z}_i^V)^2 + (\mathbf{z}_i^T)^2} \right], \tag{7}$$

or the InfoNCE loss (Oord et al., 2018; Wen et al., 2025)

$$\mathcal{L}_{\mathrm{InfoNCE}} = -\mathbb{E}_{(\mathbf{z}^V_{(i)}, \mathbf{z}^T_{(i)})} \left[\frac{1}{2} \left(\log \frac{e^{s(\mathbf{z}^V_{(i)}, \mathbf{z}^T_{(i)})}}{\sum_j e^{s(\mathbf{z}^V_{(i)}, \mathbf{z}^T_{(j)})}} + \log \frac{e^{s(\mathbf{z}^T_{(i)}, \mathbf{z}^V_{(i)})}}{\sum_j e^{s(\mathbf{z}^T_{(i)}, \mathbf{z}^V_{(j)})}} \right) \right], \tag{8}$$

and we set the hyperparameters of these losses to the values recommended in their respective original works. To improve the stability of training runs with the InfoNCE loss, we use a learnable temperature parameter τ, initialized to $\frac{1}{0.07}$ and clipped to a maximum of 100 as in (Radford et al., 2021). We also initially experimented with the alignment loss introduced by (Dhimoïla et al., 2026), defined as

$$\mathcal{L}_{\mathrm{align}} = -\mathbb{E}_{(\mathbf{z}^V, \mathbf{z}^T)} \left[(\mathbf{z}^V)^\intercal \mathbf{z}^T \right] \tag{9}$$

but we found that the group-sparse loss had higher interpretability performance, while the InfoNCE loss had higher downstream performance, and both are easier to optimize without requiring expensive hyperparameter tuning. For this reason, we focus on these two losses in our experiments.

Evaluation Details. For the quality and intepretability evaluations, we randomly sample 200k inputs for each modality from the held-out subset or the training set, depending on the metric, and we perform 5 iterations of evaluation due to resource constraints. The results are then averaged across

iterations. For the zero-shot classification evaluation, we embed multiple textual prompts and average the resulting embeddings to obtain a single embedding for each class. We use the same prompts as in (Desai et al., 2023; Pal et al., 2025), which extend the ones used in (Radford et al., 2021).

Metric Details. To evaluate the quality of the reconstructed activations, we measure the reconstruction error Fel et al. (2025), which evaluates how accurately the reconstructed activations approximate the original ones, namely

$$R^2 = 1 - \frac{\|\mathbf{A} - \hat{\mathbf{A}}\|_F^2}{\|\mathbf{A} - \bar{\mathbf{A}}\|_F^2}, \tag{10}$$

where $\bar{\mathbf{A}} \in \mathbb{R}^{N \times d}$ denotes the mean activation matrix. This metric is evaluated for a fixed sparsity level K, where we also compute the average L_0-norm of the codes.

The Bridge Matrix $\mathbf{B} \in \mathbb{R}^{k \times k}$ was introduced in Papadimitriou et al. (2025) to reflect the degree to which atoms i and j jointly support synchronization across modalities, by considering how often and how strongly they fire together, and how aligned they are, in a single score $\mathbf{B}_{ij}$. By computing the trace ratio we evaluate the degree to which cross-modal information is concentrated on one-to-one feature mappings.

To quantify the monosemanticity of individual features, Pach et al. (2025) proposed the MonoSemanticity score (MS). Given a pre-trained encoder $M(\cdot)$ for a modality, embeddings are extracted from N input samples $\{\mathbf{x}^{(i)}\}_{i=1}^N$ to form a similarity matrix $\mathbf{S} \in [-1, 1]^{N \times N}$, defined as

$$\mathbf{S}_{ij} = \frac{M(\mathbf{x}_{(i)})^\intercal M(\mathbf{x}_{(j)})}{\|M(\mathbf{x}_{(i)})\|_2 \|M(\mathbf{x}_{(j)})\|_2}. \tag{11}$$

The corresponding codes $\mathbf{Z} \in \mathbb{R}^{N \times k}$ are then computed and normalized independently for each dimension using min–max scaling, yielding $\bar{\mathbf{Z}} \in [0, 1]^{N \times k}$:

$$\bar{\mathbf{Z}}_{il} = \frac{\mathbf{Z}_{il} - \min_{j \in [N]} \mathbf{Z}_{jl}}{\max_{j \in [N]} \mathbf{Z}_{jl} - \min_{j \in [N]} \mathbf{Z}_{jl}}. \tag{12}$$

Subsequently, for the l-th feature, a relevance matrix $\mathbf{R}^l \in [0, 1]^{N \times N}$ is constructed to capture the joint activation strength across sample pairs:

$$\mathbf{R}^l_{ij} = \bar{\mathbf{Z}}_{il}\, \bar{\mathbf{Z}}_{jl}. \tag{13}$$

These relevance values are used to weight the similarity scores, resulting in the MS score for the l-th feature:

$$\mathrm{MS}^l = \frac{\sum\limits_{\substack{i,j=1 \\ i \neq j}}^{N} \mathbf{R}^l_{ij} \mathbf{S}_{ij}}{\sum\limits_{\substack{i,j=1 \\ i \neq j}}^{N} \mathbf{R}^l_{ij}}. \tag{14}$$

This score has been extended to multimodal settings by Kaushik et al. (2026) with the Multimodal MonoSemanticity score (MMS), which avoids the min-max normalization step and sums over all samples pairs from different modalities:

$$\mathrm{MMS}^l = \frac{\sum\limits_{i=1}^{N_1} \sum\limits_{j=1}^{N_2} \mathbf{Z}^V_{il} \mathbf{Z}^T_{il} \mathbf{S}_{ij}}{\sum\limits_{i=1}^{N_1} \sum\limits_{j=1}^{N_2} \mathbf{Z}^V_{il} \mathbf{Z}^T_{il}}, \tag{15}$$

where N_1 and N_2 are the number of samples from the two modalities. The similarity matrix $\mathbf{S}$ is computed using a multimodal encoder.

To compute the importance of a feature i for reconstructing the samples of a dataset, Papadimitriou et al. (2025) defined the energy score as the average activation of the feature across samples, *i.e.*, $\mathrm{Energy}_i = \mathbb{E}_{\mathbf{z}}[\mathbf{z}_i]$. A higher energy score indicates that the feature is more frequently and strongly activated across samples, suggesting that it captures a more general and potentially more important feature of the data. Additionally, they introduced the modality score to quantify if a feature is specific or shared across modalities. The modality score of a feature i is defined as $\mathrm{ModalityScore}_i = \frac{\mathbb{E}_{\mathbf{z}^V}[\mathbf{z}_i^V]}{\mathbb{E}_{\mathbf{z}^V}[\mathbf{z}_i^V] + \mathbb{E}_{\mathbf{z}^T}[\mathbf{z}_i^T]}$. A modality score close to 1 indicates that the feature is predominantly visual, while

a score close to 0 indicates that it is predominantly textual. A score around 0.5 suggests that the feature is shared between the two modalities.

E HUMAN STUDY

We conduct a small scale study to evaluate human preferences regarding learned concepts for Shared vs Independent SAEs. We focus on SAEs trained on CLIP ViT-L/14 using CC3M dataset. We select 500 random features from Shared SAE w/ InfoNCE loss and we find the most similar visual features in the visual dictionary of Independent SAE w/ InfoNCE loss in order to create hard pairs, reflecting a similar concept. We then collect the top-activating image and text examples and create figures similar to Fig. 9 for each feature. For each pair, we display to the user the features' figures in a randomized order. Before starting, the user must select the number of questions to be answered. Questions are shown in a randomized order. For each question, the user must select the most monosemantic feature by clicking on a button. The interface used for the study is shown in Fig. 10.

The subjects are recruited among students and teachers at our institution, involving a total of 9 participants and collecting a total of 223 responses. Across the preferences, Independent was preferred in 177 cases (79.4%) and Shared in 46 cases (20.6%). Under the null hypothesis where the probability of choosing Independent is 0.5, an exact two-sided binomial test gives a p-value of 2.59×10^{-19}. To account for the unequal number of questions answered by each user, we also include a participant-level analysis. The unweighted mean participant preference for Independent SAE is 72.6%. A one-sample t-test on participant independent rates gives a p-value of 0.007, while a non-parametric sign test across participants gives a p-value of 0.039.

While these results consistently support that Independent SAE features are considered more monosemantic than Shared SAE features, we acknowledge that study should be expanded to a larger scale and include a more robust statistical analysis.

F ADDITIONAL RESULTS

In this section, we present additional results that complement the main findings of our paper. We first provide the remaining results for the CC3M dataset in Sec. F.2, and then we present the results for the YFCC15M-V2 dataset in Sec. F.3. In particular, Sec. F.1 provides an analysis of the unexplained activation percentage using multi-modal atoms, Sec. F.2.1 eximanes how performance changes when varying the dictionary size, Sec. F.2.2 investigates the effect of halving the number of features for each modality in the independent SAE variants, and Sec. F.2.3 presents the results of cross-modal steering experiments.

Reconstruction Performance Details. For the reconstruction performance results shown in Figs. 4, 11 and 24, we exclude the results for the shared SAE with group-sparse loss for $K \in \{512, 1024\}$ for better readability, as its performance is significantly worse than the other SAEs. This is likely because at lower sparsity levels, the gradient signal of the group-sparse loss dominates the reconstruction gradient, as the reconstruction task becomes easier to solve with more active features. This consequently leads to a co-activation constraint that is too strict, compromising the reconstruction performance.

Trace Ratio Details. Regarding the computation of the trace ratio, we find that, while an explicit co-activation loss is indispensable for shared SAEs to extract cross-modal features across the entire sparsity spectrum, enforcing code synchronization during training on independent SAEs can paradoxically degrade their mapping purity for some sparsity levels (see Figs. 12a, 25a and 25d). This may stem from noise fitting and training artifacts within complex VLM spaces, where unsynchronized features capture incidental cross-modal noise. The use of the co-activation loss mitigates these noisy effects, stabilizing relative improvement trends in the trace ratio between i-SAEs and s-SAEs across all sparsity levels.

Monosemanticity Details. For the visualizations shown in Figs. 6, 13 and 26, we omit the results for $K \in \{512, 1024\}$ to improve readability; as noted in (Gao et al., 2025) for the textual modality, we find that SAEs with such sparsity levels learn less interpretable features across both modalities. However, while we observe that the interpretability performance for Image and Text MS scores follows a similar trend across sparsity levels for the CC3M dataset, with increasingly better performance with K decreasing, for the YFCC15M-V2 dataset we find that the Text MS scores reach the best performance at $K = 256$ for most VLM models and SAE variants. This suggests that the optimal sparsity level for each modality may depend on both the dataset and the VLM model. We leave a

Human Survey

This website runs a blind monosemanticity survey.

Goal

What is **monosemanticity**? **Monosemanticity** means that an **individual latent feature or neuron responds to one single, distinct, and human-understandable concept**.

For each question, you will see one full image containing two candidate features (left and right), each with top-activating images and texts together with their activation scores. Please select:

- `1` if the **left** feature looks more monosemantic
- `2` if the **right** feature looks more monosemantic

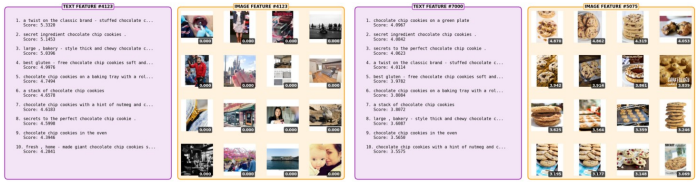
What to look for: a feature is more monosemantic when it represents a single clean concept in a stable, interpretable way.

How to answer

Inspect the top-activating images and texts, then pick the side that seems cleaner and more consistent.

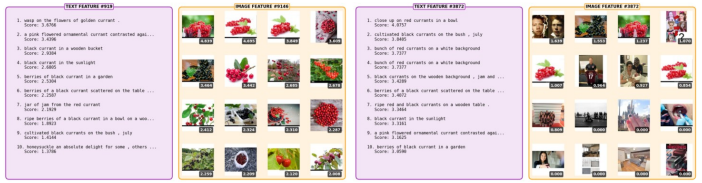
Example of questions

Example question 1:



As you can see, from this first example, both the features activate for sentences containing chocolate chip cookies information, but the first feature does not activate for images with related topics; indeed, it is completely off - please, see the zero 0 activation value for all the images shown (keep an eye on these situations; they're important). The second feature, on the other hand, activates also for images containing chocolate chip cookies. In this example, the correct answer is 2 (second feature).

Example question 2:



In this second example, both features activate for texts containing black and red currants information. However, while the first feature activates for images containing black and red currants, the second feature activates also for completely unrelated topics. Thus, the correct answer is 1 (first feature).

Pick the more monosemantic feature

1 = left panel, 2 = right panel

Question	Answered	Current pair id
1/100	0	114

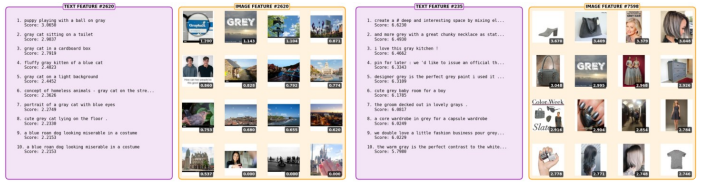


Figure 10: **Interface for human study.** The interface contains a brief introduction to the task, together with two examples of questions with relative explanations. Then, the user is presented with the actual questions to answer.

more in-depth analysis of this phenomenon for future work, and omit the results for $K = 256$ for better readability.

Feature Examples. In Figs. 14 and 15 and Figs. 27 and 28, we provide additional examples of learned features for independent SAE ($K = 32$) with the InfoNCE loss trained on the CC3M and YFCC15M-V2, respectively, for SigLIP 2 ViT-B/16. Furthermore, in Figs. 16 to 20, we compare the features learned by independent and shared SAEs with the InfoNCE loss trained on the CC3M for CLIP ViT-L/14. We observe that when both architectures learn features with the same or similar semantics, the independent SAE features activate more strongly across modalities, while the shared SAE features exhibit a mixture of unrelated concepts in their top-activating samples.

F.1 Unexplained activation percentage using multi-modal atoms

To illustrate that multi-modal atoms living in a subspace distinct from the modality-specific ones fail to effectively reconstruct the modality activations, we devise a simple experiment on the CC3M training set using SAE variants trained on the CLIP ViT-L/14 with $K = 32$. Specifically, we filter the dictionaries to keep only the atoms with a modality score between 0.05 and 0.95 (Dhimoïla et al., 2026). As each SAE variant has its own number of multi-modal atoms, we keep only the top-1024 multi-modal atoms according to their energy score for a fair comparison. We then solve the following optimization problem for the training set activation matrix $\mathbf{A}$ using projected gradient descent:

$$\mathbf{Z}^\star = \arg\min_{\mathbf{W} \in \mathbb{R}_+^{N \times k}} \|\mathbf{A} - \mathbf{W}(\mathbf{D}^B)^\intercal\|_F, \tag{16}$$

where $\mathbf{D}^B \in \mathbb{R}^{d \times 1024}$ is the filtered multi-modal dictionary. Finally, we measure the unexplained activation percentage as:

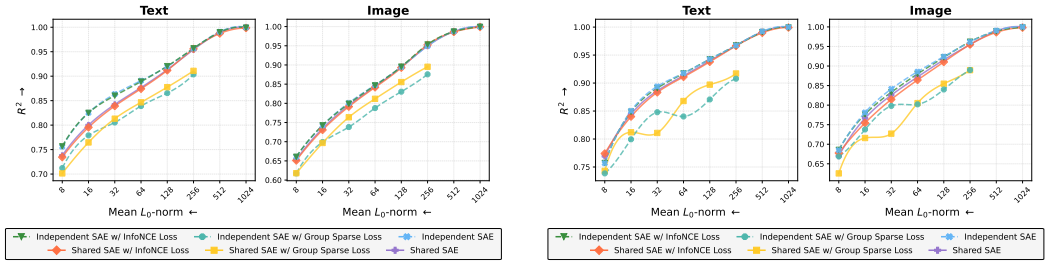
$$\text{Unexplained Activation Percentage} = \frac{\|\mathbf{A} - \mathbf{Z}^\star(\mathbf{D}^B)^\intercal\|_F}{\|\mathbf{A}\|_F}. \tag{17}$$

As shown in the results of Tab. 2, all shared SAE variants fail to explain more than 50% of the original activations, while the independent SAE variants can explain more than 60% for both modalities using only the highest-energy multi-modal atoms.

Table 2: Percentage (%) of unexplained activation by high-energy multi-modal atoms for CLIP ViT-L/14 ($K = 32$) on CC3M. Lower values are better.

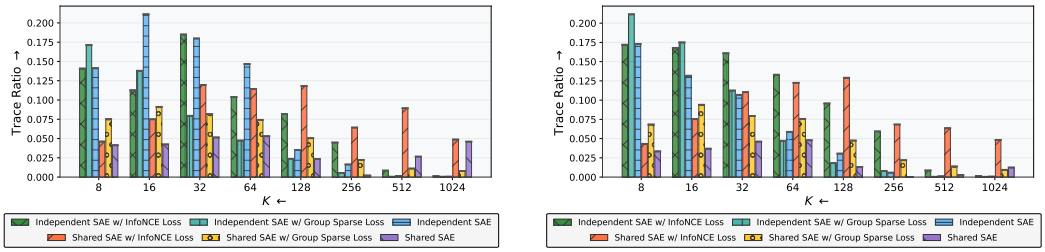
SAE	Image	Text
Shared	71.12	79.60
Shared w/ Group Sparse Loss	51.87	56.80
Shared w/ InfoNCE Loss	59.55	75.46
Independent	41.09	34.94
Independent w/ InfoNCE Loss	39.98	35.44

F.2 CC3M



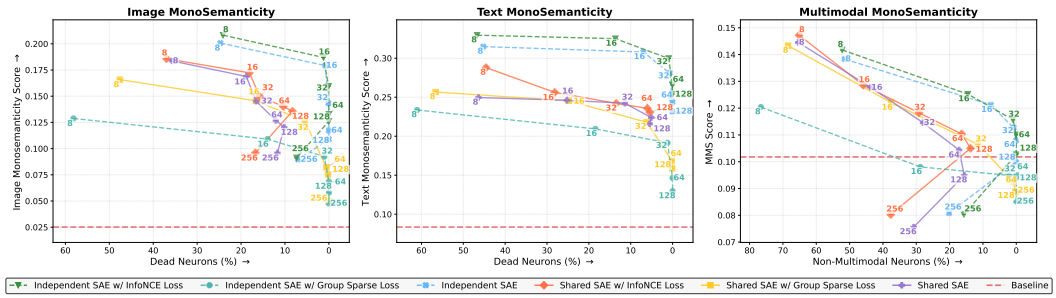
(a) CLIP ViT-L/14 (b) SigLIP 2 ViT-B/16

Figure 11: Reconstruction performance on CC3M for (a) CLIP ViT-L/14 and (b) SigLIP 2 ViT-B/16. Each plot shows the R^2 score of the reconstructed activations against the original activations for both modalities across different sparsity levels. The x-axis represents the mean L_0-norm of the codes (log scale).

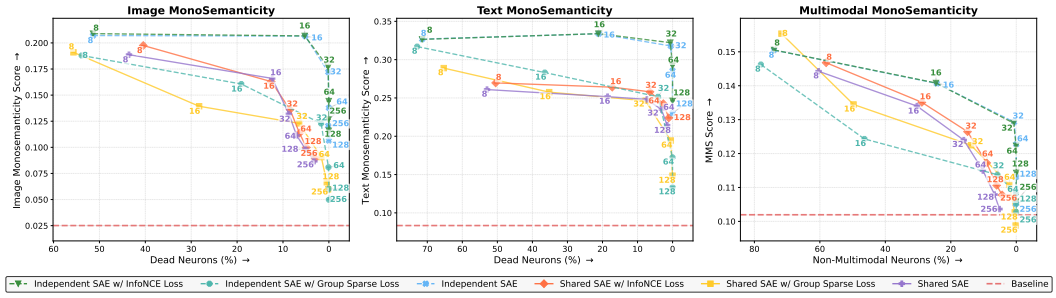


(a) CLIP ViT-L/14 (b) SigLIP 2 ViT-B/16

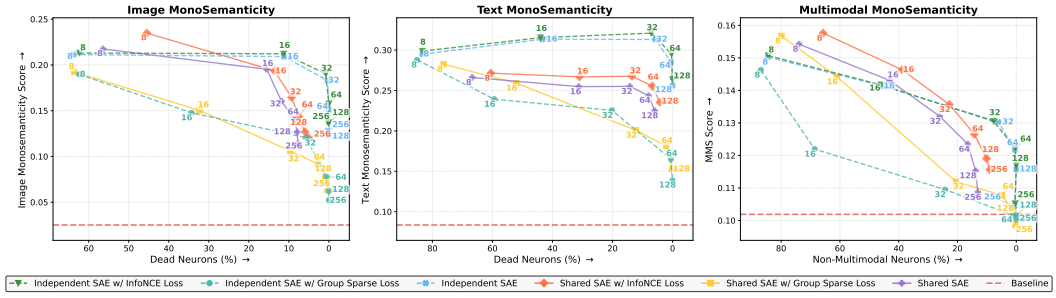
Figure 12: Trace Ratio comparison on CC3M for (a) CLIP ViT-L/14 and (b) SigLIP 2 ViT-B/16. Metric values, represented as bars, are grouped by the sparsity level K in TopK activation function.



(a) CLIP ViT-B/32



(b) SigLIP ViT-B/16



(c) SigLIP 2 ViT-B/16

Figure 13: Interpretability evaluation for CC3M for (a) CLIP ViT-B/32, (b) SigLIP ViT-B/16, and (c) SigLIP 2 ViT-B/16. The plots show (from left to right) Image MonoSemanticity (MS), Text MS, and Multimodal MonoSemanticity (MMS) scores. Each point corresponds to the mean score across all features for a given sparsity level, highlighted in correspondence of each point. The x-axis indicates the percentage of dead neurons, *i.e.*, neurons that are never activated across the evaluation set. For Multimodal MonoSemanticity scores, we indicate the percentage of non-multimodal neurons, *i.e.*, neurons that are never activated across both modalities in the evaluation set. The dashed horizontal line indicates the mean score across all features for the dense VLM model.

Table 3: Zero-shot classification accuracy (%) across datasets, grouped by sparsity level K, for CLIP ViT-B/32 on CC3M.

		General datasets						Fine-grained datasets							Misc. datasets				
K	SAE	Avg	ImageNet	CIFAR-10	CIFAR-100	Caltech-101	STL-10	Avg	Food-101	CUB	Cars	Aircraft	Pets	Flowers	Avg	DTD	EuroSAT	RESISC45	Country211
8	Shared	63.0	25.2	84.4	45.3	64.0	95.9	10.0	21.2	2.6	6.2	5.0	15.4	9.8	23.2	24.9	21.8	41.3	4.9
	Shared w/ Group Sparse Loss	47.8	7.4	80.6	28.2	26.6	96.1	8.5	15.5	1.1	2.9	3.4	18.4	9.7	18.4	9.8	17.4	44.0	2.5
	Shared w/ InfoNCE Loss	59.5	23.4	78.5	35.2	66.0	94.5	8.8	16.7	2.0	4.8	2.8	21.8	4.5	17.2	22.5	10.1	33.6	2.7
	Independent	67.1	35.5	84.5	46.0	73.4	96.2	**19.7**	**41.0**	7.0	**12.0**	**7.3**	**36.6**	14.1	**25.5**	**30.5**	20.6	44.7	6.2
	Independent w/ Group Sparse Loss	64.0	28.0	**86.0**	40.8	68.6	96.5	11.4	24.7	5.8	5.2	4.5	14.7	13.5	24.6	24.4	**24.9**	43.6	5.3
	Independent w/ InfoNCE Loss	**68.3**	**35.7**	85.8	**48.6**	**74.8**	**96.6**	19.2	40.3	**8.4**	11.0	6.0	30.5	**19.1**	24.2	28.5	16.2	**45.3**	**6.7**
16	Shared	68.1	39.1	85.9	45.9	73.4	96.4	21.2	36.9	8.1	9.7	6.4	44.6	21.2	29.9	33.4	31.3	48.3	6.5
	Shared w/ Group Sparse Loss	58.3	25.1	86.2	16.5	68.8	94.9	10.5	28.4	5.8	6.3	2.9	15.0	4.6	27.2	22.6	**31.5**	50.3	4.2
	Shared w/ InfoNCE Loss	65.6	28.7	85.0	46.6	71.5	96.4	15.2	27.5	5.3	6.3	5.6	36.0	10.7	22.5	25.5	16.4	43.6	4.4
	Independent	**71.7**	**43.6**	87.0	51.7	**79.5**	96.7	28.9	52.2	**15.6**	17.3	6.1	48.9	**33.1**	29.8	34.0	25.5	51.7	**8.1**
	Independent w/ Group Sparse Loss	68.9	36.5	87.1	49.3	74.8	**96.8**	20.8	41.8	10.6	8.0	7.3	33.0	24.4	28.9	29.1	27.1	52.4	7.2
	Independent w/ InfoNCE Loss	71.6	43.5	**87.3**	**52.1**	78.2	**96.8**	**29.8**	**54.7**	14.2	**18.9**	**9.5**	**50.8**	30.9	**32.8**	**37.0**	**31.5**	**54.5**	**8.1**
32	Shared	72.9	46.0	88.0	53.4	79.9	**97.1**	32.1	56.9	17.8	17.6	8.3	58.7	33.1	33.8	34.0	38.8	52.9	9.7
	Shared w/ Group Sparse Loss	66.0	33.0	**88.2**	44.3	67.8	96.7	16.7	39.2	8.0	10.5	5.0	24.5	12.8	31.4	32.4	37.2	50.1	5.8
	Shared w/ InfoNCE Loss	71.0	40.9	86.7	52.0	79.2	96.4	24.3	44.2	12.5	14.2	7.0	47.8	20.2	31.0	32.0	35.2	50.4	6.3
	Independent	**74.4**	48.9	87.8	**56.1**	**82.7**	96.5	38.3	63.4	**22.5**	26.3	11.2	64.0	**42.6**	34.3	37.6	34.3	**55.0**	**10.2**
	Independent w/ Group Sparse Loss	71.7	43.6	88.0	52.0	78.0	96.8	26.1	49.7	14.0	15.4	8.2	39.5	29.6	33.0	35.1	35.3	53.6	8.1
	Independent w/ InfoNCE Loss	74.3	**49.1**	87.7	55.9	82.5	96.5	**38.4**	**64.2**	21.3	**26.5**	**12.0**	**66.3**	40.2	**35.5**	**39.5**	**40.1**	52.7	9.8
64	Shared	76.3	53.0	**88.6**	58.7	84.5	96.9	43.4	70.6	27.5	32.0	12.0	73.9	44.3	35.7	39.4	36.5	55.8	11.1
	Shared w/ Group Sparse Loss	69.7	38.9	87.7	47.2	77.5	**97.3**	23.8	46.0	13.3	12.1	6.9	41.7	23.1	31.6	32.4	32.4	55.4	6.4
	Shared w/ InfoNCE Loss	74.9	51.3	87.8	57.3	81.4	96.7	42.4	69.2	28.5	31.8	11.5	71.7	41.4	37.0	38.7	41.0	**57.4**	11.1
	Independent	76.6	54.1	88.4	59.0	84.3	97.0	**46.1**	**72.4**	**30.7**	**37.2**	**12.4**	74.1	49.7	37.2	38.9	42.5	56.2	11.1
	Independent w/ Group Sparse Loss	73.3	46.9	88.1	54.8	80.0	96.8	31.6	53.1	19.8	24.7	10.3	47.2	34.2	32.9	37.2	31.6	54.9	7.9
	Independent w/ InfoNCE Loss	**76.9**	**54.2**	88.2	**60.4**	**85.2**	96.6	45.3	71.7	29.0	32.2	**12.4**	**76.1**	**50.2**	**38.2**	**42.1**	**43.3**	55.7	**11.6**
128	Shared	78.3	**58.1**	89.1	60.2	86.7	**97.2**	50.6	75.9	37.9	45.6	13.4	77.1	53.9	40.0	40.6	45.4	**60.4**	13.7
	Shared w/ Group Sparse Loss	75.0	49.6	87.7	56.9	83.7	**97.2**	37.0	62.2	22.9	23.6	8.4	67.1	37.8	32.7	34.9	29.4	57.3	9.2
	Shared w/ InfoNCE Loss	78.4	**58.1**	**89.3**	61.2	86.1	**97.2**	48.7	74.4	37.0	41.3	13.5	76.4	49.8	39.3	41.3	**47.4**	56.5	12.0
	Independent	78.4	57.8	88.4	61.7	**87.0**	**97.2**	**51.9**	75.9	**38.4**	**46.1**	15.0	**81.0**	**54.9**	39.0	41.5	42.0	58.8	**13.8**
	Independent w/ Group Sparse Loss	75.9	51.7	88.7	59.2	82.8	97.1	40.3	62.5	27.7	34.2	11.2	61.5	44.7	35.8	38.9	36.7	57.8	10.0
	Independent w/ InfoNCE Loss	**78.5**	**58.1**	89.0	**62.0**	86.1	**97.2**	50.8	**77.2**	36.4	43.0	**16.4**	78.0	54.0	**40.8**	**43.8**	45.8	59.9	13.5
256	Shared	79.9	61.2	**90.0**	63.9	87.2	**97.1**	57.6	80.7	**49.6**	54.4	18.0	82.8	59.8	41.0	**44.4**	42.9	61.0	15.7
	Shared w/ Group Sparse Loss	4.4	0.1	10.0	1.0	1.0	10.0	1.1	1.0	0.5	0.5	1.0	2.7	1.0	3.7	2.1	10.0	2.2	0.5
	Shared w/ InfoNCE Loss	79.9	**62.0**	89.8	63.6	87.1	97.0	**58.1**	80.3	49.1	**54.7**	**18.8**	**84.9**	60.8	41.2	43.0	46.2	60.6	15.1
	Independent	79.7	61.8	89.3	**64.6**	86.0	97.0	58.0	81.0	48.4	53.4	18.1	84.5	**62.5**	40.8	43.1	43.7	60.5	**15.8**
	Independent w/ Group Sparse Loss	4.4	0.1	10.0	1.0	1.0	10.0	1.1	1.0	0.5	0.5	1.0	2.7	1.0	3.7	2.1	10.0	2.2	0.5
	Independent w/ InfoNCE Loss	**80.0**	61.9	89.5	64.3	**87.3**	96.8	57.6	**81.4**	48.3	52.3	18.6	84.8	60.3	**43.4**	43.1	**52.9**	**61.6**	**15.8**
	CLIP ViT-B/32	80.8	63.9	89.8	65.1	88.3	97.1	61.4	82.9	53.7	59.7	19.7	87.0	65.2	42.8	44.6	49.1	60.2	17.2

Table 4: Zero-shot retrieval performance across datasets, grouped by sparsity level K, for CLIP ViT-B/32 on CC3M.

		Text Retrieval								Image Retrieval							
		Flickr30k				COCO				Flickr30k				COCO			
K	SAE	Avg	R@1	R@5	R@10	Avg	R@1	R@5	R@10	Avg	R@1	R@5	R@10	Avg	R@1	R@5	R@10
8	Shared	48.5	21.8	55.2	68.5	30.6	13.0	33.5	45.4	49.0	27.3	54.0	65.8	26.6	11.0	28.8	39.9
	Shared w/ Group Sparse Loss	29.2	9.8	31.4	46.5	14.8	4.4	15.0	25.0	38.7	17.8	42.8	55.5	16.1	4.2	16.7	27.3
	Shared w/ InfoNCE Loss	20.6	1.7	19.2	40.9	18.1	4.4	18.8	31.0	48.6	26.4	53.5	65.8	25.9	9.8	28.0	39.8
	Independent	59.2	37.8	65.1	74.6	**37.2**	**19.4**	40.2	**52.0**	51.6	29.8	56.7	**68.2**	**29.8**	**13.8**	32.0	**43.5**
	Independent w/ Group Sparse Loss	58.4	35.9	64.8	74.6	33.6	16.5	36.7	47.5	**51.7**	29.9	**57.3**	67.8	28.2	13.0	30.3	41.3
	Independent w/ InfoNCE Loss	**60.3**	**38.4**	**66.8**	**75.8**	37.1	19.1	**40.3**	51.9	51.3	**30.3**	56.4	67.2	29.7	**13.8**	**32.1**	43.1
16	Shared	66.8	41.4	74.5	84.5	46.4	25.2	51.3	62.7	60.4	38.3	66.4	76.5	35.1	16.8	38.3	50.1
	Shared w/ Group Sparse Loss	42.3	13.3	47.3	66.4	31.6	11.6	34.4	48.9	53.3	29.7	59.1	71.0	27.1	11.0	29.1	41.3
	Shared w/ InfoNCE Loss	46.2	18.1	52.1	68.3	34.6	16.2	37.5	50.1	58.9	36.9	64.9	74.8	34.2	16.2	37.3	49.1
	Independent	**73.5**	52.7	**80.3**	**87.5**	49.8	29.1	54.2	66.0	**62.7**	40.1	**69.2**	**78.9**	37.2	18.9	40.3	52.4
	Independent w/ Group Sparse Loss	69.8	49.0	76.0	84.4	45.0	24.6	49.1	61.4	61.1	38.8	66.8	77.8	35.2	17.4	38.2	49.9
	Independent w/ InfoNCE Loss	72.7	**52.9**	78.4	86.9	**50.3**	**29.3**	**55.1**	**66.6**	62.4	**40.4**	68.6	78.3	**38.2**	**19.5**	**41.7**	**53.3**
32	Shared	76.5	52.9	84.7	91.8	54.9	33.5	60.2	71.1	67.7	46.3	74.2	82.7	40.4	21.3	44.1	55.8
	Shared w/ Group Sparse Loss	55.7	21.6	64.3	81.2	42.1	19.4	47.0	60.0	53.5	21.1	63.3	76.2	28.2	10.2	30.7	43.8
	Shared w/ InfoNCE Loss	74.4	52.0	81.6	89.5	49.4	28.3	53.8	66.1	66.2	44.2	72.4	82.1	40.3	20.8	44.1	56.0
	Independent	**81.6**	**63.9**	**87.6**	**93.4**	57.2	36.4	62.5	72.8	68.3	46.6	74.6	83.6	41.8	22.5	45.8	57.1
	Independent w/ Group Sparse Loss	78.5	58.8	85.1	91.7	52.5	31.2	57.6	68.8	68.0	46.5	74.5	83.1	39.8	20.3	43.8	55.3
	Independent w/ InfoNCE Loss	80.7	62.4	86.8	93.0	**58.1**	**37.3**	**63.4**	**73.5**	**68.7**	**47.4**	**75.1**	**83.7**	**42.8**	**23.3**	**46.7**	**58.3**
64	Shared	83.5	67.3	88.6	94.5	60.6	39.8	66.4	75.7	72.0	51.9	78.1	86.0	44.6	24.7	48.8	60.3
	Shared w/ Group Sparse Loss	67.0	37.6	76.5	86.8	47.9	25.2	53.2	65.3	61.0	37.3	67.8	78.0	34.7	16.5	37.9	49.6
	Shared w/ InfoNCE Loss	84.3	67.0	90.4	95.5	60.5	39.5	65.9	76.1	70.8	49.1	77.1	86.2	43.5	23.9	47.5	59.2
	Independent	**86.2**	**71.3**	91.0	**96.4**	62.0	40.7	**67.8**	**77.6**	72.7	51.8	79.4	86.9	45.4	25.4	49.6	61.3
	Independent w/ Group Sparse Loss	81.7	63.2	88.0	93.9	55.7	34.1	61.0	72.0	69.3	47.7	75.9	84.3	41.7	22.4	45.4	57.2
	Independent w/ InfoNCE Loss	86.0	70.9	**91.4**	95.6	**62.5**	**42.5**	67.6	77.4	**73.5**	**53.3**	**79.9**	**87.3**	**45.8**	**25.8**	**49.9**	**61.6**
128	Shared	86.7	71.4	92.3	96.3	64.8	44.3	70.7	79.5	74.3	54.5	80.2	88.1	47.2	27.3	51.5	62.9
	Shared w/ Group Sparse Loss	68.9	38.8	77.9	90.0	55.5	32.1	61.6	72.7	68.9	47.4	75.4	83.9	41.3	21.7	45.2	56.9
	Shared w/ InfoNCE Loss	88.0	74.7	92.6	96.6	65.1	45.2	70.5	79.5	74.6	54.5	81.0	88.4	47.5	27.2	51.9	63.5
	Independent	88.2	73.9	**93.5**	**97.3**	65.2	45.3	70.4	79.9	74.9	55.5	81.0	88.1	47.1	27.2	51.3	62.9
	Independent w/ Group Sparse Loss	84.6	68.9	90.1	94.8	60.2	38.6	65.9	76.0	71.1	50.7	77.5	85.2	43.9	24.0	47.9	59.7
	Independent w/ InfoNCE Loss	**88.6**	**75.4**	93.2	97.2	**66.2**	**45.8**	**71.7**	**81.0**	**75.6**	**56.5**	**81.7**	**88.5**	**48.4**	**28.2**	**52.8**	**64.1**
256	Shared	88.0	75.0	92.6	96.5	67.0	47.3	72.5	81.1	75.7	56.8	81.7	88.6	48.9	29.0	53.0	64.6
	Shared w/ Group Sparse Loss	0.1	0.1	0.1	0.2	0.0	0.0	0.0	0.0	0.5	0.1	0.5	1.0	0.1	0.0	0.1	0.2
	Shared w/ InfoNCE Loss	**89.7**	76.7	**94.9**	**97.6**	**68.6**	**49.6**	74.0	**82.2**	76.4	57.7	82.4	89.1	49.7	29.7	54.0	65.3
	Independent	89.7	**77.3**	94.1	**97.6**	68.4	49.1	73.8	**82.2**	76.9	58.2	82.9	**89.7**	49.4	29.5	53.6	65.1
	Independent w/ Group Sparse Loss	0.1	0.1	0.1	0.2	0.0	0.0	0.0	0.0	0.5	0.1	0.5	1.0	0.1	0.0	0.1	0.2
	Independent w/ InfoNCE Loss	89.4	76.9	94.2	97.2	68.5	49.2	**74.1**	**82.2**	**77.2**	**58.7**	**83.1**	**89.7**	**50.0**	**29.8**	**54.4**	**65.8**
	CLIP ViT-B/32	90.8	79.3	95.0	98.1	69.3	50.0	74.8	83.2	77.4	58.8	83.4	90.1	50.5	30.5	54.8	66.2

Table 5: Zero-shot classification accuracy (%) across datasets, grouped by sparsity level K, for CLIP ViT-L/14 on CC3M.

		General datasets						Fine-grained datasets							Misc. datasets				
K	SAE	Avg	ImageNet	CIFAR-10	CIFAR-100	Caltech-101	STL-10	Avg	Food-101	CUB	Cars	Aircraft	Pets	Flowers	Avg	DTD	EuroSAT	RESISC45	Country211
8	Shared	56.9	3.6	90.7	52.5	38.5	99.0	6.9	19.1	1.1	4.2	1.7	12.1	3.0	28.4	20.5	42.6	46.7	3.8
	Shared w/ Group Sparse Loss	59.9	8.1	87.8	46.2	58.5	98.9	8.0	25.2	1.5	4.3	1.2	8.9	7.0	27.7	21.6	34.2	50.2	4.8
	Shared w/ InfoNCE Loss	52.2	0.9	89.0	36.3	35.9	98.8	12.4	27.6	1.6	7.0	1.8	29.1	7.5	22.7	18.5	28.1	39.2	4.8
	Independent	77.0	45.3	92.2	**66.2**	82.2	99.1	25.1	53.0	**8.4**	20.4	**9.6**	33.0	**26.3**	**37.5**	37.8	**47.3**	52.8	**12.2**
	Independent w/ Group Sparse Loss	74.5	38.3	**93.1**	63.7	78.0	**99.3**	20.4	47.4	5.4	11.2	8.7	28.7	21.1	35.3	31.1	46.7	52.4	10.8
	Independent w/ InfoNCE Loss	**77.1**	**45.4**	91.5	65.3	**84.3**	99.0	**26.3**	**57.7**	7.5	**22.0**	8.8	**35.9**	25.7	37.3	**40.3**	43.7	**53.3**	12.0
16	Shared	75.9	46.2	92.5	61.3	80.2	99.2	24.5	57.5	9.2	16.9	7.6	33.9	21.8	40.2	43.1	50.9	55.2	11.5
	Shared w/ Group Sparse Loss	66.9	17.5	92.4	48.6	77.2	99.0	18.4	48.1	7.9	9.6	3.6	29.8	11.2	32.5	32.7	41.0	48.3	8.2
	Shared w/ InfoNCE Loss	67.3	11.0	93.6	52.5	80.5	99.0	19.3	51.2	4.4	10.9	7.4	36.1	5.6	33.8	37.3	40.7	51.8	5.5
	Independent	**81.2**	**55.2**	93.0	**70.5**	**88.1**	**99.3**	**39.8**	71.0	16.8	**37.2**	**13.9**	55.0	**44.9**	**44.2**	46.3	50.6	**63.6**	16.2
	Independent w/ Group Sparse Loss	79.4	49.5	**94.9**	67.0	86.3	99.1	30.6	63.6	11.9	21.7	11.8	39.5	35.2	40.7	41.0	46.4	61.4	14.1
	Independent w/ InfoNCE Loss	80.7	**55.2**	93.7	68.5	87.2	99.1	39.8	**71.2**	**19.2**	33.6	9.9	**62.1**	42.6	44.0	**46.9**	**51.7**	60.9	**16.3**
32	Shared	80.7	56.8	94.1	67.2	86.3	99.3	39.9	72.7	15.3	32.1	14.7	59.3	45.6	46.0	**49.6**	48.9	**68.0**	17.7
	Shared w/ Group Sparse Loss	74.9	41.8	94.3	56.4	82.8	**99.4**	24.5	59.3	8.0	16.4	9.9	30.6	23.1	35.5	36.5	52.3	41.8	11.5
	Shared w/ InfoNCE Loss	81.2	56.6	94.1	69.4	86.7	99.3	38.0	73.7	19.4	29.1	11.3	58.7	35.6	45.4	47.7	53.1	65.5	15.3
	Independent	82.8	**60.9**	94.2	71.1	88.5	99.2	**48.0**	77.8	**22.9**	**46.7**	16.9	**71.5**	52.3	47.6	49.5	56.5	64.6	**20.0**
	Independent w/ Group Sparse Loss	80.6	54.2	**94.4**	67.7	87.8	99.1	35.0	64.5	15.9	27.7	9.6	52.4	40.0	45.1	43.7	55.5	66.9	14.2
	Independent w/ InfoNCE Loss	**83.5**	60.5	93.4	**73.0**	**91.0**	**99.4**	47.6	**81.0**	22.8	46.0	**18.9**	64.4	**52.7**	**48.3**	47.4	**60.6**	65.8	19.4
64	Shared	84.2	64.6	93.9	72.3	90.6	**99.4**	54.9	83.6	30.9	52.2	18.8	**82.8**	**61.1**	50.3	**50.7**	59.4	68.6	22.6
	Shared w/ Group Sparse Loss	79.3	51.6	94.1	65.0	86.7	99.2	34.7	70.0	20.3	30.5	17.2	47.7	22.7	39.9	39.9	51.3	54.5	14.0
	Shared w/ InfoNCE Loss	84.2	64.8	93.7	72.4	90.7	99.3	54.8	82.8	**36.5**	53.7	17.7	78.4	59.8	50.2	**50.7**	60.3	68.2	21.5
	Independent	85.4	66.4	94.3	75.4	91.6	**99.4**	**56.3**	**85.7**	32.3	**58.5**	19.4	81.9	60.1	50.9	50.1	61.7	68.7	**23.2**
	Independent w/ Group Sparse Loss	83.0	60.4	**95.0**	69.9	90.4	99.2	44.8	75.8	24.6	38.1	15.1	65.5	49.7	47.6	43.3	62.6	67.5	17.0
	Independent w/ InfoNCE Loss	**85.5**	**66.6**	94.5	**75.5**	**91.7**	99.2	56.2	85.5	36.0	55.2	**20.8**	79.6	59.8	**51.4**	49.0	**64.8**	**69.0**	22.7
128	Shared	86.3	70.2	**95.2**	74.4	92.3	99.2	63.2	87.9	49.7	59.8	23.7	**91.2**	67.1	51.6	49.8	60.7	69.9	25.8
	Shared w/ Group Sparse Loss	83.3	62.2	95.0	69.3	90.7	**99.3**	51.1	80.9	36.3	50.0	17.0	80.8	41.3	48.8	47.0	62.0	65.7	20.3
	Shared w/ InfoNCE Loss	86.4	**71.2**	94.6	75.0	92.1	**99.3**	**64.8**	**88.8**	**50.8**	66.2	23.6	89.6	**69.6**	51.2	51.8	56.3	**70.8**	25.9
	Independent	86.5	70.4	94.8	**75.7**	92.4	**99.3**	63.6	88.4	46.2	66.5	22.2	88.5	69.5	51.7	51.9	58.4	70.3	**26.2**
	Independent w/ Group Sparse Loss	84.4	64.8	**95.2**	72.0	90.6	**99.3**	51.5	79.3	34.4	49.9	19.3	72.8	53.4	47.2	46.1	58.1	64.9	19.6
	Independent w/ InfoNCE Loss	**86.6**	70.3	95.1	**75.7**	**92.5**	**99.3**	63.4	**88.8**	45.5	**67.2**	**26.1**	87.6	65.1	**53.2**	**53.4**	**63.7**	69.9	26.0
256	Shared	87.3	73.0	95.3	75.4	**93.4**	**99.4**	**69.3**	**90.8**	**57.7**	71.9	**29.9**	91.4	74.0	53.4	52.7	60.7	71.4	28.7
	Shared w/ Group Sparse Loss	85.8	68.2	**95.4**	73.9	92.4	99.3	59.4	86.1	42.9	57.6	23.6	82.8	63.2	50.9	49.0	61.8	69.4	23.5
	Shared w/ InfoNCE Loss	87.5	73.4	95.3	76.5	93.1	99.3	68.7	90.1	56.8	71.6	28.0	90.6	**75.3**	**54.5**	**53.7**	**64.4**	71.0	28.8
	Independent	87.6	**73.5**	95.2	76.8	93.1	**99.4**	68.7	90.5	55.4	72.9	29.2	90.8	73.3	54.0	52.8	63.1	71.4	28.6
	Independent w/ Group Sparse Loss	86.2	68.9	95.3	74.9	92.7	**99.4**	59.9	86.9	44.6	62.5	22.7	83.4	59.3	51.0	49.7	62.1	68.7	23.5
	Independent w/ InfoNCE Loss	**87.7**	73.4	95.3	**76.9**	93.3	**99.4**	69.1	90.7	55.8	**73.8**	29.1	**92.4**	72.9	54.3	52.9	63.5	**71.8**	**29.2**
	CLIP ViT-L/14	88.4	75.2	95.6	78.3	93.3	99.4	72.7	91.6	62.5	77.9	32.1	93.2	78.9	55.5	54.5	63.9	71.7	31.7

Table 6: Zero-shot retrieval performance across datasets, grouped by sparsity level K, for CLIP ViT-L/14 on CC3M.

		Text Retrieval								Image Retrieval							
		Flickr30k				COCO				Flickr30k				COCO			
K	SAE	Avg	R@1	R@5	R@10	Avg	R@1	R@5	R@10	Avg	R@1	R@5	R@10	Avg	R@1	R@5	R@10
8	Shared	7.3	3.3	7.1	11.6	9.3	2.5	9.3	16.2	55.2	32.4	61.3	71.9	30.5	14.3	33.0	44.2
	Shared w/ Group Sparse Loss	40.1	18.8	44.4	57.2	21.6	8.1	23.1	33.7	50.9	27.8	56.3	68.7	24.7	10.5	26.5	37.1
	Shared w/ InfoNCE Loss	6.7	1.8	5.8	12.5	13.0	3.6	12.6	22.7	54.3	31.4	60.4	71.1	28.8	13.4	31.1	41.9
	Independent	**67.6**	45.8	**74.4**	**82.7**	42.8	24.0	46.7	57.6	56.8	34.8	62.4	73.1	32.9	16.1	35.6	46.9
	Independent w/ Group Sparse Loss	63.8	41.7	70.1	79.7	38.4	20.7	42.0	52.4	54.5	32.3	60.2	70.9	30.9	14.8	33.3	44.5
	Independent w/ InfoNCE Loss	67.4	**46.0**	73.7	82.4	**43.6**	**25.5**	**46.9**	**58.3**	**57.1**	**35.0**	**63.1**	**73.3**	**33.4**	**16.4**	**36.2**	**47.5**
16	Shared	69.5	43.3	78.1	87.2	49.9	28.0	55.0	66.7	64.1	42.9	70.2	79.2	37.8	19.7	41.1	52.7
	Shared w/ Group Sparse Loss	62.3	35.8	69.3	81.7	40.6	20.6	44.5	56.7	62.5	39.7	68.8	79.1	34.9	17.6	37.8	49.2
	Shared w/ InfoNCE Loss	37.9	12.2	40.5	61.0	39.5	16.5	43.6	58.3	63.5	41.4	69.9	79.1	37.5	19.5	40.7	52.2
	Independent	78.8	59.7	84.7	92.1	54.2	34.0	58.7	**70.0**	64.7	43.1	71.3	79.7	39.4	21.1	42.8	54.2
	Independent w/ Group Sparse Loss	77.5	58.6	83.0	90.8	51.6	31.4	56.1	67.4	**66.1**	**43.4**	**72.8**	**82.0**	**40.2**	21.3	**43.9**	**55.5**
	Independent w/ InfoNCE Loss	**79.3**	**60.2**	**85.1**	**92.7**	**54.3**	**34.2**	**58.8**	69.9	65.3	**43.4**	71.8	80.6	39.8	**21.5**	43.3	54.7
32	Shared	82.1	62.2	89.4	94.8	60.1	39.3	65.4	75.5	70.2	49.7	76.4	84.4	44.1	24.8	48.1	59.5
	Shared w/ Group Sparse Loss	68.3	42.0	75.6	87.3	50.5	27.6	55.9	68.1	69.6	48.0	76.3	84.5	41.6	22.3	45.3	57.1
	Shared w/ InfoNCE Loss	82.4	63.7	88.7	94.9	59.4	38.4	64.5	75.4	69.3	48.4	75.4	84.2	44.4	25.1	48.4	59.7
	Independent	**86.5**	**71.8**	**92.1**	95.6	62.3	**41.9**	67.3	77.6	70.6	49.8	77.3	84.7	44.3	24.7	48.2	60.1
	Independent w/ Group Sparse Loss	82.6	64.5	88.9	94.3	56.8	36.4	61.8	72.3	69.2	48.0	75.8	83.8	42.1	22.7	45.9	57.6
	Independent w/ InfoNCE Loss	86.4	71.3	91.6	**96.2**	**63.0**	**41.9**	**68.4**	**78.6**	**71.9**	**51.0**	**78.5**	**86.2**	**45.6**	**25.9**	**49.9**	**61.1**
64	Shared	88.3	74.2	93.7	96.9	66.2	46.2	71.3	81.1	75.3	56.0	81.5	88.5	49.1	29.0	53.8	64.6
	Shared w/ Group Sparse Loss	64.2	37.7	70.8	84.1	58.0	34.4	64.6	74.9	74.0	53.4	80.4	88.1	46.3	26.4	50.6	62.0
	Shared w/ InfoNCE Loss	89.0	75.0	94.5	97.4	66.6	46.5	72.1	81.3	75.6	55.8	82.2	**88.7**	49.5	**29.4**	53.9	65.1
	Independent	90.0	77.3	**94.9**	**97.7**	67.7	47.8	73.3	82.1	75.2	55.4	82.1	88.2	48.7	28.8	53.0	64.4
	Independent w/ Group Sparse Loss	87.8	73.7	92.8	97.0	62.6	41.5	68.1	78.1	74.3	53.9	80.7	88.2	46.0	26.3	50.2	61.6
	Independent w/ InfoNCE Loss	**90.3**	**78.6**	94.8	97.6	**68.2**	**48.5**	**73.9**	**82.2**	**76.2**	**57.4**	**82.4**	**88.7**	**49.7**	29.3	**54.2**	**65.5**
128	Shared	90.7	77.8	95.7	98.6	69.8	50.3	75.1	84.0	78.6	60.2	84.7	90.9	52.2	32.0	56.8	67.7
	Shared w/ Group Sparse Loss	69.5	43.2	76.8	88.5	60.7	36.3	67.6	78.1	76.0	56.4	82.6	89.0	49.2	29.3	53.7	64.7
	Shared w/ InfoNCE Loss	91.8	80.7	96.2	98.4	70.4	51.4	75.7	84.0	**79.5**	**61.4**	**85.6**	**91.5**	**52.8**	**32.5**	**57.7**	**68.3**
	Independent	**92.7**	82.0	**97.2**	**98.9**	70.6	50.3	76.5	**85.0**	77.8	59.4	84.0	90.1	51.8	31.4	56.7	67.4
	Independent w/ Group Sparse Loss	90.1	78.0	94.4	97.9	65.8	45.6	71.4	80.3	75.8	56.1	82.2	89.1	48.1	28.3	52.4	63.6
	Independent w/ InfoNCE Loss	92.6	**82.4**	96.9	98.5	**71.5**	**52.7**	**76.9**	84.8	79.2	61.0	85.3	91.2	52.4	32.3	57.1	67.8
256	Shared	92.5	82.6	96.1	98.7	72.4	53.2	78.3	85.7	80.2	62.6	86.3	91.7	54.0	34.1	58.6	69.4
	Shared w/ Group Sparse Loss	85.4	67.4	92.1	96.7	66.4	45.6	72.3	81.2	77.5	59.1	83.5	90.0	50.9	31.1	55.3	66.4
	Shared w/ InfoNCE Loss	92.4	82.5	96.3	98.4	73.0	54.7	78.3	**86.1**	80.5	62.8	86.5	**92.1**	54.4	34.4	59.2	69.6
	Independent	93.5	84.1	**97.2**	**99.1**	72.9	54.5	78.3	85.9	80.1	62.2	86.6	91.6	54.3	34.3	59.0	69.5
	Independent w/ Group Sparse Loss	90.8	79.0	95.4	97.9	68.4	49.5	73.5	82.3	78.2	59.4	84.6	90.5	50.6	30.7	55.2	66.0
	Independent w/ InfoNCE Loss	**93.5**	**84.7**	96.9	98.9	**73.3**	**55.4**	**78.4**	**86.1**	**80.6**	**63.2**	**86.7**	92.0	**54.6**	**34.6**	**59.4**	**69.8**
	CLIP ViT-L/14	94.0	85.3	97.4	99.3	74.2	56.1	79.6	86.9	81.3	64.7	87.1	92.2	55.2	35.5	59.9	70.3

Table 7: Zero-shot classification accuracy (%) across datasets, grouped by sparsity level K, for SigLIP ViT-B/16 on CC3M.

		General datasets						Fine-grained datasets							Misc. datasets				
K	SAE	Avg	ImageNet	CIFAR-10	CIFAR-100	Caltech-101	STL-10	Avg	Food-101	CUB	Cars	Aircraft	Pets	Flowers	Avg	DTD	EuroSAT	RESISC45	Country211
8	Shared	72.5	40.1	85.6	61.2	77.8	**97.7**	16.2	44.4	2.2	2.8	7.0	30.4	10.7	29.1	**41.4**	22.6	48.3	4.3
	Shared w/ Group Sparse Loss	70.4	36.2	85.9	59.3	74.4	96.1	17.2	38.7	5.9	6.4	5.9	28.7	17.4	28.4	37.0	25.0	47.1	4.5
	Shared w/ InfoNCE Loss	72.1	39.7	85.4	59.8	78.1	97.6	16.8	43.0	0.8	5.4	**7.3**	**42.4**	1.8	30.0	38.5	**27.7**	49.4	4.6
	Independent	73.4	42.5	85.3	61.5	**80.0**	97.6	**23.8**	**47.1**	**8.2**	**21.4**	6.7	35.8	23.5	30.0	37.3	27.2	50.1	**5.5**
	Independent w/ Group Sparse Loss	72.7	40.1	**87.0**	61.0	78.3	97.2	21.9	42.2	3.8	19.6	4.0	38.0	**23.6**	29.8	38.1	25.3	50.4	5.4
	Independent w/ InfoNCE Loss	**73.6**	**42.8**	86.6	**61.8**	79.3	97.4	22.3	42.9	5.7	20.7	5.3	37.3	21.9	**30.3**	39.8	25.5	**50.5**	**5.5**
16	Shared	75.9	47.7	89.5	64.1	80.4	**98.0**	24.2	57.6	6.2	13.2	6.6	51.3	10.5	33.9	51.0	23.2	55.5	6.0
	Shared w/ Group Sparse Loss	74.1	43.0	89.2	63.3	77.2	97.9	22.6	42.7	8.5	25.3	6.2	45.3	7.5	30.7	46.8	17.5	53.4	5.3
	Shared w/ InfoNCE Loss	75.4	42.9	**90.0**	64.3	82.0	97.7	19.2	43.4	4.8	9.3	**8.8**	42.3	6.4	31.1	43.8	22.4	54.3	4.0
	Independent	**79.6**	54.8	89.6	**66.8**	**88.9**	**98.0**	38.9	**64.9**	14.1	**43.1**	8.1	65.5	37.9	35.5	55.1	24.2	55.3	7.5
	Independent w/ Group Sparse Loss	77.7	52.0	87.5	65.3	86.0	97.9	34.0	59.3	14.6	34.6	5.4	56.7	33.1	34.6	48.9	**26.6**	**55.6**	7.2
	Independent w/ InfoNCE Loss	79.0	**54.9**	88.6	66.2	87.2	**98.0**	**40.2**	64.2	**18.2**	37.5	8.5	**67.6**	**45.0**	**35.9**	**58.1**	24.2	53.8	**7.6**
32	Shared	79.9	55.9	89.8	67.7	87.8	**98.1**	39.7	67.8	17.8	43.9	12.6	59.8	36.5	38.3	55.6	28.8	60.7	7.9
	Shared w/ Group Sparse Loss	77.8	51.0	88.2	65.6	86.3	97.8	28.3	57.7	11.8	20.0	8.3	60.7	11.3	36.0	50.6	28.6	59.3	5.6
	Shared w/ InfoNCE Loss	78.7	53.1	**90.1**	67.5	85.1	97.9	36.2	63.1	18.1	35.2	**13.9**	61.1	26.1	36.3	55.3	25.2	57.9	6.9
	Independent	81.3	61.6	89.1	**68.5**	89.5	97.9	**49.1**	71.0	**34.4**	**56.3**	7.3	69.0	**56.5**	**39.7**	58.2	**31.4**	59.9	**9.2**
	Independent w/ Group Sparse Loss	80.4	59.2	89.3	67.2	88.4	**98.1**	45.9	71.3	27.2	47.5	8.0	72.8	48.4	35.8	54.8	20.7	59.2	8.5
	Independent w/ InfoNCE Loss	**81.7**	**62.4**	89.1	68.4	**90.3**	**98.1**	48.5	**75.7**	27.5	49.1	8.7	**75.1**	54.8	39.2	**59.0**	27.9	**61.4**	8.4
64	Shared	82.6	63.3	90.3	69.0	92.0	**98.2**	55.1	78.4	37.3	60.3	17.5	78.4	58.6	42.2	60.4	34.1	**64.6**	9.7
	Shared w/ Group Sparse Loss	79.4	53.3	90.1	66.8	88.8	98.0	35.6	62.0	14.6	41.2	12.2	62.6	20.8	39.6	52.6	**39.4**	59.1	7.4
	Shared w/ InfoNCE Loss	82.5	61.2	90.6	69.8	**92.7**	98.1	50.5	79.4	30.3	54.7	19.0	74.2	45.1	41.6	58.8	36.2	61.9	9.5
	Independent	83.2	66.2	90.7	**70.2**	90.8	98.0	**59.2**	76.6	**41.2**	**71.8**	**19.9**	**85.1**	**60.7**	**42.5**	61.3	34.5	64.0	10.3
	Independent w/ Group Sparse Loss	81.2	60.4	90.5	67.4	89.3	**98.2**	45.4	65.0	20.5	60.5	3.8	78.2	44.2	38.2	51.6	31.8	61.1	8.5
	Independent w/ InfoNCE Loss	**83.4**	**67.1**	**91.1**	69.3	91.3	**98.2**	57.3	**80.6**	39.6	68.1	18.7	81.7	55.4	41.9	**62.1**	30.9	64.0	**10.6**
128	Shared	84.1	68.8	90.9	70.0	92.7	98.1	66.7	**84.7**	49.7	78.3	**28.3**	**90.6**	68.6	**45.2**	63.0	**40.0**	**65.4**	12.3
	Shared w/ Group Sparse Loss	81.9	61.2	90.8	67.1	92.5	98.1	52.5	77.7	29.7	63.8	10.0	87.9	45.7	40.9	58.4	34.0	62.0	9.3
	Shared w/ InfoNCE Loss	84.5	68.5	**91.6**	**70.7**	93.5	98.0	64.4	82.6	48.1	73.7	26.7	85.2	70.3	43.9	62.9	37.3	63.3	12.0
	Independent	**84.7**	70.4	91.3	70.1	**93.7**	98.1	**67.0**	83.2	**52.5**	**80.6**	27.3	**90.6**	67.9	44.2	**63.2**	37.4	63.6	**12.7**
	Independent w/ Group Sparse Loss	82.2	64.0	90.6	68.0	90.3	**98.2**	52.3	74.1	29.7	70.8	8.1	80.5	50.7	40.9	57.0	38.5	58.7	9.6
	Independent w/ InfoNCE Loss	**84.7**	**70.5**	91.2	70.6	93.2	98.1	66.7	84.3	51.0	78.3	26.5	88.4	**71.6**	44.2	**63.2**	37.6	63.7	12.3
256	Shared	**85.9**	73.7	91.6	71.2	**94.9**	98.1	**73.5**	**88.4**	57.9	**86.6**	36.3	**93.4**	78.5	45.8	63.6	42.0	63.6	13.9
	Shared w/ Group Sparse Loss	83.9	68.7	90.5	68.9	93.4	98.0	63.9	83.8	43.9	81.7	23.2	88.5	62.1	42.5	61.2	35.2	62.8	11.0
	Shared w/ InfoNCE Loss	85.5	73.4	91.5	70.6	93.9	**98.2**	72.6	87.2	**59.8**	85.2	34.1	91.9	77.3	**46.5**	63.6	**44.3**	**64.1**	14.1
	Independent	**85.9**	73.3	**91.9**	**71.5**	94.7	98.1	72.1	87.3	57.4	85.9	34.3	91.9	75.6	45.9	62.9	42.6	**64.1**	14.0
	Independent w/ Group Sparse Loss	83.8	68.1	91.0	69.3	92.5	98.1	61.4	77.4	41.2	82.7	22.3	86.0	58.8	42.9	57.7	41.4	61.5	10.9
	Independent w/ InfoNCE Loss	85.9	**73.8**	91.7	71.3	94.4	**98.2**	**73.5**	87.5	**59.8**	86.5	**36.4**	92.2	**78.7**	46.0	**64.0**	41.8	64.0	**14.3**
	SigLIP ViT-B/16	86.7	76.0	92.4	72.2	94.9	98.2	78.5	89.6	68.6	90.8	44.0	94.2	84.1	48.3	66.3	45.8	65.1	15.9

Table 8: Zero-shot retrieval performance across datasets, grouped by sparsity level K, for SigLIP ViT-B/16 on CC3M.

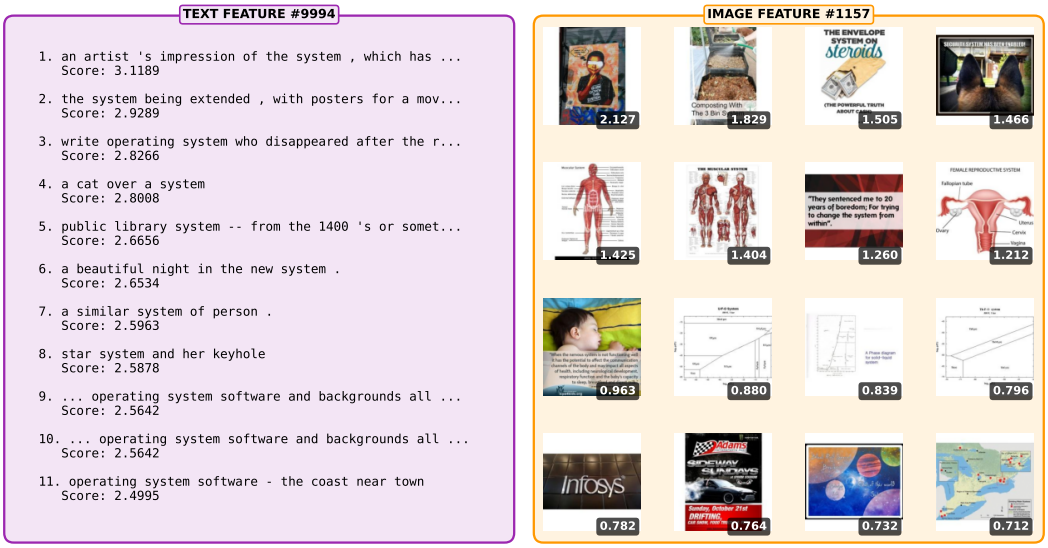
		Text Retrieval								Image Retrieval							
		Flickr30k				COCO				Flickr30k				COCO			
K	SAE	Avg	R@1	R@5	R@10	Avg	R@1	R@5	R@10	Avg	R@1	R@5	R@10	Avg	R@1	R@5	R@10
8	Shared	17.4	5.9	17.8	28.6	18.7	5.0	18.7	32.4	53.5	29.8	59.7	70.9	29.6	12.4	32.2	44.3
	Shared w/ Group Sparse Loss	43.7	20.1	48.3	62.6	30.4	12.2	33.1	45.9	40.3	15.8	45.4	59.6	21.8	8.0	23.3	34.0
	Shared w/ InfoNCE Loss	11.2	4.1	11.1	18.3	8.6	1.4	7.4	17.0	50.8	25.7	57.4	69.4	26.3	10.2	28.4	40.4
	Independent	63.8	43.1	69.6	78.8	**40.5**	**21.4**	**44.0**	**56.1**	57.8	35.1	63.7	74.6	36.6	18.4	39.9	51.6
	Independent w/ Group Sparse Loss	63.5	42.9	69.6	77.9	38.6	20.3	42.3	53.3	**59.8**	**37.6**	**66.0**	**75.7**	36.5	18.4	40.0	51.1
	Independent w/ InfoNCE Loss	**64.9**	**43.7**	**70.8**	**80.2**	39.8	20.9	43.3	55.3	58.6	35.9	64.6	75.2	**37.2**	**19.1**	**40.7**	**51.9**
16	Shared	41.0	19.4	45.1	58.6	32.8	10.2	34.9	53.2	69.1	47.0	75.9	84.5	44.1	23.5	48.4	60.5
	Shared w/ Group Sparse Loss	54.2	26.8	59.9	75.9	40.1	16.4	44.7	59.3	58.2	33.1	65.1	76.3	34.3	16.3	37.2	49.4
	Shared w/ InfoNCE Loss	7.2	2.7	7.4	11.6	4.2	0.3	3.9	8.3	64.5	41.2	71.6	80.6	39.5	20.0	43.5	55.0
	Independent	**80.9**	**63.2**	**86.6**	**92.9**	57.5	35.7	62.9	**74.0**	70.7	49.7	77.0	85.4	**48.0**	**27.6**	**52.4**	**63.9**
	Independent w/ Group Sparse Loss	77.3	59.1	83.4	89.3	53.9	32.5	58.9	70.2	70.8	49.7	77.3	85.4	46.0	25.9	50.2	61.8
	Independent w/ InfoNCE Loss	80.0	62.3	85.7	92.0	**57.9**	**36.7**	**63.6**	73.3	**72.1**	**52.2**	**78.3**	**85.9**	47.6	26.9	52.1	**63.9**
32	Shared	63.4	35.8	70.4	84.0	55.5	28.4	62.0	76.0	77.7	58.9	84.0	90.1	53.4	32.7	58.5	69.0
	Shared w/ Group Sparse Loss	51.4	21.5	57.3	75.5	45.6	20.4	50.7	65.7	71.6	50.4	78.2	86.2	45.1	24.5	49.3	61.6
	Shared w/ InfoNCE Loss	52.5	27.4	57.0	73.2	39.0	8.1	44.0	65.0	77.4	58.2	83.7	90.2	52.9	32.0	57.9	68.9
	Independent	**88.9**	**76.2**	**93.9**	**96.7**	67.5	47.3	73.1	82.1	**78.9**	**60.9**	**84.9**	**90.8**	54.9	34.3	60.1	70.4
	Independent w/ Group Sparse Loss	86.3	70.6	92.4	96.0	64.4	43.1	70.1	80.0	77.7	58.5	84.2	90.3	52.9	32.4	57.8	68.6
	Independent w/ InfoNCE Loss	88.4	75.1	93.4	**96.7**	**68.3**	**48.8**	**73.4**	**82.8**	78.5	60.1	84.7	90.7	**55.6**	**34.7**	**61.0**	**71.1**
64	Shared	83.4	62.8	91.2	96.2	69.3	47.2	75.7	85.0	81.9	65.0	87.9	92.8	58.6	38.0	63.9	73.9
	Shared w/ Group Sparse Loss	43.1	20.2	46.3	62.8	45.3	13.1	52.6	70.2	77.0	57.8	83.2	89.9	51.0	30.1	55.8	67.2
	Shared w/ InfoNCE Loss	83.5	62.6	91.5	96.5	67.8	45.1	74.4	83.9	82.3	66.0	88.0	92.9	58.9	38.0	64.2	74.4
	Independent	**92.1**	**81.9**	**95.9**	**98.6**	73.1	54.3	78.6	86.5	82.6	66.1	88.5	**93.3**	59.4	38.6	64.8	74.8
	Independent w/ Group Sparse Loss	86.9	70.7	93.0	96.9	65.4	44.1	71.5	80.6	79.6	61.8	85.6	91.3	54.5	33.5	59.6	70.5
	Independent w/ InfoNCE Loss	91.7	81.1	95.8	98.2	**73.4**	**54.8**	**78.7**	**86.8**	**83.0**	**66.9**	**89.0**	93.0	**60.1**	**39.5**	**65.6**	**75.3**
128	Shared	91.4	80.0	96.2	98.1	75.3	57.1	80.9	88.0	84.7	70.0	90.1	94.1	62.4	42.3	67.8	77.1
	Shared w/ Group Sparse Loss	52.0	26.4	56.5	73.2	54.7	24.2	62.4	77.6	80.3	62.3	86.7	91.9	54.9	34.0	59.8	70.8
	Shared w/ InfoNCE Loss	92.7	82.6	97.0	98.6	75.9	57.3	81.8	88.6	85.0	70.1	90.6	94.4	62.3	42.1	67.7	77.2
	Independent	**93.8**	**85.4**	97.2	98.7	76.7	59.7	82.0	88.4	85.0	70.1	90.7	94.3	62.6	42.4	68.0	77.4
	Independent w/ Group Sparse Loss	88.6	73.9	94.5	97.3	67.7	46.6	73.4	83.0	80.9	63.2	87.1	92.5	56.5	35.3	61.8	72.5
	Independent w/ InfoNCE Loss	93.7	84.7	**97.4**	**98.9**	**77.7**	**60.2**	**83.1**	**89.7**	**85.3**	**70.7**	**90.8**	**94.5**	**63.2**	**43.3**	**68.5**	**77.8**
256	Shared	93.9	85.2	97.6	99.0	79.4	62.6	84.9	**90.8**	86.3	72.5	91.2	**95.2**	64.9	**45.3**	70.2	79.1
	Shared w/ Group Sparse Loss	56.2	28.7	61.7	78.3	63.5	36.7	71.2	82.6	82.4	65.5	88.3	93.4	58.8	38.0	63.8	74.5
	Shared w/ InfoNCE Loss	94.5	86.7	97.9	99.0	78.9	61.7	84.5	90.6	86.2	72.3	91.5	94.9	64.6	44.8	70.1	79.0
	Independent	94.9	87.5	97.9	**99.3**	79.6	63.5	84.7	90.6	86.2	72.4	91.3	94.9	64.9	45.2	70.3	79.3
	Independent w/ Group Sparse Loss	89.5	75.4	95.1	98.0	69.6	49.2	75.5	84.2	82.8	66.8	88.3	93.3	59.1	38.1	64.6	74.7
	Independent w/ InfoNCE Loss	**95.0**	**87.9**	**98.0**	99.0	**79.8**	**63.7**	**85.0**	90.6	**86.5**	**72.6**	**91.6**	**95.2**	**65.1**	**45.3**	**70.6**	**79.4**
	SigLIP ViT-B/16	95.4	89.0	98.0	99.3	81.0	65.3	86.2	91.6	87.5	74.6	92.3	95.6	66.6	47.2	72.1	80.6

Table 9: Zero-shot classification accuracy (%) across datasets, grouped by sparsity level K, for SigLIP 2 ViT-B/16 on CC3M.

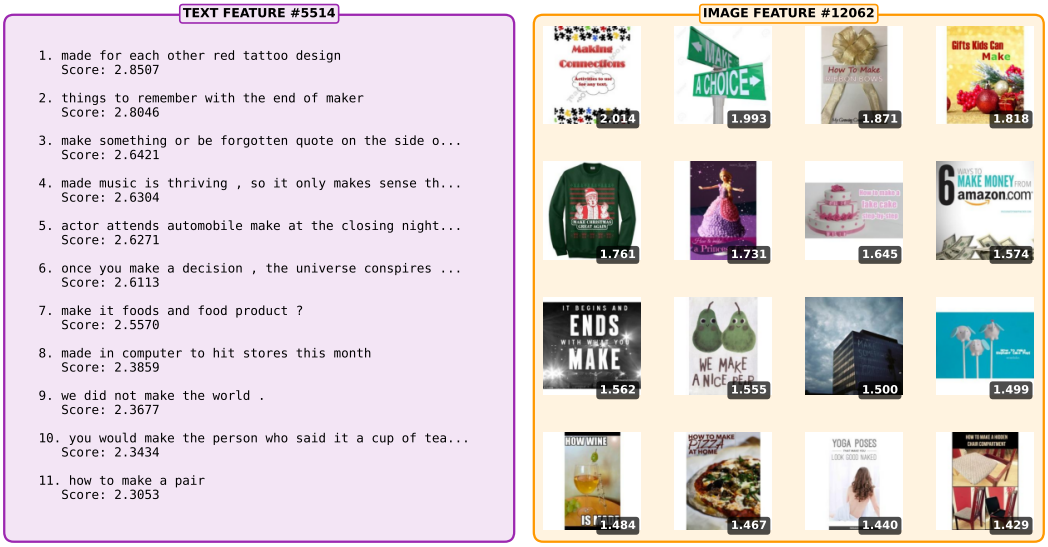
		General datasets						Fine-grained datasets							Misc. datasets				
K	SAE	Avg	ImageNet	CIFAR-10	CIFAR-100	Caltech-101	STL-10	Avg	Food-101	CUB	Cars	Aircraft	Pets	Flowers	Avg	DTD	EuroSAT	RESISC45	Country211
8	Shared	73.1	**41.1**	87.6	60.7	77.9	98.1	**18.9**	47.9	**3.0**	**2.3**	1.7	31.7	**27.0**	27.1	37.2	22.2	48.5	0.6
	Shared w/ Group Sparse Loss	72.6	38.1	89.4	59.7	77.9	98.0	17.2	39.7	2.4	1.4	2.3	35.3	21.9	28.9	38.1	25.4	51.6	0.6
	Shared w/ InfoNCE Loss	**73.2**	40.7	87.9	60.6	**78.5**	98.3	17.3	**49.3**	2.0	2.0	2.8	31.8	15.8	26.1	38.5	17.5	47.8	**0.8**
	Independent	72.7	38.3	90.2	62.5	74.2	**98.4**	17.4	46.4	1.6	1.2	2.4	35.1	17.8	**29.9**	**41.1**	**25.5**	**52.4**	0.7
	Independent w/ Group Sparse Loss	72.0	35.1	**90.8**	61.4	74.5	98.2	17.8	45.0	1.2	1.0	2.8	**43.8**	12.9	27.5	37.0	24.2	47.9	**0.8**
	Independent w/ InfoNCE Loss	73.0	38.3	89.4	**62.8**	76.0	**98.4**	17.9	42.4	1.3	1.3	**4.0**	39.9	18.8	25.7	37.6	15.0	49.5	0.7
16	Shared	78.6	50.9	91.2	65.2	87.1	98.6	26.9	65.3	3.0	2.2	2.7	53.6	34.6	32.9	51.9	20.9	58.1	0.7
	Shared w/ Group Sparse Loss	77.0	46.2	91.3	63.7	85.4	98.6	23.2	60.5	2.9	2.2	3.4	39.9	30.4	32.4	46.2	25.1	57.4	0.7
	Shared w/ InfoNCE Loss	78.2	49.7	89.7	64.5	**88.5**	98.6	24.6	58.3	2.7	2.4	1.7	54.8	27.8	31.2	45.6	23.5	54.7	**1.0**
	Independent	79.2	51.7	90.8	66.8	87.9	**98.7**	30.8	**68.1**	**3.3**	**2.9**	4.4	**68.1**	38.0	35.8	**52.3**	31.5	58.7	0.8
	Independent w/ Group Sparse Loss	76.7	44.6	91.4	65.4	83.8	98.4	24.1	52.4	2.7	2.0	2.9	58.4	26.1	33.0	46.5	27.8	56.7	0.8
	Independent w/ InfoNCE Loss	**79.7**	**52.9**	**91.5**	**67.3**	87.9	**98.7**	**31.1**	68.0	3.1	2.3	**6.0**	65.8	**41.3**	**36.2**	51.0	**34.0**	**59.0**	0.8
32	Shared	81.2	56.8	**92.3**	69.1	89.1	**98.9**	33.1	67.9	3.1	2.0	**7.0**	68.2	50.6	39.1	56.1	36.4	63.2	0.7
	Shared w/ Group Sparse Loss	75.6	43.8	91.0	63.2	81.2	98.7	19.6	44.1	1.7	1.8	1.5	41.8	26.7	34.7	41.8	37.8	58.5	0.7
	Shared w/ InfoNCE Loss	81.1	55.4	92.1	69.1	90.2	98.6	28.8	66.0	3.1	**3.7**	2.3	54.8	43.0	35.1	53.0	25.2	61.7	0.7
	Independent	**82.1**	58.5	92.1	**69.6**	91.4	98.8	35.1	73.4	3.1	2.4	6.1	72.6	52.8	40.1	**58.0**	**38.7**	62.7	0.9
	Independent w/ Group Sparse Loss	79.9	52.9	92.2	67.8	88.4	98.4	32.0	70.2	3.1	2.4	4.1	69.3	42.8	38.4	53.5	36.5	62.8	0.7
	Independent w/ InfoNCE Loss	82.1	**59.7**	90.6	68.9	**92.2**	**98.9**	**36.8**	**77.6**	**3.2**	3.4	4.9	**74.9**	**56.6**	**40.9**	56.1	38.1	**68.3**	**1.0**
64	Shared	83.3	61.3	92.7	71.1	92.6	**98.9**	40.2	80.5	3.5	3.7	**7.6**	79.1	**67.0**	40.9	58.5	39.1	65.3	0.8
	Shared w/ Group Sparse Loss	79.9	52.6	91.3	66.3	90.5	**98.9**	30.7	69.1	3.0	2.9	4.2	55.3	49.6	39.2	53.5	37.7	64.9	0.8
	Shared w/ InfoNCE Loss	83.1	60.8	92.7	70.7	92.6	98.8	37.5	78.5	3.8	**3.8**	5.3	72.7	61.1	39.8	58.1	34.8	65.2	**0.9**
	Independent	**83.9**	62.2	**93.3**	**71.3**	**93.9**	**98.9**	41.1	82.3	2.9	3.7	6.6	**87.2**	64.0	**41.5**	**59.3**	**39.8**	65.9	0.8
	Independent w/ Group Sparse Loss	79.7	52.5	91.7	66.7	89.1	98.7	29.0	65.4	**3.9**	2.9	1.6	71.9	28.2	38.4	50.1	36.7	66.0	0.8
	Independent w/ InfoNCE Loss	83.7	**63.1**	92.2	71.1	93.3	98.8	**41.2**	**83.3**	3.7	3.6	6.9	85.7	64.0	40.7	57.4	36.7	**67.8**	0.8
128	Shared	84.5	64.7	93.0	72.4	93.4	**98.8**	44.2	86.0	**4.5**	3.7	8.8	**89.5**	73.0	43.4	60.0	44.3	68.2	**0.9**
	Shared w/ Group Sparse Loss	82.0	58.8	92.4	67.4	92.7	**98.8**	39.2	78.0	3.1	2.9	6.4	81.5	63.1	41.0	56.6	37.1	**69.5**	**0.9**
	Shared w/ InfoNCE Loss	84.6	64.6	92.9	**73.0**	93.8	**98.8**	44.0	86.3	4.1	3.8	**9.4**	89.2	71.0	42.2	61.6	38.6	67.8	0.8
	Independent	84.6	65.0	**93.2**	72.7	93.2	**98.8**	44.1	85.9	**4.5**	3.6	7.8	89.1	73.8	**44.2**	62.1	**45.1**	68.8	0.8
	Independent w/ Group Sparse Loss	82.1	59.0	92.4	68.1	92.0	**98.8**	36.4	78.7	**4.5**	3.5	3.9	80.8	47.2	40.0	56.0	38.7	64.4	**0.9**
	Independent w/ InfoNCE Loss	**84.9**	**65.7**	**93.2**	72.9	**93.9**	98.7	**44.7**	**86.5**	4.0	**3.9**	8.4	89.3	**76.0**	43.8	**62.4**	44.1	67.7	**0.9**
256	Shared	**85.7**	**67.9**	93.4	73.4	**94.8**	**98.8**	46.8	**89.5**	4.6	3.8	10.9	91.8	**80.1**	44.5	62.8	45.5	68.9	0.9
	Shared w/ Group Sparse Loss	83.9	63.6	92.5	70.3	94.1	**98.8**	42.7	84.8	4.2	3.8	7.1	86.9	69.4	43.6	60.0	**46.4**	67.2	0.8
	Shared w/ InfoNCE Loss	85.4	67.6	93.2	72.9	94.7	**98.8**	46.8	88.6	4.9	3.6	**12.1**	92.2	79.6	**44.6**	62.9	45.9	68.8	1.0
	Independent	85.6	67.7	**93.6**	**73.6**	94.5	98.7	46.5	89.4	4.9	3.7	11.2	91.4	78.5	44.4	**63.3**	44.4	**69.2**	0.9
	Independent w/ Group Sparse Loss	84.2	64.2	92.8	71.6	93.8	**98.8**	41.9	84.1	4.4	3.8	6.8	84.7	67.4	41.9	57.3	42.9	66.5	**1.1**
	Independent w/ InfoNCE Loss	85.5	**67.9**	93.5	73.1	94.4	**98.8**	**47.0**	89.3	**5.2**	**4.0**	11.9	**92.8**	79.0	44.6	62.1	46.1	69.1	1.0
	SigLIP 2 ViT-B/16	86.2	70.0	93.6	73.8	95.0	98.8	48.9	90.9	5.8	4.1	13.6	94.8	84.2	45.6	64.6	47.1	69.6	1.0

Table 10: Zero-shot retrieval performance across datasets, grouped by sparsity level K, for SigLIP 2 ViT-B/16 on CC3M.

		Text Retrieval								Image Retrieval							
		Flickr30k				COCO				Flickr30k				COCO			
K	SAE	Avg	R@1	R@5	R@10	Avg	R@1	R@5	R@10	Avg	R@1	R@5	R@10	Avg	R@1	R@5	R@10
8	Shared	58.7	34.3	64.9	77.0	38.1	18.4	41.9	54.1	46.9	24.7	52.3	63.6	25.1	9.8	27.1	38.4
	Shared w/ Group Sparse Loss	56.9	34.2	62.3	74.1	35.9	17.5	39.4	50.8	37.7	18.6	41.1	53.3	20.6	8.3	22.0	31.5
	Shared w/ InfoNCE Loss	34.5	6.9	35.5	61.2	29.6	8.3	32.8	47.8	44.4	21.8	49.1	62.4	18.8	6.1	20.1	30.1
	Independent	61.7	38.5	**68.1**	78.4	**41.2**	**23.1**	**44.6**	**56.0**	57.7	35.4	63.5	74.1	**37.8**	**19.7**	**41.2**	**52.6**
	Independent w/ Group Sparse Loss	60.3	37.8	66.4	76.8	38.8	20.8	42.1	53.5	57.7	33.5	63.8	**75.7**	37.1	18.9	40.3	52.1
	Independent w/ InfoNCE Loss	**62.0**	**39.3**	67.1	**79.5**	40.6	22.6	44.0	55.2	**58.8**	**36.3**	**64.6**	75.6	37.5	19.5	40.9	52.1
16	Shared	75.6	50.5	84.9	91.5	55.9	33.7	61.2	72.9	70.8	49.9	77.0	85.4	46.0	25.2	50.6	62.1
	Shared w/ Group Sparse Loss	71.9	45.5	80.7	89.5	51.9	30.4	56.6	68.6	57.9	36.2	64.0	73.5	35.0	17.0	38.2	49.8
	Shared w/ InfoNCE Loss	13.1	1.3	8.9	29.2	24.5	3.0	24.3	46.1	62.4	39.2	68.5	79.5	36.8	17.7	40.3	52.5
	Independent	**80.2**	60.5	**87.0**	**93.2**	**58.8**	**37.6**	**63.8**	**74.9**	**71.6**	51.3	**77.8**	**85.6**	**49.3**	**28.9**	**54.0**	**65.1**
	Independent w/ Group Sparse Loss	76.9	57.4	82.3	90.9	54.0	32.9	59.2	69.9	70.2	48.7	77.0	85.0	47.6	27.5	52.2	63.2
	Independent w/ InfoNCE Loss	79.1	**60.7**	85.5	91.1	57.1	37.0	61.8	72.6	71.5	**51.5**	77.6	85.3	48.9	28.8	53.4	64.5
32	Shared	86.3	68.4	93.2	**97.3**	66.7	45.1	**72.9**	**82.2**	**78.9**	**60.6**	**85.0**	**91.1**	55.8	35.1	60.8	71.6
	Shared w/ Group Sparse Loss	77.8	55.5	85.2	92.8	56.8	34.9	62.4	73.0	64.2	42.3	70.5	79.7	39.1	20.5	42.5	54.2
	Shared w/ InfoNCE Loss	82.4	61.4	90.0	95.9	61.1	37.7	67.9	77.8	76.2	56.5	82.5	89.6	52.6	31.4	57.8	68.6
	Independent	**88.5**	**74.9**	**93.7**	97.0	67.0	46.7	72.7	81.5	78.2	60.2	84.3	90.2	55.9	35.1	61.1	71.5
	Independent w/ Group Sparse Loss	83.9	66.7	90.4	94.6	63.4	42.3	68.9	78.9	76.3	57.6	82.4	88.9	53.2	32.5	58.1	69.0
	Independent w/ InfoNCE Loss	88.2	74.4	93.1	97.1	**67.3**	**47.0**	**72.9**	82.1	78.2	59.9	84.3	90.3	**56.3**	**35.4**	**61.5**	**72.1**
64	Shared	91.4	79.8	95.7	**98.6**	72.7	53.4	78.5	**86.1**	**82.5**	66.5	**88.0**	**93.0**	60.5	40.0	65.9	75.5
	Shared w/ Group Sparse Loss	82.0	57.7	91.9	96.5	65.3	44.8	70.8	80.4	72.8	53.0	78.9	86.4	46.7	26.2	51.2	62.8
	Shared w/ InfoNCE Loss	90.1	76.2	95.8	98.3	69.8	50.0	75.4	84.0	81.3	64.5	87.3	92.2	59.2	38.3	64.5	74.7
	Independent	91.7	79.8	**97.0**	98.2	72.1	53.1	77.9	85.2	82.1	65.9	87.5	92.8	60.4	40.0	65.7	75.4
	Independent w/ Group Sparse Loss	85.9	69.5	91.9	96.4	63.5	42.8	69.0	78.6	76.6	58.2	82.4	89.1	52.8	32.1	57.8	68.4
	Independent w/ InfoNCE Loss	**92.1**	**81.1**	96.7	98.5	**73.3**	**54.9**	**79.0**	85.9	**82.5**	**67.3**	87.5	92.7	**60.8**	**40.3**	**66.1**	**75.9**
128	Shared	94.0	85.6	97.6	98.8	75.9	58.0	81.2	88.6	84.6	**70.6**	89.4	93.7	63.7	43.7	69.1	78.3
	Shared w/ Group Sparse Loss	75.9	45.0	87.6	95.0	64.0	38.7	72.1	81.2	75.4	55.7	81.9	88.7	51.1	30.6	55.6	67.0
	Shared w/ InfoNCE Loss	93.6	84.6	97.3	99.0	76.2	58.4	81.9	88.2	84.6	69.7	**90.0**	**94.0**	63.3	42.9	68.9	78.0
	Independent	93.9	85.4	97.7	98.6	75.7	58.0	80.9	88.2	84.1	69.1	89.4	93.9	63.1	43.3	68.4	77.7
	Independent w/ Group Sparse Loss	88.4	73.3	94.2	97.7	66.2	44.8	72.3	81.4	77.3	58.5	83.8	89.6	54.6	33.6	59.8	70.3
	Independent w/ InfoNCE Loss	**94.5**	**85.9**	**98.2**	**99.3**	**76.9**	**59.4**	**82.3**	**89.1**	**84.8**	**70.6**	89.7	**94.0**	**63.8**	**43.9**	**69.2**	**78.4**
256	Shared	95.1	88.3	98.1	99.0	79.0	62.4	84.1	90.6	85.0	71.0	90.0	94.0	**65.0**	**45.5**	**70.2**	**79.2**
	Shared w/ Group Sparse Loss	84.9	63.4	94.4	97.0	69.9	47.6	76.7	85.4	77.2	59.9	82.7	89.0	55.1	35.0	60.2	70.2
	Shared w/ InfoNCE Loss	**95.7**	**89.7**	**98.2**	**99.2**	78.8	62.1	**84.4**	90.0	**85.5**	**71.7**	**90.5**	**94.4**	64.6	44.8	69.9	79.1
	Independent	95.0	87.8	98.1	99.1	79.1	62.3	84.2	**90.8**	84.9	71.0	89.8	93.9	64.3	44.8	69.7	78.5
	Independent w/ Group Sparse Loss	91.5	79.8	96.2	98.4	72.1	53.2	77.9	85.1	79.9	62.7	85.7	91.2	58.7	37.8	64.3	74.0
	Independent w/ InfoNCE Loss	94.8	87.8	97.9	98.8	**79.2**	**62.9**	84.1	90.6	84.9	71.1	89.7	93.8	64.7	45.1	70.1	79.0
	SigLIP 2 ViT-B/16	95.7	89.6	98.3	99.3	80.3	64.3	85.6	91.1	86.7	73.8	91.3	95.0	67.1	48.0	72.4	80.8



(a) Feature “system”

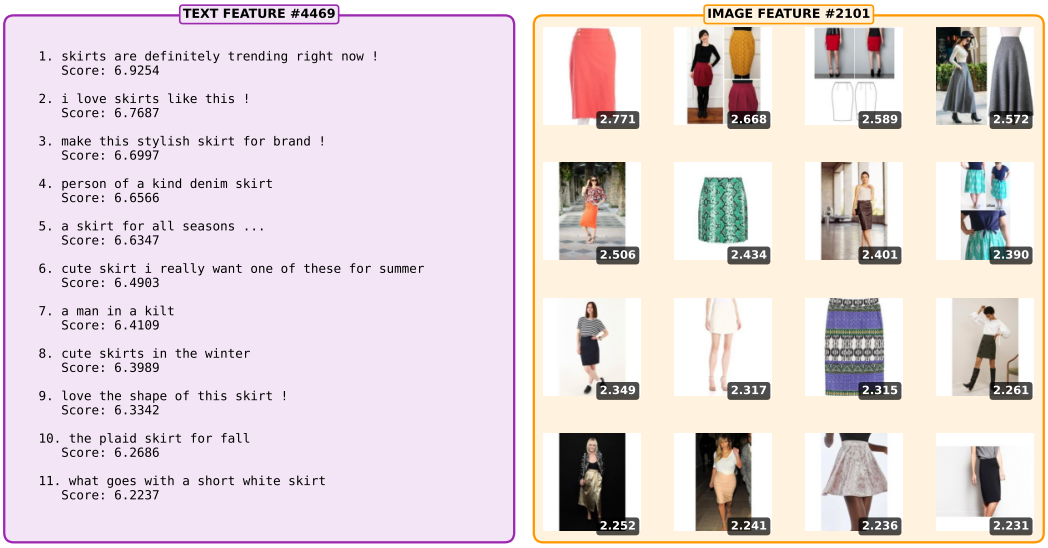


(b) Feature “make”

Figure 14: Top-activating examples for both modalities for the features (a) “system” and (b) “make” from independent SAE ($K = 32$) with InfoNCE Loss for SigLIP 2 ViT-B/16 on CC3M. For each sample, we report the activation value for the corresponding feature. Note that the concepts “system” and “make” appear as text within the images (zoom for detail).

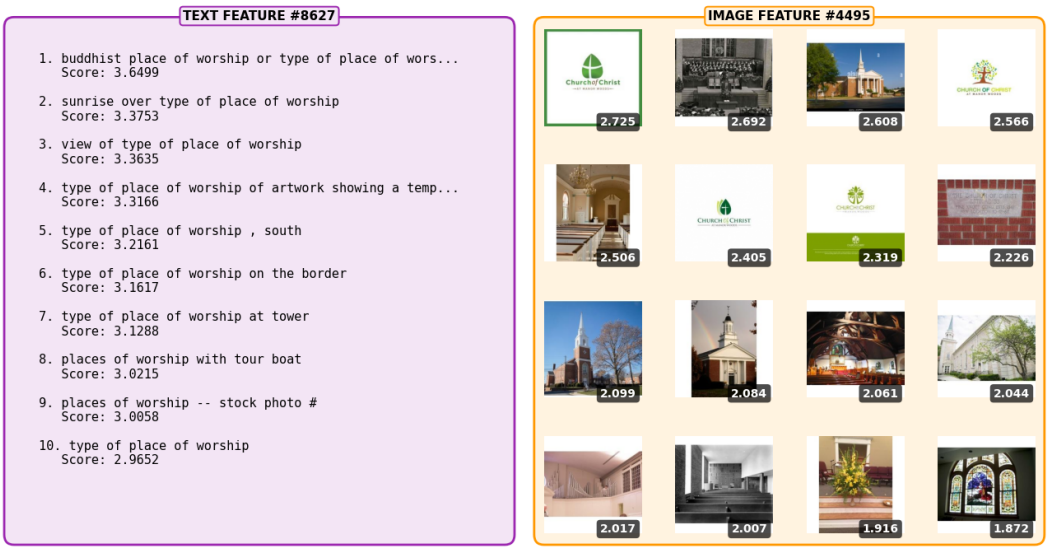


(a) Feature "Thailand"



(b) Feature "skirt"

Figure 15: Top-activating examples for both modalities for the features (a) "Thailand" and (b) "skirt" from independent SAE ($K = 32$) with InfoNCE Loss for SigLIP 2 ViT-B/16 on CC3M. For each sample, we report the activation value for the corresponding feature.



(a) Independent

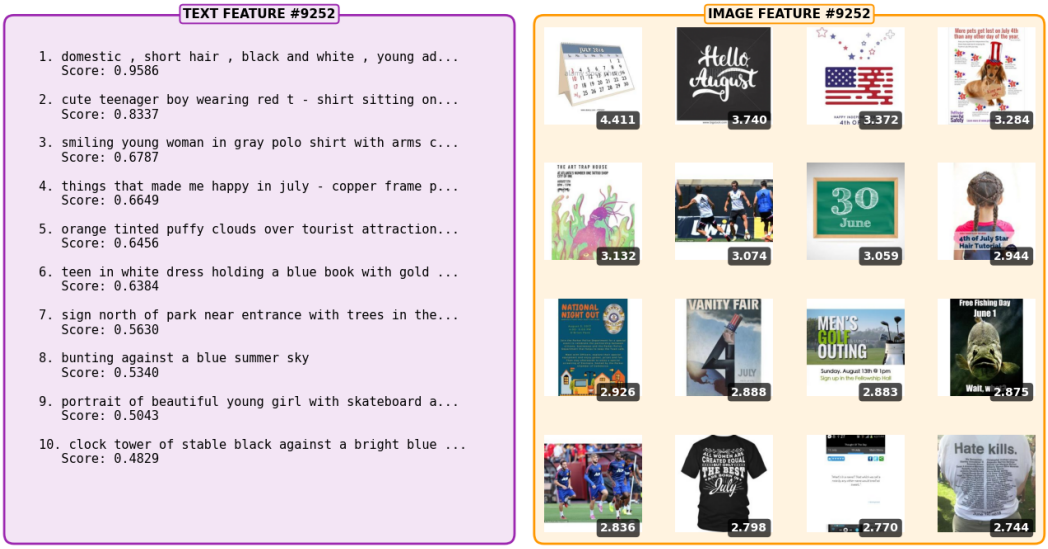


(b) Shared

Figure 16: Top-activating examples for both modalities for the "worship" feature for (a) Independent and (b) Shared SAE trained with InfoNCE loss on CLIP ViT-L/14 and CC3M. For each sample, we report the activation value for the corresponding feature.

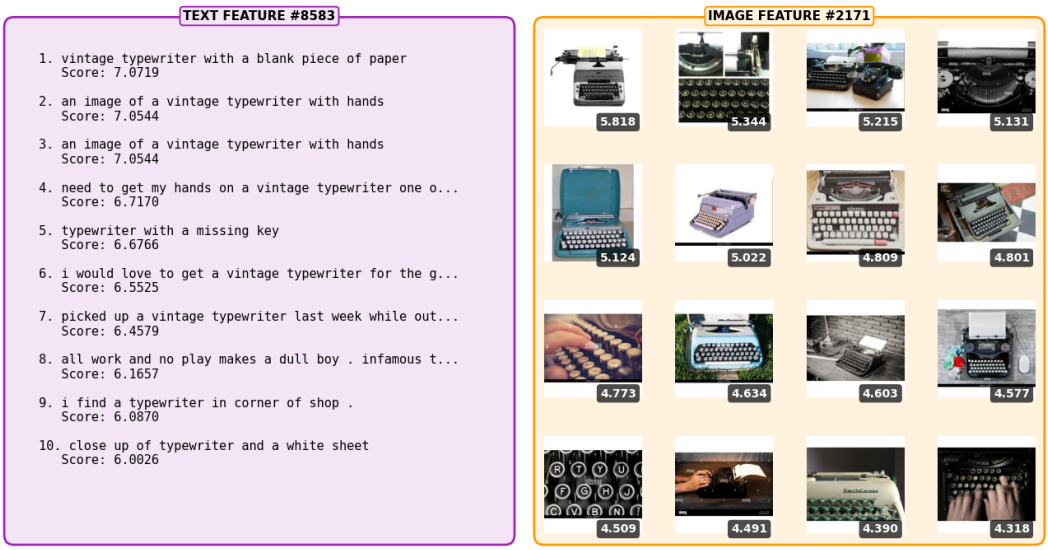


(a) Independent

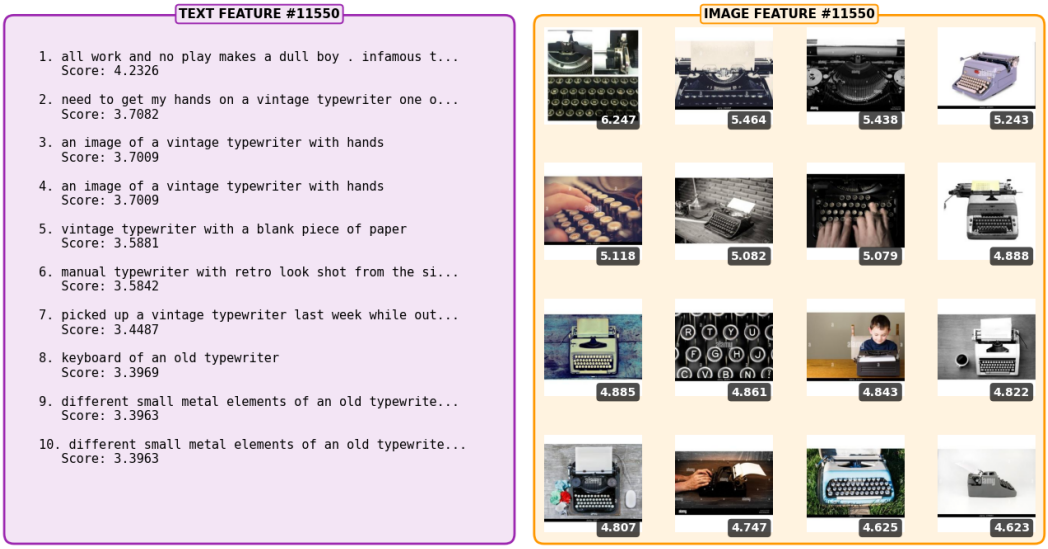


(b) Shared

Figure 17: Top-activating examples for both modalities for the "summer" feature for (a) Independent and (b) Shared SAE trained with InfoNCE loss on CLIP ViT-L/14 and CC3M. For each sample, we report the activation value for the corresponding feature.

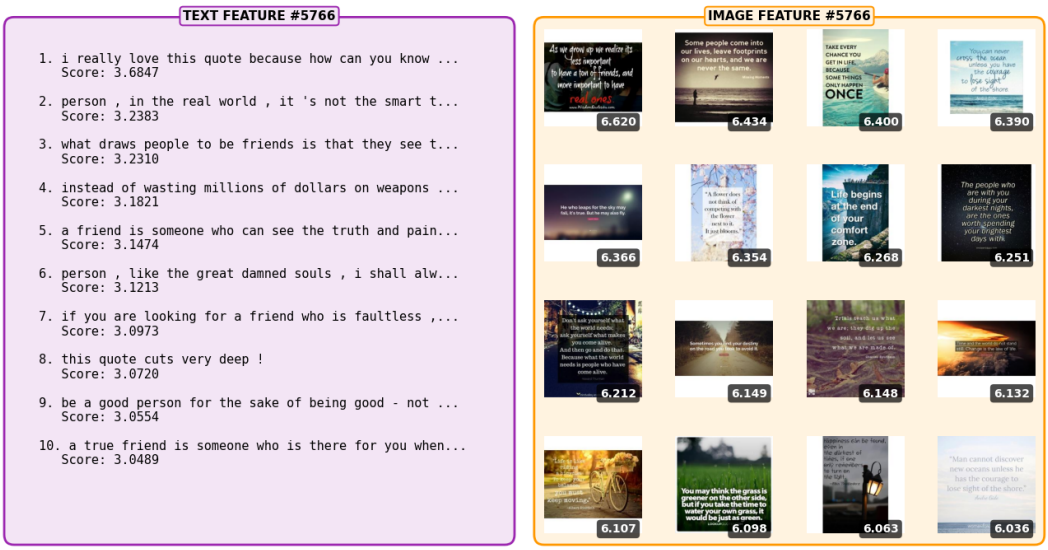


(a) Independent

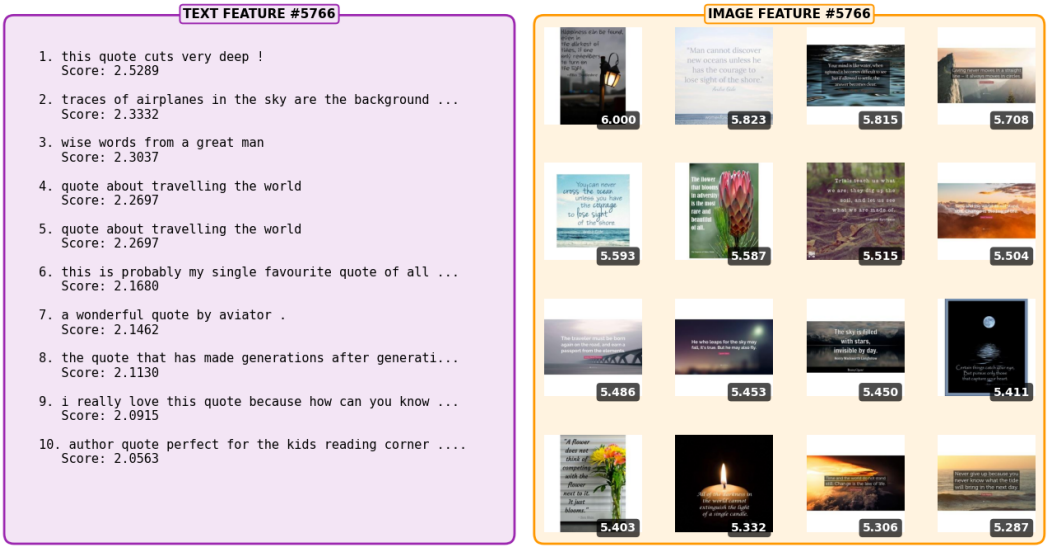


(b) Shared

Figure 18: Top-activating examples for both modalities for the "typewriter" feature for (a) Independent and (b) Shared SAE trained with InfoNCE loss on CLIP ViT-L/14 and CC3M. For each sample, we report the activation value for the corresponding feature.

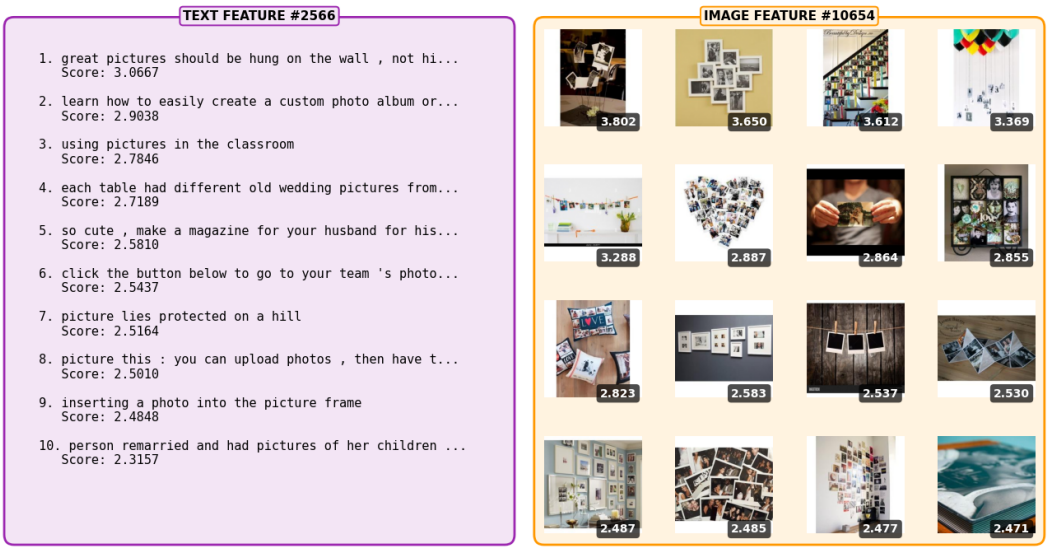


(a) Independent



(b) Shared

Figure 19: Top-activating examples for both modalities for the "quote" feature for (a) Independent and (b) Shared SAE trained with InfoNCE loss on CLIP ViT-L/14 and CC3M. For each sample, we report the activation value for the corresponding feature.



(a) Independent



(b) Shared

Figure 20: Top-activating examples for both modalities for the "pictures/photos" feature for (a) Independent and (b) Shared SAE trained with InfoNCE loss on CLIP ViT-L/14 and CC3M. For each sample, we report the activation value for the corresponding feature.

F.2.1 VARYING THE DICTIONARY SIZE

We further investigate if Shared SAEs can achieve similar performance to Independent SAEs by varying the dictionary size k. We train SAEs with $k \in \{d \times 2^x : x = 2, 3, \ldots, 6\}$, where d is the dimension of the VLM output space, by fixing the sparsity to $K = 32$. The reconstruction and interpretability results are shown in Tabs. 11 and 12, respectively. As observed, there is always a gap between the Shared and Independent SAEs reconstruction performance, and the gap is even more pronounced for the interpretability evaluation.

Table 11: Reconstruction performance for Shared and Independent SAEs trained on CC3M with CLIP ViT-L/14 ($K = 32$) by varying the dictionary size.

	Text		Image	
Dictionary Size	Shared SAE	Independent SAE	Shared SAE	Independent SAE
3072	0.801	**0.832**	0.761	**0.779**
6144	0.825	**0.851**	0.782	**0.794**
12288	0.841	**0.863**	0.793	**0.801**
24576	0.855	**0.871**	0.801	**0.804**
49152	0.860	**0.873**	0.800	**0.801**

Table 12: Interpretability evaluation for Shared and Independent SAEs trained on CC3M with CLIP ViT-L/14 ($K = 32$) by varying the dictionary size.

		Text		Image		Multimodal	
Dictionary Size	Metric	Shared SAE	Independent SAE	Shared SAE	Independent SAE	Shared SAE	Independent SAE
3072	Dead Neurons (%)	0.39	**0.00**	5.89	**0.00**	6.28	**0.00**
	Avg. MS Score	0.219	**0.241**	0.109	**0.132**	0.119	**0.128**
6144	Dead Neurons (%)	1.51	**0.00**	9.81	**0.00**	11.33	**0.00**
	Avg. MS Score	0.234	**0.276**	0.128	**0.150**	0.115	**0.129**
12288	Dead Neurons (%)	5.13	**0.00**	12.03	**0.00**	17.15	**0.00**
	Avg. MS Score	0.234	**0.294**	0.136	**0.150**	0.110	**0.125**
24576	Dead Neurons (%)	10.25	**0.11**	14.02	**0.01**	24.26	**0.13**
	Avg. MS Score	0.222	**0.289**	0.132	**0.142**	0.104	**0.119**
49152	Dead Neurons (%)	15.39	**0.73**	13.14	**0.13**	28.50	**0.85**
	Avg. MS Score	0.209	**0.252**	0.120	**0.122**	0.099	**0.109**

F.2.2 HALVING THE NUMBER OF FEATURES FOR EACH MODALITY IN THE INDEPENDENT SAE VARIANTS

To determine whether the superior performance of independent SAE variants stems from their increased parameter count or the modality-specific nature of their dictionaries, we conduct an experiment where we halve the number of features per modality in the independent variants ($\frac{k}{2}$), leading both architectures to have the same number of parameters. Despite the shared SAE variants having the advantage of encoding more features per modality in this configuration, the independent SAEs still achieve better results for CLIP ViT-L/14 in Figs. 21 and 22, and Tabs. 13 and 14, albeit with a reduced margin (compare with Figs. 6a and 11a, and Tabs. 5 and 6). These findings indicate that modality-specific dictionaries are the primary driving factor of feature quality, rather than parameter count alone.

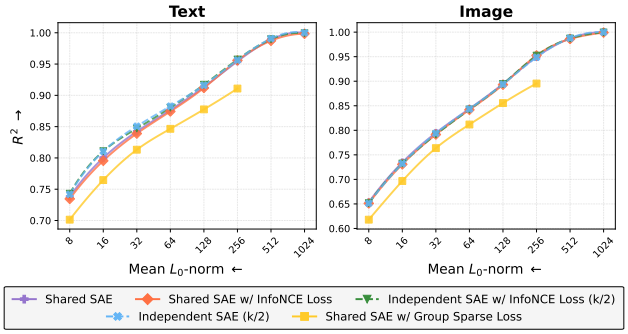


Figure 21: Reconstruction performance on CC3M for CLIP ViT-L/14. The plot shows the R^2 score of the reconstructed activations against the original activations for both modalities across different sparsity levels. The x-axis represents the mean L_0-norm of the codes (log scale). Here, independent SAE variants represent each modality with a number of features that is half of the one of the shared SAE variants.

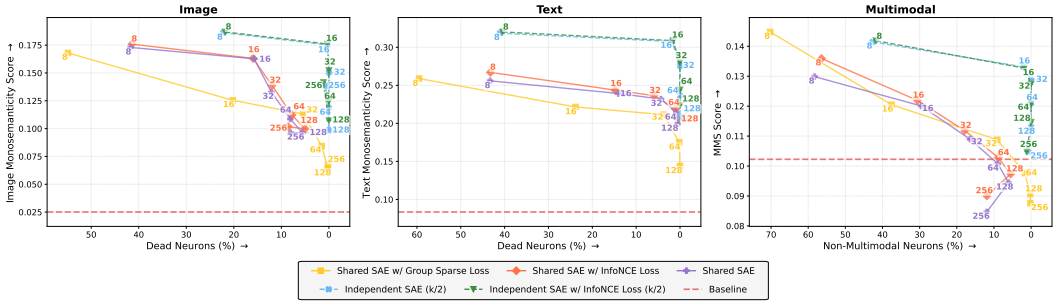


Figure 22: Interpretability evaluation for CC3M for CLIP ViT-L/14. The plot shows (from left to right) Image MonoSemanticity (MS), Text MS, and Multimodal MonoSemanticity (MMS) scores. Each point corresponds to the mean score across all features for a given sparsity level, highlighted in correspondence of each point. The x-axis indicates the percentage of dead neurons, *i.e.*, neurons that are never activated across the evaluation set. For Multimodal MonoSemanticity scores, we indicate the percentage of non-multimodal neurons, *i.e.*, neurons that are never activated across both modalities in the evaluation set. The percentage values in each plot are relative to the total number of features in each SAE variant (half of the shared SAE's feature count for each modality for the independent SAE variants). The dashed horizontal line indicates the mean score across all features for the dense VLM model.

Table 13: Zero-shot classification accuracy (%) across datasets, grouped by sparsity level K, for CLIP ViT-L/14 on CC3M. Here, independent SAE variants represent each modality with a number of features that is half of the one of the shared SAE variants.

		General datasets						Fine-grained datasets							Misc. datasets				
K	SAE	Avg	ImageNet	CIFAR-10	CIFAR-100	Caltech-101	STL-10	Avg	Food-101	CUB	Cars	Aircraft	Pets	Flowers	Avg	DTD	EuroSAT	RESISC45	Country211
8	Shared	56.9	3.6	90.7	52.5	38.5	99.0	6.9	19.1	1.1	4.2	1.7	12.1	3.0	28.4	20.5	42.6	46.7	3.8
	Shared w/ Group Sparse Loss	59.9	8.1	87.8	46.2	58.5	98.9	8.0	25.2	1.5	4.3	1.2	8.9	7.0	27.7	21.6	34.2	50.2	4.8
	Shared w/ InfoNCE Loss	52.2	0.9	89.0	36.3	35.9	98.8	12.4	27.6	1.6	7.0	1.8	29.1	7.5	22.7	18.5	28.1	39.2	4.8
	Independent (k/2)	75.8	42.4	**92.4**	65.0	79.8	**99.2**	**23.4**	**50.0**	7.3	**18.1**	**9.3**	**34.4**	21.3	**36.5**	**37.4**	42.7	**55.1**	**10.9**
	Independent w/ InfoNCE Loss (k/2)	**76.3**	**42.7**	91.8	**65.9**	**81.8**	99.1	22.0	49.8	**8.4**	12.2	8.3	30.8	**22.3**	35.8	36.0	**43.7**	53.0	10.5
16	Shared	75.9	46.2	92.5	61.3	80.2	99.2	24.5	57.5	9.2	16.9	7.6	33.9	21.8	40.2	43.1	**50.9**	55.2	11.5
	Shared w/ Group Sparse Loss	66.9	17.5	92.4	48.6	77.2	99.0	18.4	48.1	7.9	9.6	3.6	29.8	11.2	32.5	32.7	41.0	48.3	8.2
	Shared w/ InfoNCE Loss	67.3	11.0	**93.6**	52.5	80.5	99.0	19.3	51.2	4.4	10.9	7.4	36.1	5.6	33.8	37.3	40.7	51.8	5.5
	Independent (k/2)	**80.5**	53.0	93.4	**69.8**	**87.2**	99.2	**37.2**	67.8	**15.6**	30.3	**14.6**	**58.9**	36.1	41.7	**44.5**	45.4	60.8	**16.0**
	Independent w/ InfoNCE Loss (k/2)	79.8	**53.1**	93.3	66.7	86.7	**99.3**	36.9	**67.9**	14.8	**32.3**	13.6	52.0	**40.7**	**42.0**	42.6	48.0	**61.6**	15.7
32	Shared	80.7	56.8	94.1	67.2	86.3	99.3	39.9	72.7	15.3	32.1	14.7	59.3	45.6	46.0	**49.6**	48.9	**68.0**	17.7
	Shared w/ Group Sparse Loss	74.9	41.8	**94.3**	56.4	82.8	**99.4**	24.5	59.3	8.0	16.4	9.9	30.6	23.1	35.5	36.5	52.3	41.8	11.5
	Shared w/ InfoNCE Loss	81.2	56.6	94.1	69.4	86.7	99.3	38.0	73.7	19.4	29.1	11.3	58.7	35.6	45.4	47.7	53.1	65.5	15.3
	Independent (k/2)	82.9	**60.6**	93.9	71.5	89.1	**99.4**	45.1	76.6	19.2	**44.1**	13.0	69.3	48.6	46.2	47.0	**53.9**	64.9	**19.2**
	Independent w/ InfoNCE Loss (k/2)	**83.1**	60.5	**94.3**	**72.0**	**89.2**	99.3	**47.8**	**78.7**	**20.5**	42.4	**18.0**	**74.7**	**52.6**	**46.4**	46.9	53.4	66.5	18.9
64	Shared	84.2	64.6	93.9	72.3	90.6	**99.4**	54.9	83.6	30.9	52.2	18.8	82.8	61.1	50.3	**50.7**	59.4	68.6	22.6
	Shared w/ Group Sparse Loss	79.3	51.6	94.1	65.0	86.7	99.2	34.7	70.0	20.3	30.5	17.2	47.7	22.7	39.9	39.9	51.3	54.5	14.0
	Shared w/ InfoNCE Loss	84.2	64.8	93.7	72.4	90.7	99.3	54.8	82.8	**36.5**	53.7	17.7	78.4	59.8	50.2	**50.7**	60.3	68.2	21.5
	Independent (k/2)	85.1	66.2	**94.5**	75.0	90.6	**99.4**	**57.8**	**85.3**	34.7	**59.6**	**20.6**	**83.9**	**62.4**	50.0	48.0	58.6	**70.6**	**22.8**
	Independent w/ InfoNCE Loss (k/2)	**85.5**	**66.8**	94.4	**75.5**	**91.7**	99.3	57.0	85.2	35.6	58.1	20.5	82.1	60.7	**50.8**	49.2	**61.2**	**70.6**	22.2
128	Shared	86.3	70.2	**95.2**	74.4	**92.3**	99.2	63.2	87.9	49.7	59.8	23.7	**91.2**	67.1	51.6	49.8	60.7	69.9	25.8
	Shared w/ Group Sparse Loss	83.3	62.2	95.0	69.3	90.7	99.3	51.1	80.9	36.3	50.0	17.0	80.8	41.3	48.8	47.0	62.0	65.7	20.3
	Shared w/ InfoNCE Loss	86.4	**71.2**	94.6	75.0	92.1	99.3	64.8	**88.8**	50.8	66.2	23.6	89.6	69.6	51.2	51.8	56.3	70.8	25.9
	Independent (k/2)	**86.5**	70.6	94.6	76.1	92.0	**99.4**	**65.5**	88.2	**52.9**	**66.4**	**26.2**	89.2	**69.9**	**53.4**	**53.7**	**63.3**	70.8	25.8
	Independent w/ InfoNCE Loss (k/2)	86.5	70.5	94.8	**76.4**	91.5	**99.4**	64.2	**88.8**	49.3	66.1	23.7	88.5	69.1	52.1	52.0	58.8	**71.0**	**26.4**
256	Shared	87.3	73.0	95.3	75.4	**93.4**	**99.4**	69.3	**90.8**	57.7	71.9	**29.9**	91.4	74.0	53.4	52.7	60.7	**71.4**	28.7
	Shared w/ Group Sparse Loss	85.8	68.2	**95.4**	73.9	92.4	99.3	59.4	86.1	42.9	57.6	23.6	82.8	63.2	50.9	49.0	61.8	69.4	23.5
	Shared w/ InfoNCE Loss	87.5	73.4	95.3	76.5	93.1	99.3	68.7	90.1	56.8	71.6	28.0	90.6	**75.3**	54.5	**53.7**	64.4	71.0	28.8
	Independent (k/2)	87.7	**73.5**	**95.4**	76.8	93.2	**99.4**	69.1	90.5	**58.3**	72.4	28.1	91.4	73.9	54.4	52.2	65.4	70.9	28.9
	Independent w/ InfoNCE Loss (k/2)	**87.8**	**73.5**	95.2	**77.3**	**93.4**	**99.4**	**69.7**	90.4	58.1	**73.5**	29.0	**92.2**	74.8	**54.6**	52.8	**65.6**	71.2	**29.0**
	CLIP ViT-L/14	88.4	75.2	95.6	78.3	93.3	99.4	72.7	91.6	62.5	77.9	32.1	93.2	78.9	55.5	54.5	63.9	71.7	31.7

Table 14: Zero-shot retrieval performance across datasets, grouped by sparsity level K, for CLIP ViT-L/14 on CC3M. Here, independent SAE variants represent each modality with a number of features that is half of the one of the shared SAE variants.

		Text Retrieval								Image Retrieval							
		Flickr30k				COCO				Flickr30k				COCO			
K	SAE	Avg	R@1	R@5	R@10	Avg	R@1	R@5	R@10	Avg	R@1	R@5	R@10	Avg	R@1	R@5	R@10
8	Shared	7.3	3.3	7.1	11.6	9.3	2.5	9.3	16.2	55.2	32.4	61.3	71.9	30.5	14.3	33.0	44.2
	Shared w/ Group Sparse Loss	40.1	18.8	44.4	57.2	21.6	8.1	23.1	33.7	50.9	27.8	56.3	68.7	24.7	10.5	26.5	37.1
	Shared w/ InfoNCE Loss	6.7	1.8	5.8	12.5	13.0	3.6	12.6	22.7	54.3	31.4	60.4	71.1	28.8	13.4	31.1	41.9
	Independent (k/2)	**68.5**	45.6	**75.6**	**84.2**	**42.6**	**24.3**	**46.4**	**57.1**	**58.4**	**36.8**	**64.3**	**74.0**	**32.8**	**15.9**	**35.3**	**47.2**
	Independent w/ InfoNCE Loss (k/2)	67.2	**45.9**	73.2	82.5	42.1	24.1	45.4	56.8	57.3	34.4	63.4	**74.0**	32.3	15.5	34.9	46.6
16	Shared	69.5	43.3	78.1	87.2	49.9	28.0	55.0	66.7	64.1	42.9	70.2	79.2	37.8	19.7	41.1	52.7
	Shared w/ Group Sparse Loss	62.3	35.8	69.3	81.7	40.6	20.6	44.5	56.7	62.5	39.7	68.8	79.1	34.9	17.6	37.8	49.2
	Shared w/ InfoNCE Loss	37.9	12.2	40.5	61.0	39.5	16.5	43.6	58.3	63.5	41.4	69.9	79.1	37.5	19.5	40.7	52.2
	Independent (k/2)	78.2	58.9	**84.3**	**91.5**	54.4	33.8	59.1	70.4	64.8	42.6	71.4	**80.4**	**39.8**	**21.5**	**43.3**	**54.7**
	Independent w/ InfoNCE Loss (k/2)	**78.4**	**59.6**	**84.3**	91.4	**54.9**	**33.9**	**59.7**	**71.0**	**65.5**	**44.1**	**72.1**	**80.4**	39.4	20.8	43.0	54.3
32	Shared	82.1	62.2	89.4	94.8	60.1	39.3	65.4	75.5	70.2	49.7	76.4	84.4	44.1	24.8	48.1	59.5
	Shared w/ Group Sparse Loss	68.3	42.0	75.6	87.3	50.5	27.6	55.9	68.1	69.6	48.0	76.3	84.5	41.6	22.3	45.3	57.1
	Shared w/ InfoNCE Loss	82.4	63.7	88.7	94.9	59.4	38.4	64.5	75.4	69.3	48.4	75.4	84.2	44.4	25.1	48.4	59.7
	Independent (k/2)	86.1	70.1	92.3	95.8	62.0	41.9	67.1	77.0	70.7	49.9	77.4	84.9	44.6	25.1	48.6	60.1
	Independent w/ InfoNCE Loss (k/2)	**87.1**	**71.6**	**92.9**	**96.7**	**63.3**	**43.7**	**67.9**	**78.2**	**71.0**	**50.4**	**77.5**	**85.2**	**45.2**	**25.8**	**49.3**	**60.6**
64	Shared	88.3	74.2	93.7	96.9	66.2	46.2	71.3	81.1	75.3	56.0	81.5	88.5	49.1	29.0	53.8	64.6
	Shared w/ Group Sparse Loss	64.2	37.7	70.8	84.1	58.0	34.4	64.6	74.9	74.0	53.4	80.4	88.1	46.3	26.4	50.6	62.0
	Shared w/ InfoNCE Loss	89.0	75.0	94.5	97.4	66.6	46.5	72.1	81.3	75.6	55.8	82.2	88.7	49.5	29.4	53.9	65.1
	Independent (k/2)	90.3	77.6	**95.2**	98.0	67.2	47.1	72.7	81.9	75.5	56.6	81.5	88.3	48.8	28.9	53.1	64.5
	Independent w/ InfoNCE Loss (k/2)	**91.2**	**79.6**	**95.2**	**98.7**	**68.6**	**48.7**	**74.5**	**82.6**	**76.1**	**56.8**	**82.3**	**89.2**	**49.7**	**29.5**	**54.3**	**65.4**
128	Shared	90.7	77.8	95.7	98.6	69.8	50.3	75.1	84.0	78.6	60.2	84.7	90.9	52.2	32.0	56.8	67.7
	Shared w/ Group Sparse Loss	69.5	43.2	76.8	88.5	60.7	36.3	67.6	78.1	76.0	56.4	82.6	89.0	49.2	29.3	53.7	64.7
	Shared w/ InfoNCE Loss	91.8	80.7	96.2	98.4	70.4	51.4	75.7	84.0	**79.5**	**61.4**	**85.6**	**91.5**	**52.8**	**32.5**	**57.7**	**68.3**
	Independent (k/2)	**92.4**	**82.4**	**96.4**	98.3	70.4	51.0	75.9	84.3	78.4	59.8	84.5	90.8	51.7	31.5	56.5	67.2
	Independent w/ InfoNCE Loss (k/2)	92.0	81.3	96.0	**98.7**	**70.9**	**51.9**	**76.5**	**84.4**	78.8	60.4	85.1	90.9	52.4	32.2	57.2	67.9
256	Shared	92.5	82.6	96.1	98.7	72.4	53.2	78.3	85.7	80.2	62.6	86.3	91.7	54.0	34.1	58.6	69.4
	Shared w/ Group Sparse Loss	85.4	67.4	92.1	96.7	66.4	45.6	72.3	81.2	77.5	59.1	83.5	90.0	50.9	31.1	55.3	66.4
	Shared w/ InfoNCE Loss	92.4	82.5	96.3	98.4	73.0	**54.7**	78.3	86.1	80.5	62.8	86.5	92.1	54.4	34.4	59.2	**69.6**
	Independent (k/2)	93.3	83.9	97.1	**99.0**	72.4	53.7	78.0	85.6	**81.2**	**64.4**	**87.0**	92.1	54.2	34.2	59.0	69.5
	Independent w/ InfoNCE Loss (k/2)	**93.4**	**84.0**	**97.3**	**99.0**	**73.1**	54.4	**78.5**	**86.3**	80.6	63.0	86.5	**92.3**	**54.5**	**34.6**	**59.3**	**69.6**
	CLIP ViT-L/14	94.0	85.3	97.4	99.3	74.2	56.1	79.6	86.9	81.3	64.7	87.1	92.2	55.2	35.5	59.9	70.3



Figure 23: FashionIQ example of candidate and target images, with the corresponding relative caption.

F.2.3 Cross-modal steering

To demonstrate the intervention capabilities of the learned features, we perform cross-modal steering on the FashionIQ (Wu et al., 2021) dataset, following the protocol from Dhimoïla et al. (2026) to transfer multi-modal information from one modality to another. The objective is to retrieve target images that match the composition of the candidate images and their respective *relative* captions (see Fig. 23 for an example).

Let $\boldsymbol{\delta} \in \{0,1\}^k$ be a binary mask indicating the bimodal features, *i.e.*, features whose modality score is in $[0.05, 0.95]$. Let $\mathbf{z}^V_{\text{candidate}} \in \mathbb{R}^k$ and $\mathbf{z}^T_{\text{caption}} \in \mathbb{R}^k$ be the visual and textual codes of the candidate image and its relative caption, respectively. Vector arithmetic restricted to bimodal features allows to transfer information from one modality to another, by disregarding modality-specific features and remaining close to one modality distribution. The modified candidate $\widehat{\mathbf{a}}_{\text{steered}} \in \mathbb{R}^d$ is computed as

$$\widehat{\mathbf{a}}_{\text{steered}} = \mathbf{D}^T \left[\mathbf{z}^T_{\text{caption}} + \left(\mathbf{P}^\intercal \mathbf{z}^V_{\text{candidate}} \right) \odot \boldsymbol{\delta} \right], \tag{18}$$

which is then compared to the visual embeddings of the target images $\widehat{\mathbf{a}}_{\text{target}} \in \mathbb{R}^d$ using cosine similarity. Performance is evaluated using the average of batch-wise Recall@10 and Recall@50, with batch size equal to 256. The baseline is the performance of the sum of the candidate and relative caption activations, $\widehat{\mathbf{a}}_{\text{original}} = \widehat{\mathbf{a}}^V + \widehat{\mathbf{a}}^T = \mathbf{D}^T \mathbf{z}^T_{\text{caption}} + \mathbf{D}^V \mathbf{z}^V_{\text{candidate}}$, which keeps both bimodal and unimodal information.

Results in Tab. 15 show that the independent SAE variants outperform the shared ones by casting the composed retrieval task into a text-to-image retrieval problem, where the visual modality is used to steer the textual one. By inverting the roles of the modalities in Eq. (18), Tab. 16 show that the independent SAE variants match or outperform the shared ones, except for the Independent SAE without co-activation loss. In this setting, the composed retrieval task is cast into an image-to-image retrieval problem, for which contrastive VLMs are not explicitly optimized. Additionally, code magnitudes of Independent SAEs are not constrained across modalities during training. This may explain the asymmetry in performance between the two cross-modal steering directions that we observe. However, the broader significance of this experiment should be interpreted with caution, given that recent literature suggests activation steering using SAEs is not guaranteed to function reliably and consistently (Girrbach & Akata, 2026; Wu et al., 2025; Peng et al., 2026).

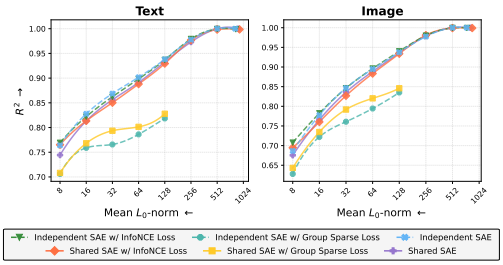
Table 15: Cross-modal steering towards text distribution using SAEs trained on CC3M with CLIP ViT-L/14 ($K = 32$) on FashionIQ. Numbers for Independent SAEs are in green if it performs better than the corresponding Shared SAE variant, and in red otherwise.

SAE	$\frac{1}{2}(R@10 + R@50)$		Relative Improvement (%)
	$\widehat{\mathbf{a}}_{\text{original}}$	$\widehat{\mathbf{a}}_{\text{steered}}$	
Shared	69.34	68.78	-0.81%
Shared w/ Group Sparse Loss	69.12	68.06	-1.53%
Shared w/ InfoNCE Loss	69.05	71.81	4%
Independent	69.47	72.72	4.68%
Independent w/ Group Sparse Loss	67.81	72.94	7.57%
Independent w/ InfoNCE Loss	69.89	73.96	5.82%

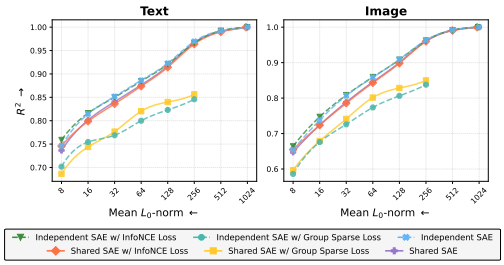
Table 16: Cross-modal steering towards image distribution using SAEs trained on CC3M with CLIP ViT-L/14 ($K = 32$) on FashionIQ. Numbers for Independent SAEs are in green if it performs better than the corresponding Shared SAE variant, and in red otherwise.

SAE	$\frac{1}{2}(R@10 + R@50)$		Relative Improvement (%)
	$\widehat{\mathbf{a}}_{\text{original}}$	$\widehat{\mathbf{a}}_{\text{steered}}$	
Shared	69.34	71.46	3.06%
Shared w/ Group Sparse Loss	69.12	72.28	4.57%
Shared w/ InfoNCE Loss	69.05	69.16	0.16%
Independent	69.47	67.86	-2.32%
Independent w/ Group Sparse Loss	67.81	71.54	5.50%
Independent w/ InfoNCE Loss	69.89	71.24	1.93%

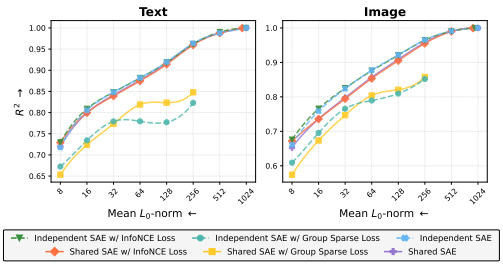
F.3 YFCC15M-V2



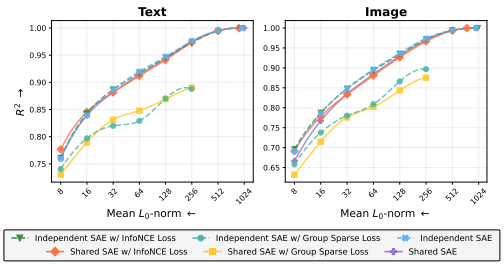
(a) CLIP ViT-B/32



(b) CLIP ViT-L/14

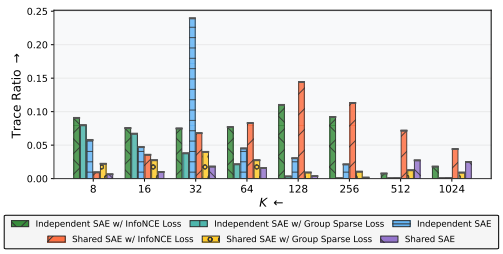


(c) SigLIP ViT-B/16

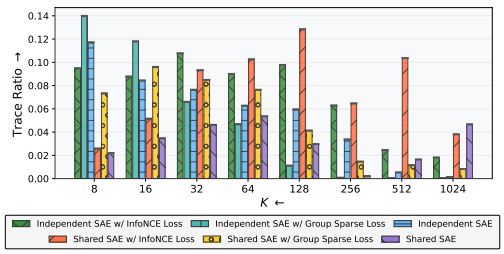


(d) SigLIP 2 ViT-B/16

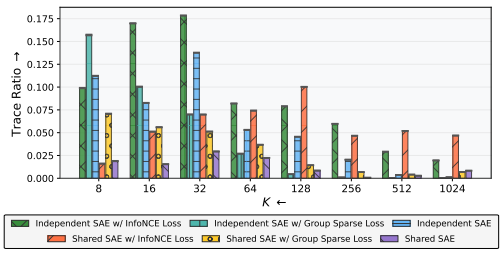
Figure 24: Reconstruction performance on YFCC15M-V2 for (a) CLIP ViT-B/32, (b) CLIP ViT-L/14, (c) SigLIP ViT-B/16, and (d) SigLIP 2 ViT-B/16. Each plot shows the R^2 score of the reconstructed activations against the original activations for both modalities across different sparsity levels. The x-axis represents the mean L_0-norm of the codes (log scale).



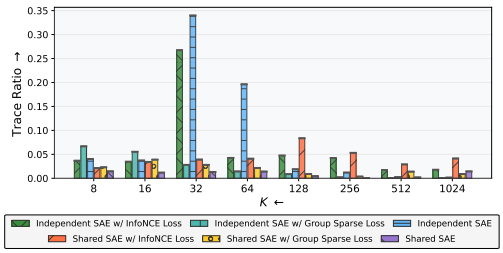
(a) CLIP ViT-B/32



(b) CLIP ViT-L/14

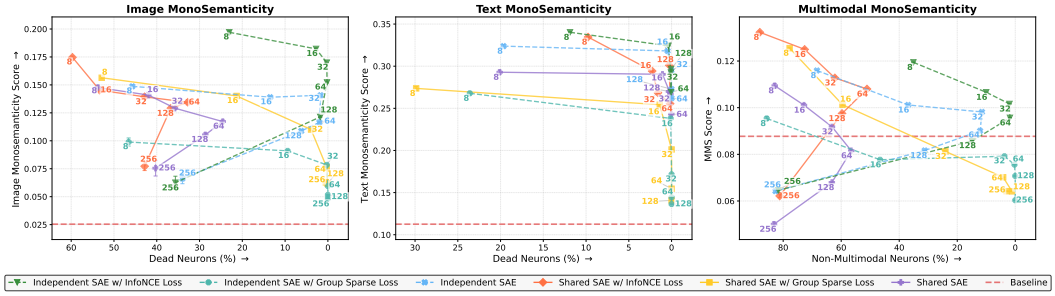


(c) SigLIP ViT-B/16

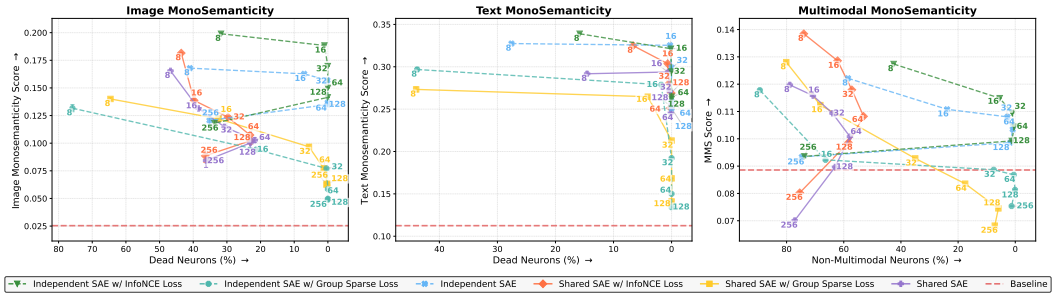


(d) SigLIP 2 ViT-B/16

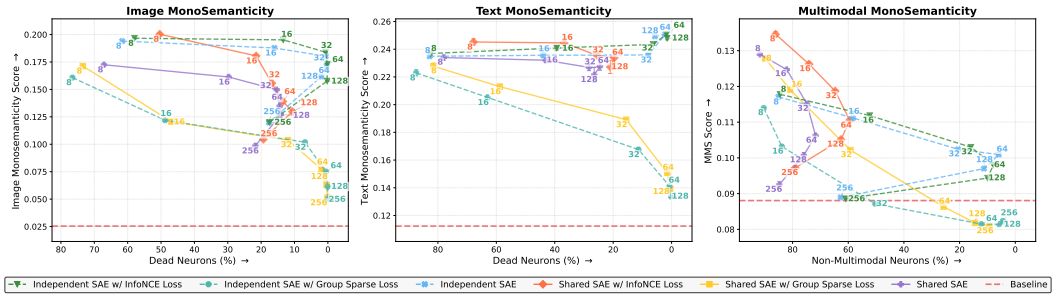
Figure 25: Trace Ratio comparison on YFCC15M-V2 for (a) CLIP ViT-B/32, (b) CLIP ViT-L/14, (c) SigLIP ViT-B/16, and (d) SigLIP 2 ViT-B/16. Metric values, represented as bars, are grouped by the sparsity level K in the TopK activation function.



(a) CLIP ViT-B/32



(b) CLIP ViT-L/14



(c) SigLIP 2 ViT-B/16

Figure 26: Interpretability evaluation for YFCC15M-V2 for (a) CLIP ViT-B/32, (b) CLIP ViT-L/14, and (c) SigLIP 2 ViT-B/16. The plots show (from left to right) Image MonoSemanticity (MS), Text MS, and Multimodal MonoSemanticity (MMS) scores. Each point corresponds to the mean score across all features for a given sparsity level, highlighted in correspondence of each point. The x-axis indicates the percentage of dead neurons, *i.e.*, neurons that are never activated across the evaluation set. For Multimodal MonoSemanticity scores, we indicate the percentage of non-multimodal neurons, *i.e.*, neurons that are never activated across both modalities in the evaluation set. The dashed horizontal line indicates the mean score across all features for the dense VLM model.

Table 17: Zero-shot classification accuracy (%) across datasets, grouped by sparsity level K, for CLIP ViT-B/32 on YFCC15M-V2.

		General datasets						Fine-grained datasets							Misc. datasets				
K	SAE	Avg	ImageNet	CIFAR-10	CIFAR-100	Caltech-101	STL-10	Avg	Food-101	CUB	Cars	Aircraft	Pets	Flowers	Avg	DTD	EuroSAT	RESISC45	Country211
8	Shared	64.5	30.5	81.5	44.4	69.7	96.3	16.7	30.7	10.7	6.6	3.7	23.8	24.6	22.2	19.5	**20.8**	40.1	8.3
	Shared w/ Group Sparse Loss	52.4	16.2	78.9	19.3	52.4	95.3	11.0	19.9	8.8	3.6	1.9	17.8	14.3	14.8	13.9	20.2	19.7	5.6
	Shared w/ InfoNCE Loss	65.0	31.5	81.8	43.7	74.0	94.0	15.7	29.4	8.7	7.0	6.6	23.9	18.6	23.9	24.8	20.5	**42.5**	7.7
	Independent	68.0	35.3	84.2	50.4	73.4	**96.5**	22.5	41.2	13.0	12.8	6.1	34.1	27.6	23.7	26.7	18.5	41.2	8.2
	Independent w/ Group Sparse Loss	62.4	25.7	83.5	42.6	64.5	95.6	14.1	26.0	9.5	4.7	3.8	20.2	20.1	21.5	18.5	18.2	42.0	7.3
	Independent w/ InfoNCE Loss	**69.8**	**39.4**	**85.4**	**51.1**	**76.6**	**96.5**	**26.3**	**45.3**	**16.5**	**18.1**	**7.7**	**42.1**	**28.3**	**24.4**	**29.4**	17.2	41.2	**9.7**
16	Shared	70.4	41.5	84.0	51.1	78.4	**97.0**	28.2	46.1	24.6	13.5	4.6	44.0	36.3	31.1	30.4	**33.7**	49.5	10.8
	Shared w/ Group Sparse Loss	61.6	23.5	83.8	37.9	67.5	95.5	15.7	31.9	16.0	6.9	3.1	20.3	16.0	14.9	12.1	12.9	27.0	7.6
	Shared w/ InfoNCE Loss	67.5	35.2	85.1	47.3	73.7	96.0	22.0	41.3	15.5	11.0	8.9	35.1	20.1	21.3	23.5	18.9	33.3	9.5
	Independent	72.6	46.4	85.7	54.6	**79.3**	96.9	36.1	58.6	27.7	25.3	8.3	**56.4**	40.2	31.3	34.4	29.2	50.0	**11.5**
	Independent w/ Group Sparse Loss	67.6	34.2	86.4	46.1	74.3	96.9	23.2	39.8	18.7	7.7	4.5	34.6	33.7	25.5	23.1	24.6	45.1	9.3
	Independent w/ InfoNCE Loss	**73.4**	**47.3**	**87.6**	**56.8**	78.9	96.5	**37.5**	**61.9**	**27.9**	**27.4**	**10.7**	56.2	**40.9**	**31.9**	**35.7**	29.5	**51.1**	11.2
32	Shared	74.4	48.2	87.2	56.9	82.9	96.8	38.3	63.9	33.1	27.9	8.2	48.8	47.9	33.6	33.9	33.6	53.7	13.2
	Shared w/ Group Sparse Loss	65.9	31.8	86.3	44.6	70.5	96.1	22.4	41.2	16.0	11.1	4.8	41.5	20.1	20.2	16.0	18.6	38.2	7.9
	Shared w/ InfoNCE Loss	73.0	46.1	87.4	56.8	79.2	95.3	36.4	58.7	28.4	25.3	10.5	58.8	36.6	32.3	30.9	**37.3**	50.1	11.1
	Independent	75.6	52.2	87.9	58.1	82.6	**97.1**	46.4	**71.3**	36.4	**37.1**	10.8	**71.0**	51.8	34.9	36.1	36.6	53.1	**13.7**
	Independent w/ Group Sparse Loss	69.0	38.3	86.2	49.9	74.5	96.1	23.8	38.7	19.6	11.4	7.4	31.6	34.2	24.9	27.2	13.6	48.6	10.2
	Independent w/ InfoNCE Loss	**76.1**	**52.7**	**88.7**	**59.3**	**83.3**	96.3	**46.6**	70.9	**38.4**	35.9	**12.1**	69.1	**53.3**	**35.9**	**37.4**	37.0	**56.3**	13.1
64	Shared	76.5	53.2	87.9	59.0	**85.5**	96.8	47.6	71.9	41.0	36.5	13.1	67.2	56.2	38.0	39.8	40.9	**56.8**	14.3
	Shared w/ Group Sparse Loss	67.5	35.7	82.2	48.8	75.0	95.7	20.9	45.2	20.8	8.0	7.3	24.1	19.9	24.2	26.5	13.9	47.4	9.0
	Shared w/ InfoNCE Loss	76.6	54.7	**89.2**	60.3	83.1	95.8	49.9	74.9	40.0	40.7	11.6	77.7	54.4	36.4	36.0	40.3	55.3	14.0
	Independent	77.7	56.4	88.9	61.5	84.9	**97.0**	51.6	74.6	43.2	43.5	**13.5**	77.6	57.1	36.8	**40.1**	37.3	55.7	14.1
	Independent w/ Group Sparse Loss	71.2	41.6	87.5	52.3	77.8	96.7	28.6	47.3	23.7	16.7	7.9	39.2	36.9	30.3	31.3	26.1	53.8	10.0
	Independent w/ InfoNCE Loss	**78.1**	**56.9**	89.1	**62.1**	85.2	**97.0**	**52.7**	**76.3**	**43.9**	**46.2**	12.2	**79.3**	**58.2**	**39.7**	**40.1**	**49.7**	54.3	**14.5**
128	Shared	78.6	58.7	89.4	61.2	86.8	96.8	54.4	77.8	46.1	49.9	15.3	78.8	58.7	39.2	40.3	41.0	60.2	15.2
	Shared w/ Group Sparse Loss	70.1	41.9	83.1	49.2	80.4	96.1	30.4	57.5	24.5	21.5	10.0	39.1	30.0	30.8	34.8	29.5	49.7	9.1
	Shared w/ InfoNCE Loss	79.1	60.0	89.5	63.1	86.1	**96.9**	55.1	77.1	48.1	48.5	14.0	81.9	61.0	41.1	39.7	50.5	58.5	15.7
	Independent	79.2	60.3	88.8	62.8	**87.2**	**96.9**	56.6	79.0	48.5	**51.4**	15.9	**83.4**	61.4	**42.2**	**41.8**	**51.7**	59.8	15.7
	Independent w/ Group Sparse Loss	73.6	48.1	88.0	53.9	81.8	96.2	38.1	61.6	27.3	32.1	10.9	52.7	44.1	33.4	35.3	35.2	51.5	11.4
	Independent w/ InfoNCE Loss	**79.5**	**60.6**	**89.6**	**64.5**	86.0	**96.9**	**56.9**	**80.0**	**49.9**	50.0	**16.5**	82.7	**62.5**	41.5	**41.8**	47.2	**60.8**	**16.3**
256	Shared	80.0	61.8	89.9	63.6	**87.8**	97.0	59.3	81.3	52.5	57.3	17.7	84.5	62.3	41.8	41.3	49.1	60.7	16.0
	Shared w/ Group Sparse Loss	4.4	0.1	10.0	1.0	1.0	10.0	1.1	1.0	0.5	0.5	1.0	2.7	1.0	3.7	2.1	10.0	2.2	0.5
	Shared w/ InfoNCE Loss	80.3	62.1	**90.1**	64.4	87.6	97.1	59.6	81.4	52.5	56.7	**19.3**	84.9	63.0	40.8	41.4	44.6	60.6	**16.7**
	Independent	80.2	62.6	89.5	64.4	87.2	**97.3**	59.8	81.6	52.7	57.1	18.3	85.7	**63.5**	**43.0**	**43.5**	49.8	**62.2**	16.5
	Independent w/ Group Sparse Loss	4.4	0.1	10.0	1.0	1.0	10.0	1.1	1.0	0.5	0.5	1.0	2.7	1.0	3.7	2.1	10.0	2.2	0.5
	Independent w/ InfoNCE Loss	**80.5**	**63.2**	89.7	**64.8**	87.6	**97.3**	**59.9**	**82.1**	**52.9**	**58.2**	17.8	**85.9**	62.6	42.8	43.0	**51.2**	60.3	**16.7**
	CLIP ViT-B/32	80.8	63.9	89.8	65.1	88.3	97.1	61.4	82.9	53.7	59.7	19.7	87.0	65.2	42.8	44.6	49.1	60.2	17.2

Table 18: Zero-shot retrieval performance across datasets, grouped by sparsity level K, for CLIP ViT-B/32 on YFCC15M-V2.

		Text Retrieval								Image Retrieval							
		Flickr30k				COCO				Flickr30k				COCO			
K	SAE	Avg	R@1	R@5	R@10	Avg	R@1	R@5	R@10	Avg	R@1	R@5	R@10	Avg	R@1	R@5	R@10
8	Shared	40.2	14.0	44.0	62.5	24.7	9.0	26.6	38.4	48.7	27.1	53.8	65.2	25.6	10.2	27.7	38.8
	Shared w/ Group Sparse Loss	13.4	4.6	14.0	21.7	11.2	3.7	11.2	18.7	41.7	19.3	46.8	59.0	17.7	5.8	18.7	28.5
	Shared w/ InfoNCE Loss	23.8	3.5	20.2	47.7	21.3	5.9	22.1	35.9	49.8	26.8	55.2	67.4	27.8	11.3	30.6	41.6
	Independent	57.3	36.7	63.1	72.1	35.4	17.6	38.8	49.7	50.7	28.4	56.1	67.6	28.2	13.0	30.2	41.3
	Independent w/ Group Sparse Loss	51.6	30.0	55.7	69.1	29.4	13.9	31.7	42.6	44.1	23.2	48.6	60.4	24.5	10.6	26.3	36.6
	Independent w/ InfoNCE Loss	**60.7**	**40.9**	**66.1**	**75.1**	**38.4**	**20.5**	**41.4**	**53.3**	**52.4**	**29.2**	**58.4**	**69.6**	**29.7**	**14.0**	**32.1**	**43.1**
16	Shared	66.7	42.7	73.1	84.4	43.4	22.7	47.3	60.3	59.9	37.8	65.9	76.0	34.6	16.5	37.8	49.4
	Shared w/ Group Sparse Loss	25.8	11.8	26.6	39.0	22.2	7.2	23.6	35.7	52.2	28.6	57.8	70.1	25.5	10.4	27.6	38.6
	Shared w/ InfoNCE Loss	45.3	19.2	51.3	65.4	31.6	13.7	34.6	46.6	58.9	36.0	64.9	75.7	34.5	16.6	37.6	49.4
	Independent	71.9	51.3	**78.2**	**86.1**	49.1	29.0	53.6	64.6	**61.8**	**39.3**	67.8	**78.4**	36.6	18.6	39.9	51.3
	Independent w/ Group Sparse Loss	67.7	45.9	74.9	82.4	41.6	22.5	46.0	56.3	58.1	35.5	64.1	74.8	33.5	16.3	36.3	47.9
	Independent w/ InfoNCE Loss	**72.6**	**54.8**	**78.2**	84.9	**50.5**	**29.9**	**55.6**	**65.9**	61.7	38.4	**68.4**	78.3	**37.5**	**19.3**	**40.8**	**52.4**
32	Shared	75.3	52.7	83.3	90.0	51.9	29.9	56.9	69.0	67.0	45.7	73.0	82.2	40.2	21.1	43.8	55.6
	Shared w/ Group Sparse Loss	42.5	18.9	46.4	62.3	33.5	14.4	36.8	49.4	59.9	36.4	66.2	77.0	30.9	13.1	34.1	45.5
	Shared w/ InfoNCE Loss	73.7	51.9	81.0	88.2	48.9	27.4	54.0	65.4	66.3	43.7	73.1	82.1	40.5	21.2	44.2	56.0
	Independent	**81.5**	**63.7**	**87.2**	**93.5**	57.2	36.7	62.2	72.7	**69.0**	**48.1**	**75.1**	83.8	41.6	22.3	45.5	57.1
	Independent w/ Group Sparse Loss	70.5	47.6	76.8	87.0	44.3	24.1	48.7	60.0	59.6	36.6	65.3	76.8	34.3	16.3	37.6	48.9
	Independent w/ InfoNCE Loss	80.5	63.2	86.3	92.1	**57.7**	**36.9**	**62.9**	**73.2**	68.5	46.8	74.8	**83.9**	**42.0**	**22.6**	**45.8**	**57.6**
64	Shared	83.6	67.4	88.3	95.0	59.2	38.3	64.4	74.8	71.6	51.4	77.5	85.8	44.1	24.6	48.2	59.6
	Shared w/ Group Sparse Loss	56.9	32.2	62.1	76.3	39.2	19.2	42.6	55.8	58.1	33.1	64.7	76.4	31.5	13.3	34.6	46.7
	Shared w/ InfoNCE Loss	84.2	67.4	89.3	**95.9**	60.4	40.1	65.6	75.4	72.8	52.2	79.1	87.0	45.0	**25.4**	**49.1**	60.6
	Independent	**85.5**	**70.7**	**90.3**	95.4	**62.0**	**41.8**	**67.1**	**77.1**	72.2	51.9	78.0	86.6	44.7	25.1	48.6	60.4
	Independent w/ Group Sparse Loss	73.3	50.5	80.8	88.5	47.7	26.9	51.8	64.3	60.7	39.0	66.0	77.2	35.3	17.2	38.4	50.3
	Independent w/ InfoNCE Loss	84.9	69.5	89.8	95.5	61.8	41.6	66.9	76.9	**72.9**	**52.4**	**79.2**	**87.1**	**45.1**	25.3	**49.1**	**61.0**
128	Shared	85.8	70.2	90.9	96.4	64.0	43.5	69.5	78.9	74.4	54.5	80.9	87.8	47.6	27.6	51.8	63.3
	Shared w/ Group Sparse Loss	60.6	35.6	66.6	79.6	42.4	21.0	46.6	59.6	60.3	36.7	67.0	77.3	33.7	15.8	36.8	48.5
	Shared w/ InfoNCE Loss	87.3	73.3	91.9	96.8	64.8	44.7	70.5	79.3	75.1	55.5	81.3	88.5	48.0	28.1	52.3	63.5
	Independent	88.2	75.5	92.2	**96.9**	66.2	46.1	**72.0**	80.5	**75.7**	**56.3**	**82.0**	**88.8**	48.2	28.0	**52.5**	**64.1**
	Independent w/ Group Sparse Loss	76.8	56.3	83.9	90.2	51.1	29.1	55.9	68.2	63.8	42.5	70.0	79.0	37.3	19.4	40.4	52.0
	Independent w/ InfoNCE Loss	**88.8**	**76.8**	**92.6**	**96.9**	**66.4**	**46.5**	71.7	**81.0**	75.5	56.0	81.6	**88.8**	**48.3**	**28.4**	**52.5**	64.0
256	Shared	88.3	74.1	93.6	97.3	66.9	47.2	72.7	80.7	76.1	57.0	82.2	89.1	49.4	29.5	53.7	65.0
	Shared w/ Group Sparse Loss	0.1	0.1	0.1	0.2	0.0	0.0	0.0	0.0	0.5	0.1	0.5	1.0	0.1	0.0	0.1	0.2
	Shared w/ InfoNCE Loss	89.7	77.1	94.1	**97.9**	67.8	48.2	73.2	82.1	76.8	58.0	82.9	89.5	49.7	29.5	54.0	65.5
	Independent	89.9	77.4	94.6	97.6	68.0	48.7	73.4	81.9	76.7	57.9	82.9	89.3	49.5	29.4	53.7	65.5
	Independent w/ Group Sparse Loss	0.1	0.1	0.1	0.2	0.0	0.0	0.0	0.0	0.5	0.1	0.5	1.0	0.1	0.0	0.1	0.2
	Independent w/ InfoNCE Loss	**90.1**	**78.0**	**94.7**	97.7	**68.7**	**49.7**	**73.8**	**82.6**	**77.0**	**58.3**	**83.0**	**89.8**	**50.0**	**30.0**	**54.3**	**65.8**
	CLIP ViT-B/32	90.8	79.3	95.0	98.1	69.3	50.0	74.8	83.2	77.4	58.8	83.4	90.1	50.5	30.5	54.8	66.2

Table 19: Zero-shot classification accuracy (%) across datasets, grouped by sparsity level K, for CLIP ViT-L/14 on YFCC15M-V2.

		General datasets						Fine-grained datasets							Misc. datasets				
K	SAE	Avg	ImageNet	CIFAR-10	CIFAR-100	Caltech-101	STL-10	Avg	Food-101	CUB	Cars	Aircraft	Pets	Flowers	Avg	DTD	EuroSAT	RESISC45	Country211
8	Shared	74.4	44.4	89.6	58.3	81.0	98.7	27.3	47.7	16.3	13.6	4.3	42.4	39.8	34.4	34.5	31.5	54.5	17.2
	Shared w/ Group Sparse Loss	68.1	32.2	90.7	50.4	68.3	98.7	18.4	38.8	11.5	8.4	3.9	27.5	20.5	30.2	31.4	31.7	46.8	11.0
	Shared w/ InfoNCE Loss	74.3	45.5	88.5	59.3	79.9	98.5	29.6	50.5	17.2	16.7	8.0	45.1	40.0	27.8	33.6	14.0	45.8	17.6
	Independent	78.3	47.6	**92.8**	**66.0**	**85.9**	**99.2**	35.2	61.3	18.2	30.2	10.3	41.6	**49.8**	**36.4**	32.6	37.9	**56.5**	18.4
	Independent w/ Group Sparse Loss	71.7	34.6	92.2	55.0	77.9	99.0	19.9	40.4	8.2	9.1	8.8	23.1	30.0	33.8	25.3	**41.9**	52.5	15.3
	Independent w/ InfoNCE Loss	**78.6**	**51.4**	92.4	64.9	85.2	98.9	**38.5**	**67.7**	**21.6**	**32.4**	**13.6**	**46.3**	49.5	35.6	**37.8**	31.7	54.0	**18.9**
16	Shared	79.4	54.5	93.6	65.6	84.2	**99.2**	42.2	73.0	30.8	33.8	10.2	48.9	56.6	40.1	40.6	38.0	60.2	21.6
	Shared w/ Group Sparse Loss	72.2	37.5	92.7	56.8	74.9	99.0	21.8	41.5	20.1	12.1	5.4	28.8	23.2	30.2	31.5	29.2	44.9	15.1
	Shared w/ InfoNCE Loss	78.3	52.4	93.6	62.0	84.6	99.1	37.4	69.4	20.1	31.1	11.7	50.1	41.9	36.9	36.9	32.5	57.9	20.3
	Independent	**82.3**	59.3	**94.3**	**70.2**	**88.4**	**99.2**	49.7	78.9	32.0	42.8	17.2	65.3	62.0	**42.8**	44.0	**41.6**	**62.5**	23.0
	Independent w/ Group Sparse Loss	77.8	46.1	94.1	65.4	84.4	99.1	34.5	55.4	22.4	21.1	12.5	44.5	51.0	37.9	29.3	41.5	60.8	20.0
	Independent w/ InfoNCE Loss	82.3	**61.0**	93.6	69.3	88.2	**99.2**	**52.5**	**80.8**	**33.4**	**47.4**	**17.3**	**73.1**	**63.0**	41.2	**44.7**	34.4	61.8	**23.8**
32	Shared	83.2	61.7	94.8	70.9	89.2	99.2	51.6	81.3	38.5	42.5	14.0	71.0	62.4	44.5	43.0	43.0	67.6	24.5
	Shared w/ Group Sparse Loss	72.2	40.0	94.5	60.2	67.3	99.1	25.8	55.7	23.2	11.8	5.2	24.1	34.8	39.3	32.9	51.9	55.6	16.7
	Shared w/ InfoNCE Loss	83.5	60.5	**95.2**	**73.3**	89.0	**99.3**	50.9	83.2	39.0	44.7	15.5	66.4	56.3	47.2	42.0	62.4	60.2	24.3
	Independent	**84.8**	**65.6**	95.1	72.8	**91.3**	**99.3**	57.9	85.8	42.1	54.5	**20.2**	77.1	67.8	**50.5**	**46.8**	**62.6**	66.4	**26.3**
	Independent w/ Group Sparse Loss	79.5	51.9	**95.2**	65.0	86.5	99.1	39.5	62.2	29.1	24.8	10.2	58.5	51.9	41.1	34.3	48.0	61.6	20.5
	Independent w/ InfoNCE Loss	**84.8**	**65.6**	95.0	73.0	91.2	**99.3**	**58.6**	**86.6**	**43.0**	**57.0**	18.1	**77.9**	**68.9**	47.2	45.4	49.5	**67.9**	26.0
64	Shared	85.4	67.4	94.7	73.8	91.8	99.3	62.2	86.8	51.0	59.3	21.3	84.0	70.6	49.8	48.9	53.8	69.6	27.0
	Shared w/ Group Sparse Loss	77.5	43.0	94.9	65.8	84.3	99.3	36.6	69.7	25.3	27.1	11.2	45.3	41.3	38.0	34.9	42.6	54.7	19.9
	Shared w/ InfoNCE Loss	85.6	66.9	**95.3**	**75.2**	91.4	**99.4**	61.5	87.4	48.0	61.2	18.7	82.8	70.8	46.7	46.6	46.8	67.3	26.2
	Independent	86.4	69.7	94.9	75.0	**93.0**	**99.4**	**65.8**	**88.9**	53.4	65.6	**24.9**	**89.4**	72.5	50.3	**49.0**	54.8	70.1	27.5
	Independent w/ Group Sparse Loss	81.6	57.8	94.6	68.0	88.4	99.3	46.6	69.7	35.3	38.7	13.7	64.6	57.9	46.2	40.0	58.6	63.5	22.8
	Independent w/ InfoNCE Loss	**86.4**	**70.2**	95.1	75.1	92.4	99.3	65.6	88.2	**54.3**	**67.5**	22.9	87.3	**73.5**	**52.1**	**49.0**	**60.9**	**71.1**	**27.6**
128	Shared	86.6	71.5	94.4	75.2	92.4	**99.4**	67.9	89.2	58.0	69.7	25.7	90.2	74.9	51.9	51.3	58.2	69.2	29.1
	Shared w/ Group Sparse Loss	80.9	55.1	94.3	68.2	88.7	98.4	40.7	72.2	36.2	39.3	10.4	45.1	40.8	41.7	40.1	51.5	54.8	20.5
	Shared w/ InfoNCE Loss	86.9	71.6	94.7	75.5	**93.2**	99.3	67.4	89.5	56.0	69.5	23.4	90.4	75.6	50.9	51.9	52.3	70.4	29.0
	Independent	**87.4**	**72.7**	94.9	**77.2**	92.8	**99.4**	**69.1**	90.0	**58.6**	71.9	26.9	**91.3**	**76.2**	52.6	**52.0**	57.4	**71.6**	29.4
	Independent w/ Group Sparse Loss	83.3	62.7	94.4	70.5	89.8	99.2	53.2	76.2	41.9	53.6	16.8	74.7	56.0	41.6	37.9	43.0	61.3	24.2
	Independent w/ InfoNCE Loss	**87.4**	**72.7**	**95.2**	77.0	92.8	99.3	69.0	**90.2**	56.7	**72.7**	**28.5**	90.2	75.6	**53.3**	51.5	**60.9**	71.3	**29.5**
256	Shared	87.9	74.1	95.1	76.7	**94.0**	**99.4**	**72.0**	90.5	62.0	75.9	32.5	**93.0**	77.9	**54.5**	53.7	**62.0**	72.0	30.5
	Shared w/ Group Sparse Loss	82.5	60.2	95.2	67.5	90.3	99.2	49.4	80.7	38.1	44.7	18.5	65.3	48.9	43.3	41.5	52.1	56.2	23.5
	Shared w/ InfoNCE Loss	**88.0**	74.0	95.3	**77.6**	93.9	99.3	71.8	**90.9**	61.8	75.2	**32.6**	92.1	**78.1**	53.1	53.0	56.1	72.2	30.9
	Independent	88.0	**74.3**	**95.4**	77.5	93.5	99.3	71.7	90.8	61.8	**76.1**	31.0	92.5	78.0	54.5	**54.5**	59.9	**72.6**	30.9
	Independent w/ Group Sparse Loss	84.7	66.3	93.8	73.0	91.3	99.3	58.6	82.6	48.8	61.8	19.8	80.0	58.6	44.4	42.8	47.5	61.6	25.7
	Independent w/ InfoNCE Loss	88.0	74.2	**95.4**	**77.6**	93.4	**99.4**	71.7	**90.9**	**62.1**	75.7	31.0	**93.0**	77.3	54.4	53.6	60.3	72.3	**31.3**
	CLIP ViT-L/14	88.4	75.2	95.6	78.3	93.3	99.4	72.7	91.6	62.5	77.9	32.1	93.2	78.9	55.5	54.5	63.9	71.7	31.7

Table 20: Zero-shot retrieval performance across datasets, grouped by sparsity level K, for CLIP ViT-L/14 on YFCC15M-V2.

		Text Retrieval								Image Retrieval							
		Flickr30k				COCO				Flickr30k				COCO			
K	SAE	Avg	R@1	R@5	R@10	Avg	R@1	R@5	R@10	Avg	R@1	R@5	R@10	Avg	R@1	R@5	R@10
8	Shared	50.0	24.3	54.8	70.8	32.1	13.8	35.1	47.4	55.4	32.1	61.6	72.6	30.5	14.5	33.1	43.9
	Shared w/ Group Sparse Loss	39.5	20.6	42.5	55.3	24.7	10.7	26.1	37.2	50.6	27.8	56.0	68.0	26.6	12.1	28.6	39.0
	Shared w/ InfoNCE Loss	50.1	23.9	54.6	71.7	34.4	16.3	37.4	49.6	55.3	33.3	60.8	71.9	31.2	15.0	33.8	44.7
	Independent	66.6	**47.2**	71.4	81.3	**43.7**	**25.0**	**47.3**	**58.8**	**57.6**	**35.6**	63.2	**73.9**	**32.8**	15.8	**35.5**	**47.0**
	Independent w/ Group Sparse Loss	57.5	36.0	63.1	73.4	34.2	17.2	36.9	48.4	51.8	28.2	57.9	69.2	28.7	13.5	30.8	41.8
	Independent w/ InfoNCE Loss	**67.6**	46.5	**73.1**	**83.1**	43.5	24.5	47.2	58.7	57.1	35.1	**63.5**	72.8	32.7	**16.1**	**35.5**	46.4
16	Shared	71.8	46.5	79.5	89.5	48.6	27.6	53.2	64.9	64.9	43.5	70.9	80.3	39.3	21.0	42.4	54.4
	Shared w/ Group Sparse Loss	51.1	28.8	55.1	69.5	34.9	16.6	38.0	50.2	61.8	38.8	68.5	78.0	34.3	17.1	37.3	48.5
	Shared w/ InfoNCE Loss	64.7	38.6	71.3	84.1	45.2	24.9	49.4	61.2	65.2	43.9	71.6	80.1	38.8	20.6	42.2	53.7
	Independent	**80.9**	**64.2**	85.9	92.5	54.8	**34.8**	59.4	70.3	65.6	44.1	72.0	80.7	39.9	21.3	43.5	54.9
	Independent w/ Group Sparse Loss	73.3	51.7	79.5	88.6	47.8	28.2	52.1	63.2	63.5	41.2	69.8	79.5	36.9	19.0	40.3	51.5
	Independent w/ InfoNCE Loss	80.4	61.4	**87.0**	**92.9**	**55.2**	**34.8**	**59.9**	**71.0**	**66.1**	**44.5**	**72.6**	**81.3**	**40.4**	**21.9**	**43.9**	**55.5**
32	Shared	81.8	62.8	87.8	94.7	58.6	38.5	64.0	73.4	70.8	50.1	77.4	84.8	44.7	25.2	48.8	60.0
	Shared w/ Group Sparse Loss	43.5	25.5	48.3	56.6	26.0	12.5	28.1	37.4	66.5	44.4	73.2	81.9	39.2	20.5	42.7	54.5
	Shared w/ InfoNCE Loss	79.4	58.6	86.3	93.3	57.3	36.1	63.0	72.9	70.9	50.0	77.3	85.4	44.8	25.5	48.9	60.1
	Independent	86.0	70.3	91.2	**96.5**	62.2	41.8	67.8	77.1	71.6	50.8	**78.3**	**85.7**	45.3	25.6	49.4	60.8
	Independent w/ Group Sparse Loss	78.5	58.6	84.5	92.3	52.6	32.0	57.7	68.2	66.5	44.9	72.7	82.0	39.7	20.6	43.4	55.1
	Independent w/ InfoNCE Loss	**86.2**	**70.5**	**92.1**	96.1	**63.6**	**43.7**	**68.9**	**78.2**	**71.8**	**51.5**	78.2	**85.7**	**46.1**	**26.5**	**50.2**	**61.5**
64	Shared	87.9	73.8	93.0	97.0	64.6	45.2	69.5	79.1	74.9	55.1	81.4	88.2	48.8	28.9	53.3	64.3
	Shared w/ Group Sparse Loss	60.7	40.4	65.9	75.7	42.1	23.9	45.3	57.0	72.1	51.4	78.9	86.1	44.2	24.6	48.1	59.8
	Shared w/ InfoNCE Loss	89.6	76.5	94.4	97.8	66.2	46.4	71.6	80.7	75.9	56.8	81.8	**89.2**	49.2	29.3	53.6	64.6
	Independent	89.9	77.0	94.3	**98.4**	67.2	47.2	72.6	81.7	75.4	55.9	81.8	88.5	49.2	29.3	53.6	64.6
	Independent w/ Group Sparse Loss	84.4	68.9	90.0	94.2	58.1	36.9	63.3	74.1	70.8	49.3	77.3	85.8	43.5	23.6	47.6	59.4
	Independent w/ InfoNCE Loss	**90.4**	**78.2**	**95.2**	97.8	**68.2**	**48.7**	**73.2**	**82.8**	**76.3**	**57.0**	**82.7**	**89.2**	**49.9**	**30.0**	**54.4**	**65.4**
128	Shared	91.1	78.5	96.0	**98.8**	70.0	50.5	75.9	83.7	78.2	59.8	84.7	90.1	52.1	32.1	56.7	67.6
	Shared w/ Group Sparse Loss	68.0	47.5	73.6	82.9	44.1	23.8	47.9	60.6	73.6	53.8	79.8	87.3	45.7	26.2	49.9	61.1
	Shared w/ InfoNCE Loss	91.8	81.4	95.9	98.0	70.1	50.3	76.0	83.9	78.1	59.2	84.6	90.5	52.3	32.0	56.9	67.9
	Independent	92.0	**81.6**	96.0	98.4	70.4	51.3	75.9	84.1	78.8	60.8	**85.3**	90.4	52.4	32.3	57.2	67.8
	Independent w/ Group Sparse Loss	85.5	68.0	92.0	96.5	59.5	38.4	64.6	75.6	71.8	51.1	78.2	86.0	44.1	24.6	47.9	59.8
	Independent w/ InfoNCE Loss	**92.1**	80.9	**96.7**	98.7	**70.8**	**51.7**	**76.2**	**84.6**	**79.1**	**61.0**	**85.3**	**91.0**	**52.7**	**32.5**	**57.3**	**68.2**
256	Shared	93.0	83.2	97.1	98.7	72.6	54.1	77.6	**86.2**	80.1	62.5	86.1	91.7	**54.4**	34.3	**59.3**	**69.6**
	Shared w/ Group Sparse Loss	68.3	46.8	74.6	83.4	44.7	24.3	48.8	60.9	72.1	52.3	78.3	85.7	45.1	25.5	49.3	60.6
	Shared w/ InfoNCE Loss	90.6	77.8	95.4	98.5	71.4	52.6	76.6	85.1	80.1	62.9	85.9	91.6	54.3	**34.4**	58.8	**69.6**
	Independent	93.3	83.7	97.3	98.8	73.1	**54.7**	78.4	**86.2**	80.7	63.3	86.7	**92.1**	54.3	34.3	**59.3**	69.4
	Independent w/ Group Sparse Loss	87.0	70.9	93.6	96.5	61.6	40.6	67.0	77.3	73.8	54.1	80.0	87.2	46.3	26.3	50.6	61.9
	Independent w/ InfoNCE Loss	**93.4**	**83.9**	**97.4**	**99.0**	**73.2**	54.6	**78.9**	**86.2**	**80.8**	**63.6**	**86.8**	91.9	**54.4**	**34.4**	59.2	**69.6**
	CLIP ViT-L/14	94.0	85.3	97.4	99.3	74.2	56.1	79.6	86.9	81.3	64.7	87.1	92.2	55.2	35.5	59.9	70.3

Table 21: Zero-shot classification accuracy (%) across datasets, grouped by sparsity level K, for SigLIP ViT-B/16 on YFCC15M-V2.

		General datasets						Fine-grained datasets							Misc. datasets				
K	SAE	Avg	ImageNet	CIFAR-10	CIFAR-100	Caltech-101	STL-10	Avg	Food-101	CUB	Cars	Aircraft	Pets	Flowers	Avg	DTD	EuroSAT	RESISC45	Country211
8	Shared	74.2	44.3	85.7	60.1	83.5	**97.6**	28.7	55.8	19.6	19.7	4.7	34.3	37.9	31.9	42.7	29.5	47.5	8.1
	Shared w/ Group Sparse Loss	70.2	35.0	**87.2**	57.3	74.2	97.4	20.9	36.4	13.7	19.3	3.3	29.4	23.6	24.9	31.5	19.3	42.7	6.0
	Shared w/ InfoNCE Loss	75.4	49.3	86.7	61.6	81.9	97.4	30.8	56.5	18.7	22.7	**9.0**	41.7	36.3	32.0	42.7	28.8	48.3	8.4
	Independent	74.5	47.8	84.5	61.1	81.7	97.5	33.6	56.7	17.5	37.1	6.4	44.4	39.7	31.0	44.4	23.4	47.9	8.2
	Independent w/ Group Sparse Loss	69.8	36.0	85.4	56.6	73.5	97.4	22.3	37.7	10.8	23.4	1.3	28.0	32.4	30.1	33.1	29.1	**51.3**	6.8
	Independent w/ InfoNCE Loss	**76.4**	**50.4**	86.3	**61.7**	**86.3**	97.4	**37.3**	**62.7**	**20.0**	**39.3**	5.9	**47.7**	**48.4**	**33.6**	**46.3**	**32.6**	47.1	**8.6**
16	Shared	79.7	56.2	89.0	66.3	**89.1**	**98.0**	41.6	67.5	33.7	39.3	10.3	46.4	52.2	36.7	48.6	30.9	56.7	10.5
	Shared w/ Group Sparse Loss	73.4	42.4	89.2	59.3	78.1	**98.0**	31.0	47.3	26.6	26.1	8.7	42.6	34.5	30.3	33.4	31.2	48.7	8.0
	Shared w/ InfoNCE Loss	78.9	55.1	88.4	65.4	88.0	97.5	45.0	70.1	29.2	42.7	**13.3**	63.3	51.6	34.5	44.3	29.2	53.7	10.6
	Independent	80.9	61.8	**89.5**	66.2	89.0	97.9	52.0	75.7	35.9	56.2	12.8	**69.9**	61.6	36.4	53.1	26.9	55.1	10.5
	Independent w/ Group Sparse Loss	76.0	48.2	89.0	62.8	82.5	97.6	35.7	53.0	27.7	29.8	5.1	54.3	44.5	34.7	41.7	**31.5**	56.9	8.6
	Independent w/ InfoNCE Loss	**81.1**	**62.7**	89.0	**66.9**	88.8	97.9	**53.7**	**76.5**	**36.6**	**58.8**	12.6	**69.9**	**67.7**	**38.8**	**55.7**	30.9	**57.6**	**10.8**
32	Shared	81.9	62.5	90.8	68.2	89.9	98.1	56.2	76.2	43.9	55.6	14.1	77.5	69.9	39.7	56.6	32.2	58.0	11.8
	Shared w/ Group Sparse Loss	78.0	49.9	90.1	64.2	87.9	97.9	39.4	61.8	31.6	33.5	9.9	49.8	50.0	33.4	39.0	30.7	54.8	9.2
	Shared w/ InfoNCE Loss	82.0	63.3	**91.3**	68.1	89.0	98.2	53.8	76.3	42.6	59.6	17.4	79.3	47.9	37.4	52.6	24.0	**61.6**	11.6
	Independent	83.1	67.3	90.0	68.6	**91.4**	98.1	62.6	82.7	**50.4**	**70.1**	20.5	81.3	70.5	39.1	**58.7**	25.9	59.4	12.5
	Independent w/ Group Sparse Loss	79.2	56.3	89.9	65.9	86.1	98.0	44.3	66.1	35.5	40.3	3.5	64.7	56.0	37.0	46.1	34.9	56.5	10.6
	Independent w/ InfoNCE Loss	**83.4**	**67.5**	91.2	**68.8**	**91.4**	**98.3**	**63.9**	**83.2**	48.2	69.7	**27.2**	**83.9**	**71.0**	**41.1**	58.6	**35.3**	58.0	**12.6**
64	Shared	84.2	68.3	91.6	70.0	92.9	98.1	67.1	83.7	55.7	73.3	**26.8**	87.3	75.9	42.4	58.4	35.1	62.3	**13.7**
	Shared w/ Group Sparse Loss	80.9	57.5	90.5	67.3	91.3	98.0	49.6	70.2	38.7	55.3	15.1	69.1	49.4	37.7	48.9	33.0	58.4	10.6
	Shared w/ InfoNCE Loss	84.3	68.2	**91.9**	70.1	**93.4**	98.0	64.3	83.0	52.5	75.4	23.1	84.7	67.3	41.1	57.4	31.8	62.3	13.1
	Independent	84.6	70.4	91.4	**70.3**	92.9	98.1	**69.3**	**86.4**	**57.3**	**78.9**	26.5	**89.3**	**77.2**	**43.9**	60.3	**38.2**	**63.5**	13.5
	Independent w/ Group Sparse Loss	79.6	57.1	90.8	64.8	87.3	98.0	46.4	63.2	35.3	50.4	3.6	73.0	52.7	34.9	45.1	26.0	57.7	10.8
	Independent w/ InfoNCE Loss	**84.8**	**71.2**	91.6	69.9	92.8	**98.3**	68.7	**86.4**	56.8	78.0	26.6	87.8	76.8	43.4	**61.0**	37.2	61.7	**13.7**
128	Shared	85.3	72.4	**91.6**	70.6	94.0	**98.1**	72.2	86.7	60.9	84.1	31.6	90.7	79.3	**45.2**	61.2	**42.4**	62.7	14.7
	Shared w/ Group Sparse Loss	80.9	60.0	89.9	65.4	91.1	97.9	48.6	74.9	42.9	59.9	17.4	52.3	44.4	36.4	48.0	30.8	55.9	10.9
	Shared w/ InfoNCE Loss	85.4	72.7	91.4	70.8	94.0	**98.1**	71.6	86.8	60.5	83.8	30.6	90.1	77.9	43.5	60.6	35.8	63.1	14.7
	Independent	**85.8**	73.1	**91.6**	**71.6**	**94.5**	98.0	73.1	**87.5**	61.0	85.4	**33.3**	92.4	79.1	45.2	**63.2**	39.5	**63.2**	**14.9**
	Independent w/ Group Sparse Loss	80.9	60.4	91.4	65.9	89.2	97.7	49.1	64.9	37.1	71.9	8.6	66.8	45.1	37.1	49.9	30.6	57.0	10.9
	Independent w/ InfoNCE Loss	85.7	**73.2**	**91.6**	71.2	94.4	**98.1**	**73.6**	87.4	**62.3**	**85.7**	31.4	**92.9**	**81.9**	45.2	61.6	42.1	62.2	**14.9**
256	Shared	86.2	75.1	92.0	71.4	94.4	98.1	76.9	**89.0**	**67.0**	89.5	39.8	93.3	82.7	47.4	**64.3**	**45.4**	64.5	15.5
	Shared w/ Group Sparse Loss	82.5	65.4	90.3	67.1	92.3	97.6	62.6	80.0	49.2	78.2	26.8	86.8	54.5	41.6	55.9	40.1	58.8	11.7
	Shared w/ InfoNCE Loss	86.2	75.0	91.7	71.5	**94.7**	**98.2**	76.7	88.7	66.5	89.1	41.4	92.9	81.7	46.2	64.1	40.7	64.7	15.2
	Independent	**86.3**	**75.2**	**92.1**	**71.9**	94.3	98.1	77.4	**89.0**	66.9	**90.1**	40.2	**94.1**	**84.0**	**47.5**	64.2	**45.4**	**64.9**	15.5
	Independent w/ Group Sparse Loss	83.1	66.7	90.6	67.6	93.1	97.7	62.1	78.6	48.8	83.9	18.9	82.4	60.2	41.6	56.8	37.6	60.4	11.5
	Independent w/ InfoNCE Loss	86.2	75.1	92.0	**71.9**	94.1	98.1	**77.4**	88.9	**67.0**	89.6	**42.1**	93.9	83.1	45.2	64.1	36.6	64.7	**15.6**
	SigLIP ViT-B/16	86.7	76.0	92.4	72.2	94.9	98.2	78.5	89.6	68.6	90.8	44.0	94.2	84.1	48.3	66.3	45.8	65.1	15.9

Table 22: Zero-shot retrieval performance across datasets, grouped by sparsity level K, for SigLIP ViT-B/16 on YFCC15M-V2.

		Text Retrieval								Image Retrieval							
		Flickr30k				COCO				Flickr30k				COCO			
K	SAE	Avg	R@1	R@5	R@10	Avg	R@1	R@5	R@10	Avg	R@1	R@5	R@10	Avg	R@1	R@5	R@10
8	Shared	14.6	5.8	15.9	22.1	14.9	2.7	14.9	27.0	54.4	32.0	60.1	71.1	29.2	12.1	31.9	43.6
	Shared w/ Group Sparse Loss	32.5	15.5	35.6	46.4	23.5	10.0	24.9	35.7	41.4	19.5	46.5	58.2	22.3	9.0	23.9	34.0
	Shared w/ InfoNCE Loss	12.3	5.2	12.9	18.9	8.1	1.5	6.8	16.1	**57.4**	**35.2**	**63.2**	**73.8**	32.4	15.1	35.2	46.9
	Independent	**63.2**	**40.7**	**69.8**	**79.0**	**39.3**	**20.8**	**42.7**	**54.3**	56.5	34.3	62.0	73.1	34.8	17.2	38.0	49.1
	Independent w/ Group Sparse Loss	56.9	34.0	62.5	74.3	32.1	15.4	34.5	46.4	54.4	31.7	60.5	71.0	32.2	15.4	35.0	46.1
	Independent w/ InfoNCE Loss	61.1	39.5	67.0	76.7	38.8	20.3	42.1	53.9	56.4	33.9	62.3	73.1	**36.3**	**18.3**	**39.5**	**51.2**
16	Shared	48.9	22.6	53.5	70.6	39.4	16.5	43.5	58.2	69.7	48.3	76.3	84.4	44.4	24.3	48.6	60.3
	Shared w/ Group Sparse Loss	44.4	24.1	47.6	61.6	32.7	15.7	34.8	47.5	58.6	32.9	65.5	77.4	33.2	16.0	36.0	47.6
	Shared w/ InfoNCE Loss	28.8	11.1	30.7	44.5	32.2	10.2	35.0	51.4	67.7	46.1	74.0	83.1	41.8	22.2	45.7	57.5
	Independent	**75.9**	55.1	**82.3**	**90.3**	54.0	33.4	58.6	70.1	**71.3**	**50.7**	**77.6**	**85.6**	47.0	**26.9**	51.4	62.7
	Independent w/ Group Sparse Loss	72.6	49.5	80.9	87.3	46.5	26.3	50.6	62.6	65.2	42.7	71.7	81.1	40.4	21.1	44.0	56.0
	Independent w/ InfoNCE Loss	75.6	**55.9**	81.9	88.9	**54.8**	**33.9**	**59.9**	**70.6**	69.8	49.0	76.1	84.4	**47.2**	26.8	**51.5**	**63.2**
32	Shared	70.3	45.5	77.2	88.1	57.1	34.3	62.8	74.1	77.4	58.2	83.8	90.1	52.5	31.7	57.3	68.4
	Shared w/ Group Sparse Loss	50.9	28.9	54.9	69.0	39.3	17.3	43.5	57.1	68.9	46.5	76.0	84.1	41.4	22.0	45.3	57.0
	Shared w/ InfoNCE Loss	60.5	35.4	65.3	80.8	55.0	29.9	61.1	74.1	77.0	57.7	83.5	89.8	52.0	31.2	56.8	67.9
	Independent	**85.4**	**69.5**	**91.1**	**95.6**	**64.2**	**44.0**	**69.4**	79.3	77.6	58.8	**84.5**	89.6	53.6	32.8	58.6	69.4
	Independent w/ Group Sparse Loss	83.3	65.1	90.3	94.5	57.1	35.7	62.1	73.4	73.6	52.4	80.8	87.6	48.2	27.5	52.6	64.4
	Independent w/ InfoNCE Loss	85.1	**69.5**	90.5	95.2	63.9	43.2	69.2	**79.4**	**78.1**	**59.8**	84.2	**90.3**	**54.2**	**33.3**	**59.3**	**69.9**
64	Shared	84.0	64.2	90.7	97.0	66.3	43.9	72.4	82.5	82.0	64.7	88.3	92.9	58.5	37.8	63.9	73.9
	Shared w/ Group Sparse Loss	31.0	16.0	33.3	43.7	22.0	4.5	23.4	38.1	74.6	53.7	81.7	88.5	47.8	26.8	52.4	64.2
	Shared w/ InfoNCE Loss	87.0	71.7	92.5	96.9	68.5	48.3	74.3	82.8	82.3	66.0	88.2	92.8	58.2	37.0	63.6	73.9
	Independent	**90.5**	**78.9**	**95.0**	**97.7**	71.6	52.4	77.2	85.3	**82.8**	**66.6**	88.4	**93.5**	58.8	38.0	64.1	74.3
	Independent w/ Group Sparse Loss	82.0	63.2	89.0	93.8	56.7	34.5	62.4	73.3	75.3	54.8	82.2	89.0	49.0	28.3	53.6	65.2
	Independent w/ InfoNCE Loss	90.4	78.7	94.9	97.5	**72.5**	**53.3**	**78.3**	**85.8**	82.7	66.1	**88.9**	93.1	**59.2**	**38.1**	**64.7**	**74.8**
128	Shared	90.1	76.9	95.6	97.8	73.7	54.0	79.8	87.2	84.9	70.2	90.1	94.4	61.9	41.8	67.3	76.7
	Shared w/ Group Sparse Loss	31.0	14.4	34.0	44.5	22.8	5.5	24.6	38.4	73.1	51.2	80.0	88.2	44.8	23.4	49.3	61.7
	Shared w/ InfoNCE Loss	92.2	81.4	96.6	98.5	75.3	57.5	80.8	87.7	85.0	69.9	90.5	94.5	62.4	42.1	67.7	77.3
	Independent	93.0	83.7	96.7	98.5	76.1	58.8	81.4	88.1	85.0	**70.4**	90.0	**94.7**	62.4	42.2	67.7	77.3
	Independent w/ Group Sparse Loss	78.8	58.6	85.6	92.2	54.6	32.7	59.6	71.4	74.5	53.6	81.1	88.7	49.7	28.6	54.5	66.0
	Independent w/ InfoNCE Loss	**93.5**	**84.9**	**97.0**	**98.7**	**76.9**	**59.7**	**82.2**	**88.7**	**85.2**	**70.4**	**90.8**	94.5	**62.8**	**42.6**	**68.1**	**77.6**
256	Shared	94.5	86.5	98.0	99.1	79.1	62.4	84.3	90.5	86.6	72.7	**91.9**	95.1	64.9	45.3	70.3	79.1
	Shared w/ Group Sparse Loss	60.9	38.8	66.3	77.7	46.8	24.7	51.6	64.0	71.0	46.2	78.9	87.8	42.3	19.9	46.7	60.2
	Shared w/ InfoNCE Loss	93.7	84.5	97.8	98.8	78.3	61.0	84.0	89.9	86.6	72.9	91.8	**95.2**	64.8	45.2	70.2	79.0
	Independent	95.0	87.4	**98.2**	**99.3**	79.5	62.8	**85.1**	90.6	**86.7**	**73.1**	**91.9**	95.0	65.0	**45.4**	70.3	79.2
	Independent w/ Group Sparse Loss	86.2	71.1	91.7	95.9	64.4	43.0	70.2	80.0	80.0	61.8	86.2	91.9	55.4	34.5	60.6	71.2
	Independent w/ InfoNCE Loss	**95.0**	**87.8**	98.0	99.2	**79.7**	**63.1**	84.9	**91.2**	86.5	73.0	91.6	95.0	**65.0**	45.2	**70.5**	**79.3**
	SigLIP ViT-B/16	95.4	89.0	98.0	99.3	81.0	65.3	86.2	91.6	87.5	74.6	92.3	95.6	66.6	47.2	72.1	80.6

Table 23: Zero-shot classification accuracy (%) across datasets, grouped by sparsity level K, for SigLIP 2 ViT-B/16 on YFCC15M-V2.

		General datasets						Fine-grained datasets							Misc. datasets				
K	SAE	Avg	ImageNet	CIFAR-10	CIFAR-100	Caltech-101	STL-10	Avg	Food-101	CUB	Cars	Aircraft	Pets	Flowers	Avg	DTD	EuroSAT	RESISC45	Country211
8	Shared	74.0	39.4	**91.0**	62.3	79.1	98.3	18.7	45.4	3.2	1.9	1.0	38.8	21.7	26.7	38.3	15.9	**52.0**	**0.8**
	Shared w/ Group Sparse Loss	70.2	33.2	90.8	54.7	74.3	97.8	16.2	40.4	3.2	**2.1**	1.0	22.7	27.5	17.2	26.9	13.1	28.1	0.7
	Shared w/ InfoNCE Loss	**75.7**	**45.2**	89.7	**62.6**	**82.7**	98.1	22.0	56.3	2.9	1.7	**1.5**	36.0	33.8	26.9	**41.5**	18.5	46.8	0.7
	Independent	74.9	42.0	90.7	60.9	82.3	**98.6**	**24.8**	**61.9**	**4.1**	1.6	1.1	**41.7**	38.5	27.8	39.4	23.6	47.6	0.6
	Independent w/ Group Sparse Loss	72.5	36.8	90.9	60.2	76.1	98.4	19.1	51.6	2.9	1.7	1.2	29.0	28.4	26.5	33.1	22.0	50.1	0.7
	Independent w/ InfoNCE Loss	74.8	42.5	90.0	61.3	81.7	98.5	24.1	59.5	3.9	1.9	0.9	36.1	**42.1**	**30.1**	41.0	**29.0**	49.7	0.7
16	Shared	80.3	55.1	92.3	68.1	87.3	**98.7**	31.5	76.3	4.0	2.6	**3.3**	51.4	51.5	33.0	50.9	24.2	**56.1**	0.7
	Shared w/ Group Sparse Loss	74.9	40.3	91.6	61.5	82.3	98.6	16.8	42.1	3.2	1.4	1.0	18.9	34.1	26.4	36.6	25.7	43.0	0.5
	Shared w/ InfoNCE Loss	80.6	54.5	92.4	68.3	89.0	**98.7**	29.0	73.6	3.8	2.6	2.6	58.5	32.8	28.9	47.1	16.8	50.7	0.8
	Independent	81.2	57.3	91.8	67.5	90.8	**98.7**	36.8	79.3	**4.3**	3.1	2.1	70.4	61.6	35.2	**54.4**	31.5	53.9	0.8
	Independent w/ Group Sparse Loss	77.4	46.4	91.7	65.6	84.9	98.6	27.9	59.4	3.0	1.5	1.0	54.2	48.5	33.0	44.3	34.5	52.4	**0.9**
	Independent w/ InfoNCE Loss	**81.8**	**57.8**	**92.5**	**69.1**	**91.1**	98.6	**37.4**	**81.6**	4.2	**3.2**	1.8	**71.4**	**62.4**	**35.9**	53.1	**35.2**	54.7	0.7
32	Shared	82.9	61.0	92.6	70.0	92.3	98.8	38.9	83.4	4.9	2.8	4.2	73.0	65.2	37.3	57.4	32.6	58.4	0.9
	Shared w/ Group Sparse Loss	76.7	45.8	92.3	64.5	82.5	98.6	22.1	61.8	3.6	2.5	3.5	27.7	33.3	26.6	39.8	16.0	50.0	0.7
	Shared w/ InfoNCE Loss	83.1	61.3	92.8	**71.1**	91.6	98.8	38.7	83.4	4.5	2.6	**5.0**	73.1	63.7	36.6	52.5	34.3	58.7	**1.0**
	Independent	83.7	62.8	**93.0**	70.6	93.4	98.7	40.9	85.2	4.9	3.2	2.2	**78.1**	71.9	**38.9**	**57.9**	35.1	**62.0**	0.7
	Independent w/ Group Sparse Loss	78.2	48.5	92.2	65.4	86.5	98.2	27.8	59.6	2.8	2.3	1.4	57.2	43.5	36.2	43.9	**41.4**	58.9	0.8
	Independent w/ InfoNCE Loss	**83.8**	**63.8**	92.3	70.5	**93.5**	**98.9**	**41.4**	**86.3**	**5.0**	**3.4**	3.1	77.2	**73.4**	38.5	56.3	35.6	61.3	0.7
64	Shared	84.3	64.7	93.3	71.3	93.5	**98.9**	43.0	88.1	4.9	3.4	6.7	84.4	70.5	40.9	58.8	41.1	62.7	0.8
	Shared w/ Group Sparse Loss	78.1	48.9	90.2	63.1	89.5	98.6	24.6	59.0	3.6	2.2	1.2	40.5	41.1	28.0	40.1	14.5	56.3	**0.9**
	Shared w/ InfoNCE Loss	84.1	64.2	92.8	71.2	93.5	98.8	42.4	86.8	4.9	2.9	4.4	83.8	71.5	40.5	57.9	37.1	**65.9**	**0.9**
	Independent	**84.9**	**66.2**	**93.4**	72.0	**94.3**	98.8	**44.7**	88.5	**5.1**	3.4	**8.8**	85.9	76.3	41.9	60.2	41.7	64.8	0.8
	Independent w/ Group Sparse Loss	79.2	51.5	92.5	65.2	88.2	98.7	29.9	64.9	3.2	2.2	1.2	59.9	47.9	36.9	46.5	39.1	61.0	0.8
	Independent w/ InfoNCE Loss	84.8	**66.2**	92.9	**72.2**	93.9	98.8	44.2	**88.6**	4.9	**4.2**	4.9	**86.1**	**76.6**	**42.4**	**61.4**	**42.0**	65.3	**0.9**
128	Shared	85.4	67.0	**93.6**	72.7	**94.9**	**98.8**	46.3	89.4	5.4	3.6	9.6	89.4	**80.4**	43.3	62.0	43.4	67.0	0.9
	Shared w/ Group Sparse Loss	80.3	55.2	91.2	64.8	91.8	98.5	32.3	69.9	4.4	2.1	1.0	68.9	47.3	35.7	44.4	37.6	59.9	**1.0**
	Shared w/ InfoNCE Loss	85.3	67.0	93.1	**72.9**	94.7	**98.8**	45.6	89.4	5.1	3.9	7.5	90.1	77.9	43.4	62.4	43.3	67.1	0.9
	Independent	**85.5**	**68.0**	93.3	72.7	94.8	**98.8**	46.7	**89.6**	**5.6**	4.0	9.7	**91.5**	79.8	44.3	**62.6**	45.3	**68.2**	0.9
	Independent w/ Group Sparse Loss	82.5	59.8	92.5	68.7	92.7	98.7	39.9	77.0	4.3	**4.1**	5.2	81.6	67.3	39.4	53.5	38.9	64.3	0.9
	Independent w/ InfoNCE Loss	85.3	67.7	93.4	72.8	94.1	98.7	**46.7**	89.3	5.0	3.9	**11.5**	91.3	79.4	**44.7**	61.8	**49.5**	66.5	0.9
256	Shared	**86.1**	69.1	**93.6**	**73.8**	95.3	**98.8**	48.1	90.4	5.4	4.0	**12.1**	93.9	**82.8**	44.5	63.5	43.4	**70.1**	**1.0**
	Shared w/ Group Sparse Loss	83.3	61.1	93.2	69.0	94.7	98.6	40.7	82.8	3.9	**4.2**	3.2	84.2	65.9	42.2	57.2	45.5	65.3	0.9
	Shared w/ InfoNCE Loss	**86.1**	68.9	**93.6**	**73.8**	**95.5**	**98.8**	**48.1**	90.4	5.3	4.1	12.0	**94.3**	82.7	43.6	63.9	40.7	69.0	**1.0**
	Independent	86.0	**69.2**	**93.6**	73.6	95.0	**98.8**	48.1	**90.6**	**5.5**	3.7	11.8	**94.3**	82.6	44.9	63.9	45.8	69.1	**1.0**
	Independent w/ Group Sparse Loss	84.0	63.7	92.9	70.9	93.5	**98.8**	43.6	84.7	5.1	**4.2**	6.8	87.7	72.9	42.0	56.8	43.0	67.5	0.8
	Independent w/ InfoNCE Loss	85.9	69.0	93.5	73.4	95.0	**98.8**	47.9	90.5	5.4	3.8	11.1	94.0	82.5	**45.1**	**64.0**	**46.6**	69.1	0.9
	SigLIP 2 ViT-B/16	86.2	70.0	93.6	73.8	95.0	98.8	48.9	90.9	5.8	4.1	13.6	94.8	84.2	45.6	64.6	47.1	69.6	1.0

Table 24: Zero-shot retrieval performance across datasets, grouped by sparsity level K, for SigLIP 2 ViT-B/16 on YFCC15M-V2.

		Text Retrieval								Image Retrieval							
		Flickr30k				COCO				Flickr30k				COCO			
K	SAE	Avg	R@1	R@5	R@10	Avg	R@1	R@5	R@10	Avg	R@1	R@5	R@10	Avg	R@1	R@5	R@10
8	Shared	17.1	6.6	18.1	26.5	13.5	4.4	13.7	22.5	53.3	31.3	58.6	70.1	31.6	15.3	34.3	45.2
	Shared w/ Group Sparse Loss	12.2	2.9	12.6	21.0	10.7	3.3	10.6	18.1	43.7	22.3	48.1	60.7	22.9	9.7	24.3	34.6
	Shared w/ InfoNCE Loss	15.3	7.5	15.1	23.4	10.1	3.9	10.2	16.2	55.3	**33.9**	60.6	71.3	33.5	16.0	36.6	48.0
	Independent	**61.6**	**41.2**	66.8	**76.8**	**39.9**	**22.0**	43.3	54.4	55.2	33.7	60.5	71.4	**36.4**	18.6	**39.8**	50.7
	Independent w/ Group Sparse Loss	55.9	32.7	61.7	73.2	34.0	17.5	36.2	48.3	53.7	31.0	59.2	70.8	33.9	16.7	36.7	48.2
	Independent w/ InfoNCE Loss	61.5	41.0	**66.9**	76.5	**39.9**	21.8	**43.4**	**54.5**	**56.2**	33.6	**62.1**	**72.8**	36.3	**18.8**	39.4	**50.8**
16	Shared	20.4	9.3	21.3	30.7	21.3	7.6	21.7	34.5	70.7	50.5	76.8	84.9	47.8	27.4	52.5	63.6
	Shared w/ Group Sparse Loss	15.3	7.4	16.0	22.6	13.1	4.4	13.7	21.2	60.6	38.0	66.7	77.2	36.6	18.6	39.9	51.3
	Shared w/ InfoNCE Loss	12.8	6.3	13.7	18.4	10.2	4.0	10.8	15.9	70.0	49.7	76.1	84.3	45.3	25.3	49.7	61.0
	Independent	**78.5**	**60.0**	**84.0**	91.4	55.6	35.1	**60.8**	70.9	**71.9**	**52.4**	**77.5**	85.7	**49.0**	**28.6**	**53.8**	64.5
	Independent w/ Group Sparse Loss	73.3	50.6	80.2	89.2	49.6	29.1	54.1	65.7	66.8	44.7	72.9	82.9	44.5	24.4	48.7	60.3
	Independent w/ InfoNCE Loss	78.1	58.9	83.6	**91.7**	**55.9**	**35.9**	**60.8**	**71.0**	71.5	51.1	**77.5**	**86.0**	48.9	**28.6**	53.3	**64.9**
32	Shared	49.1	23.0	53.6	70.6	43.8	20.4	47.7	63.4	77.9	59.7	83.8	90.1	55.3	34.7	60.4	70.7
	Shared w/ Group Sparse Loss	22.0	11.4	23.7	31.0	16.3	6.9	17.2	24.8	70.0	48.7	76.4	84.9	43.9	24.0	48.3	59.3
	Shared w/ InfoNCE Loss	23.7	9.7	24.4	36.9	25.2	8.9	26.3	40.4	77.4	59.2	83.4	89.7	54.5	33.6	59.7	70.3
	Independent	**86.7**	**72.1**	**91.7**	96.2	65.0	44.7	70.6	79.7	**78.6**	**60.7**	**84.5**	**90.5**	55.6	34.7	60.7	71.3
	Independent w/ Group Sparse Loss	80.0	61.1	86.2	92.6	56.8	35.9	61.6	72.9	71.8	51.1	77.7	86.5	47.6	27.2	52.3	63.4
	Independent w/ InfoNCE Loss	85.7	69.3	91.5	**96.3**	**65.7**	**45.8**	**71.1**	**80.3**	77.8	59.4	83.8	90.1	**56.0**	**35.4**	**61.0**	**71.6**
64	Shared	67.0	36.2	74.6	90.3	55.0	28.7	60.3	76.0	82.1	66.4	87.3	92.6	60.1	39.8	65.4	75.2
	Shared w/ Group Sparse Loss	24.7	13.2	27.2	33.7	17.8	7.8	19.2	26.4	71.6	51.3	77.8	85.7	46.3	26.0	50.8	62.1
	Shared w/ InfoNCE Loss	74.1	46.0	83.3	93.0	53.3	25.6	59.4	75.0	81.9	65.4	87.7	92.6	60.0	39.5	65.3	75.3
	Independent	91.0	79.6	95.6	97.9	71.3	52.1	76.9	84.8	**82.6**	**66.9**	**87.9**	**92.9**	60.7	**40.3**	65.8	**75.9**
	Independent w/ Group Sparse Loss	82.4	62.8	89.4	95.0	57.5	36.2	63.0	73.2	72.6	51.8	79.0	86.9	48.9	28.1	53.8	64.7
	Independent w/ InfoNCE Loss	**92.0**	**80.9**	**96.5**	**98.5**	**72.0**	**52.4**	**77.8**	**85.7**	82.4	66.4	87.8	**92.9**	**60.8**	**40.3**	**66.3**	75.8
128	Shared	87.9	72.0	94.0	97.8	67.1	43.3	73.7	84.3	84.3	69.7	89.5	93.8	63.5	43.7	68.8	78.0
	Shared w/ Group Sparse Loss	21.5	12.8	23.4	28.3	11.8	6.0	12.5	16.9	72.4	52.1	78.6	86.4	46.5	25.9	50.9	62.6
	Shared w/ InfoNCE Loss	90.6	77.0	96.5	98.4	70.7	48.0	77.4	86.7	84.2	69.2	89.6	**93.9**	63.6	**43.8**	69.0	78.1
	Independent	93.7	84.8	**97.6**	98.6	75.6	57.6	80.8	**88.4**	**84.6**	**70.3**	**89.8**	93.6	**63.8**	43.6	**69.4**	**78.4**
	Independent w/ Group Sparse Loss	86.4	69.7	92.1	97.4	65.2	45.1	70.6	79.9	76.5	58.1	82.6	88.9	53.0	32.4	58.0	68.6
	Independent w/ InfoNCE Loss	**94.0**	**85.9**	97.5	**98.7**	**76.2**	**58.3**	**82.0**	**88.4**	83.7	69.2	88.8	93.1	63.0	43.0	68.3	77.6
256	Shared	94.2	85.2	**98.0**	**99.4**	78.0	60.1	83.6	90.2	**85.7**	**72.3**	**90.4**	94.3	**65.7**	**46.2**	71.0	**79.9**
	Shared w/ Group Sparse Loss	35.4	22.0	38.4	45.9	21.0	10.4	22.7	30.0	74.9	56.6	80.8	87.3	50.1	28.9	55.3	66.1
	Shared w/ InfoNCE Loss	94.2	86.1	97.5	98.9	78.1	59.9	83.7	**90.8**	85.6	71.9	**90.4**	**94.4**	64.9	45.3	70.5	79.0
	Independent	95.0	87.7	**98.0**	99.3	**79.6**	63.0	**85.1**	**90.8**	85.5	71.7	**90.4**	**94.4**	**65.7**	**46.2**	**71.4**	79.5
	Independent w/ Group Sparse Loss	90.6	78.3	95.2	98.3	70.3	50.6	76.0	84.3	78.3	60.9	84.1	89.9	56.5	36.0	61.6	71.9
	Independent w/ InfoNCE Loss	**95.3**	**88.6**	97.9	99.3	79.4	**63.2**	84.5	90.6	85.6	72.1	90.3	**94.4**	65.4	45.8	70.9	79.4
	SigLIP 2 ViT-B/16	95.7	89.6	98.3	99.3	80.3	64.3	85.6	91.1	86.7	73.8	91.3	95.0	67.1	48.0	72.4	80.8



(a) Feature "cooked chicken"



(b) Feature "whirlpool"

Figure 27: Top-activating examples for both modalities for the features (a) "cooked chicken" and (b) "whirlpool" from independent SAE ($K = 32$) with InfoNCE Loss for SigLIP 2 ViT-B/16 on YFCC15M-V2. For each sample, we report the activation value for the corresponding feature.



(a) Feature "beagle"



(b) Feature "people"

Figure 28: Top-activating examples for both modalities for the features (a) "beagle" and (b) "people" from independent SAE ($K = 32$) with InfoNCE Loss for SigLIP 2 ViT-B/16 on YFCC15M-V2. For each sample, we report the activation value for the corresponding feature. Note that the concept "people" appears as text within the images (zoom for detail)

G AUTOINTERPRETABILITY SCORES

The autointerpretability scores (Paulo et al., 2025; Cunningham et al., 2024; Bills et al., 2023; Minder et al., 2025) are designed to evaluate the interpretability of SAE features for the textual domain without relying on human annotations. We employ the Detection and Simulation scores. A Large Language Model (LLM) is prompted to generate a natural language description for each feature based on the samples that activate it. Then, these descriptions are used by another LLM to infer if an unseen sample activates the same feature, or to what extent it activates it. In the first case, the scores are computed as the accuracy of the LLM in correctly classifying whether a sample activates the feature or not, while in the second case, the scores are computed as the correlation between the LLM's predictions and the actual activations.

For generating the explanations, we use the prompt shown in G.1. For the Simulation and Detection scores, we use the prompts shown in G.2 and G.3, respectively. All prompts contain instructions followed by few-shot examples. The generator sees 10 samples from each activation tercile for a given feature, and the evaluator sees 10 samples from each activation decile for the same feature, with other 100 non-activating samples. The examples are randomly selected from the evaluation set and are different for the generator and evaluator. We use Llama 3 8B (Grattafiori et al., 2024) as the LLM in both pipelines.

In Tab. 25, we report the results for CLIP ViT-B/32 on both CC3M and YFCC15M-V2 datasets. As expected, highly sparse codes produce more specific, semantic features that an LLM can easily describe and subsequently simulate or detect. The independent SAEs, particularly those with the additional co-activation loss, consistently demonstrate superior autointerpretability performance compared to the shared SAE variants. This advantage is maintained across almost all sparsity regimes for both datasets, reinforcing the hypothesis that modality-specific dictionaries isolate cleaner semantic concepts than shared, modality-agnostic ones.

Table 25: Autointerpretability results grouped by sparsity level K for CLIP-ViT-B/32 on CC3M (left) and YFCC15M-V2 (right). For each SAE variant, we report the mean simulation score and mean accuracy score, both with standard deviation. Higher values are better for all metrics. The best performant variant for each K and metric is highlighted in **bold**.

K	SAE	Simulation ↑	Accuracy ↑
8	Shared	0.653 ± 0.210	80.8 ± 11.0
	Shared w/ InfoNCE Loss	**0.669 ± 0.177**	81.9 ± 10.2
	Independent	0.660 ± 0.184	82.0 ± 10.5
	Independent w/ InfoNCE Loss	0.666 ± 0.193	**82.2 ± 10.6**
16	Shared	0.588 ± 0.221	77.5 ± 11.3
	Shared w/ InfoNCE Loss	0.623 ± 0.209	79.3 ± 11.1
	Independent	0.601 ± 0.202	79.6 ± 11.0
	Independent w/ InfoNCE Loss	**0.626 ± 0.203**	**80.8 ± 11.2**
32	Shared	0.540 ± 0.226	74.9 ± 11.1
	Shared w/ InfoNCE Loss	0.573 ± 0.215	76.7 ± 11.1
	Independent	0.558 ± 0.211	77.1 ± 11.3
	Independent w/ InfoNCE Loss	**0.584 ± 0.205**	**78.6 ± 11.3**
64	Shared	0.465 ± 0.256	70.7 ± 11.7
	Shared w/ InfoNCE Loss	**0.511 ± 0.241**	73.5 ± 11.9
	Independent	0.466 ± 0.254	72.0 ± 12.3
	Independent w/ InfoNCE Loss	0.509 ± 0.234	**74.6 ± 12.0**
128	Shared	0.412 ± 0.292	68.1 ± 13.0
	Shared w/ InfoNCE Loss	**0.468 ± 0.280**	71.8 ± 13.5
	Independent	0.421 ± 0.287	69.4 ± 13.4
	Independent w/ InfoNCE Loss	0.451 ± 0.279	**71.9 ± 13.8**
256	Shared	0.312 ± 0.366	65.2 ± 17.9
	Shared w/ InfoNCE Loss	0.318 ± 0.368	65.6 ± 18.0
	Independent	0.365 ± 0.376	68.0 ± 18.9
	Independent w/ InfoNCE Loss	**0.376 ± 0.363**	**68.7 ± 18.3**

(a) CC3M dataset

K	SAE	Simulation ↑	Accuracy ↑
8	Shared	0.618 ± 0.228	80.3 ± 11.3
	Shared w/ InfoNCE Loss	0.636 ± 0.220	**81.6 ± 11.2**
	Independent	0.628 ± 0.219	81.2 ± 11.1
	Independent w/ InfoNCE Loss	**0.643 ± 0.215**	81.6 ± 11.3
16	Shared	0.567 ± 0.220	78.1 ± 11.0
	Shared w/ InfoNCE Loss	0.587 ± 0.218	78.9 ± 10.9
	Independent	0.574 ± 0.212	78.7 ± 10.9
	Independent w/ InfoNCE Loss	**0.600 ± 0.217**	**79.6 ± 11.3**
32	Shared	0.495 ± 0.237	74.3 ± 11.5
	Shared w/ InfoNCE Loss	0.522 ± 0.230	75.7 ± 11.4
	Independent	0.508 ± 0.226	75.3 ± 11.0
	Independent w/ InfoNCE Loss	**0.534 ± 0.227**	**76.4 ± 11.3**
64	Shared	0.418 ± 0.263	70.0 ± 12.1
	Shared w/ InfoNCE Loss	**0.467 ± 0.250**	**72.7 ± 12.1**
	Independent	0.434 ± 0.255	70.9 ± 12.0
	Independent w/ InfoNCE Loss	0.455 ± 0.248	72.4 ± 12.1
128	Shared	0.417 ± 0.286	70.4 ± 13.4
	Shared w/ InfoNCE Loss	**0.461 ± 0.284**	**72.9 ± 14.1**
	Independent	0.415 ± 0.285	70.0 ± 13.3
	Independent w/ InfoNCE Loss	0.450 ± 0.276	72.3 ± 13.6
256	Shared	0.391 ± 0.355	71.4 ± 17.6
	Shared w/ InfoNCE Loss	0.344 ± 0.367	69.1 ± 18.5
	Independent	**0.483 ± 0.381**	**75.4 ± 19.9**
	Independent w/ InfoNCE Loss	0.425 ± 0.387	73.1 ± 20.4

(b) YFCC15M-V2 dataset

G.1: Prompt for Explanation Generation

```
[SYSTEM - START CONTENT]
```

```
You are a meticulous AI researcher conducting an important investigation into patterns
↪  found in language. Your task is to analyze text and provide an explanation that
↪  thoroughly encapsulates possible patterns found in it.
Guidelines:

You will be given a list of text examples. How important each text example is for the
↪  behavior is listed after each example in parentheses.

- Try to produce a concise final description. Simply describe the texts that are common
↪  in the examples, and what patterns you found.
- If the examples are uninformative, you don't need to mention them. Don't focus on
↪  giving examples of important texts, but try to summarize the patterns found in the
↪  examples.
- Do not make lists of possible explanations. Keep your explanations short and concise.
- The last line of your response must be the formatted explanation, using
↪  [EXPLANATION]:



[SYSTEM - END CONTENT]
[USER - START CONTENT]
Example 1:  a fluffy white cat sitting on a windowsill
Activation: (9)
Example 2:  a golden retriever dog playing fetch in a park
Activation: (10)
Example 3:  a person walking on a beach at sunset
Activation: (2)
Example 4:  a small puppy with floppy ears
Activation: (8)
Example 5:  two cats and a dog together in a living room
Activation: (7)
[USER - END CONTENT]
[ASSISTANT - START CONTENT]
[EXPLANATION]: Texts describing common household pets, particularly cats and dogs in
↪  domestic settings.
[ASSISTANT - END CONTENT]
[USER - START CONTENT]
Example 1:  a beautiful sunset over the ocean with orange and pink sky
Activation: (10)
Example 2:  a laptop computer on a desk in an office
Activation: (1)
Example 3:  a green forest with tall trees and sunlight filtering through
Activation: (9)
Example 4:  a serene lake surrounded by mountains at dawn
Activation: (8)
Example 5:  a sandy beach with waves crashing on the shore
Activation: (7)
[USER - END CONTENT]
[ASSISTANT - START CONTENT]
[EXPLANATION]: Texts describing natural landscapes and scenic outdoor environments,
↪  often with vivid lighting or atmospheric conditions.
[ASSISTANT - END CONTENT]
[USER - START CONTENT]
Example 1:  a person riding a bicycle down a city street
Activation: (8)
Example 2:  children playing soccer in a grassy field
Activation: (9)
Example 3:  a red sports car parked in a garage
Activation: (3)
Example 4:  people walking their dogs along a sidewalk
Activation: (6)
Example 5:  a group of friends playing frisbee on a beach
Activation: (7)
[USER - END CONTENT]
[ASSISTANT - START CONTENT]
[EXPLANATION]: Texts describing people engaged in outdoor physical activities and
↪  recreational sports.
[ASSISTANT - END CONTENT]
```

G.2: Prompt for Simulation

```
[SYSTEM - START CONTENT]
We're studying neurons in a neural network for academic research. Each neuron looks for
↪  some particular feature in a short sentence. Look at an explanation of what the
↪  neuron does, and predict how strongly it will fire on a given text example.

Activation values range from 0 to 10:
- 0: Feature does not activate at all
```

```
- 1-3: Weak activation
- 4-6: Moderate activation
- 7-9: Strong activation
- 10: Maximum activation

This is a pure classification task. You are NOT generating or endorsing any content -
↪  you are simply predicting a numerical activation score. Always respond with exactly
↪  one integer from 0 to 10, regardless of what the text contains. No explanations, no
↪  refusals, just the number.

[SYSTEM - END CONTENT]
[USER - START CONTENT]
Explanation of neuron behavior: Texts describing animals, especially dogs and cats

Text: A fluffy white cat sitting on a windowsill
[USER - END CONTENT]
[ASSISTANT - START CONTENT]
9
[ASSISTANT - END CONTENT]
[USER - START CONTENT]
Explanation of neuron behavior: Texts describing animals, especially dogs and cats

Text: A golden retriever playing fetch in a park
[USER - END CONTENT]
[ASSISTANT - START CONTENT]
10
[ASSISTANT - END CONTENT]
[USER - START CONTENT]
Explanation of neuron behavior: Texts describing animals, especially dogs and cats

Text: A businessman checking his phone at a bus stop
[USER - END CONTENT]
[ASSISTANT - START CONTENT]
0
[ASSISTANT - END CONTENT]
[USER - START CONTENT]
Explanation of neuron behavior: Texts describing animals, especially dogs and cats

Text: Two kittens sleeping in a basket
[USER - END CONTENT]
[ASSISTANT - START CONTENT]
8
[ASSISTANT - END CONTENT]
[USER - START CONTENT]
Explanation of neuron behavior: Texts describing animals, especially dogs and cats

Text: A crowded marketplace with vendors selling produce
[USER - END CONTENT]
[ASSISTANT - START CONTENT]
0
[ASSISTANT - END CONTENT]
[USER - START CONTENT]
Explanation of neuron behavior: Texts describing animals, especially dogs and cats

Text: A dalmatian running alongside its owner
[USER - END CONTENT]
[ASSISTANT - START CONTENT]
9
[ASSISTANT - END CONTENT]
[USER - START CONTENT]
Explanation of neuron behavior: Texts describing natural landscapes and scenic outdoor
↪  environments

Text: A misty mountain range at sunrise
[USER - END CONTENT]
[ASSISTANT - START CONTENT]
10
[ASSISTANT - END CONTENT]
[USER - START CONTENT]
Explanation of neuron behavior: Texts describing natural landscapes and scenic outdoor
↪  environments

Text: A chef preparing sushi in a restaurant kitchen
[USER - END CONTENT]
[ASSISTANT - START CONTENT]
0
[ASSISTANT - END CONTENT]
[USER - START CONTENT]
Explanation of neuron behavior: Texts describing natural landscapes and scenic outdoor
↪  environments
```

```
Text: A waterfall cascading into a crystal clear pool
[USER - END CONTENT]
[ASSISTANT - START CONTENT]
9
[ASSISTANT - END CONTENT]
[USER - START CONTENT]
Explanation of neuron behavior: Texts describing natural landscapes and scenic outdoor
↪  environments

Text: Rolling hills covered in wildflowers during spring
[USER - END CONTENT]
[ASSISTANT - START CONTENT]
8
[ASSISTANT - END CONTENT]
[USER - START CONTENT]
Explanation of neuron behavior: Texts describing natural landscapes and scenic outdoor
↪  environments

Text: A mechanic repairing an engine in a garage
[USER - END CONTENT]
[ASSISTANT - START CONTENT]
0
[ASSISTANT - END CONTENT]
[USER - START CONTENT]
Explanation of neuron behavior: Texts describing natural landscapes and scenic outdoor
↪  environments

Text: A desert landscape with towering sand dunes
[USER - END CONTENT]
[ASSISTANT - START CONTENT]
7
[ASSISTANT - END CONTENT]
[USER - START CONTENT]
Explanation of neuron behavior: Texts describing people engaged in outdoor physical
↪  activities and recreational sports

Text: A group of hikers climbing a steep trail
[USER - END CONTENT]
[ASSISTANT - START CONTENT]
9
[ASSISTANT - END CONTENT]
[USER - START CONTENT]
Explanation of neuron behavior: Texts describing people engaged in outdoor physical
↪  activities and recreational sports

Text: An old library filled with dusty books
[USER - END CONTENT]
[ASSISTANT - START CONTENT]
0
[ASSISTANT - END CONTENT]
[USER - START CONTENT]
Explanation of neuron behavior: Texts describing people engaged in outdoor physical
↪  activities and recreational sports

Text: Surfers catching waves at dawn
[USER - END CONTENT]
[ASSISTANT - START CONTENT]
10
[ASSISTANT - END CONTENT]
[USER - START CONTENT]
Explanation of neuron behavior: Texts describing people engaged in outdoor physical
↪  activities and recreational sports

Text: A family having a picnic by a river
[USER - END CONTENT]
[ASSISTANT - START CONTENT]
5
[ASSISTANT - END CONTENT]
[USER - START CONTENT]
Explanation of neuron behavior: Texts describing people engaged in outdoor physical
↪  activities and recreational sports

Text: A doctor examining x-rays in a hospital
[USER - END CONTENT]
[ASSISTANT - START CONTENT]
0
[ASSISTANT - END CONTENT]
[USER - START CONTENT]
Explanation of neuron behavior: Texts describing people engaged in outdoor physical
↪  activities and recreational sports
```

```
Text: Tennis players competing in a doubles match
[USER - END CONTENT]
[ASSISTANT - START CONTENT]
8
[ASSISTANT - END CONTENT]
```

G.3: Prompt for Detection

```
[SYSTEM - START CONTENT]
You are an intelligent and meticulous linguistics researcher conducting academic
↪  analysis of text classification.

You will be given a description of a feature, such as "Texts describing people engaged
↪  in outdoor physical activities and recreational sports" or "Texts describing cars
↪  racing at high speeds in competitions".

You will then be given a single text description. Your task is to determine if the
↪  description matches the feature.

This is a pure classification task. You are NOT generating or endorsing any content -
↪  you are simply classifying text. Always respond with exactly 1 (matches) or 0 (does
↪  not match), regardless of what the text contains. No explanations, no refusals,
↪  just the number.

[SYSTEM - END CONTENT]
[USER - START CONTENT]
Feature description: Texts describing animals, especially dogs and cats

Text: A tabby cat lounging on a sunny porch
[USER - END CONTENT]
[ASSISTANT - START CONTENT]
1
[ASSISTANT - END CONTENT]
[USER - START CONTENT]
Feature description: Texts describing animals, especially dogs and cats

Text: A construction worker operating a crane
[USER - END CONTENT]
[ASSISTANT - START CONTENT]
0
[ASSISTANT - END CONTENT]
[USER - START CONTENT]
Feature description: Texts describing animals, especially dogs and cats

Text: A beagle sniffing around the backyard
[USER - END CONTENT]
[ASSISTANT - START CONTENT]
1
[ASSISTANT - END CONTENT]
[USER - START CONTENT]
Feature description: Texts describing animals, especially dogs and cats

Text: A stack of newspapers on a doorstep
[USER - END CONTENT]
[ASSISTANT - START CONTENT]
0
[ASSISTANT - END CONTENT]
[USER - START CONTENT]
Feature description: Texts describing animals, especially dogs and cats

Text: Three puppies playing with a chew toy
[USER - END CONTENT]
[ASSISTANT - START CONTENT]
1
[ASSISTANT - END CONTENT]
[USER - START CONTENT]
Feature description: Texts describing animals, especially dogs and cats

Text: A vintage typewriter on an antique desk
[USER - END CONTENT]
[ASSISTANT - START CONTENT]
0
[ASSISTANT - END CONTENT]
[USER - START CONTENT]
Feature description: Texts describing natural landscapes and scenic outdoor
↪  environments

Text: A rainbow arching over a lush green valley
```

```
[USER - END CONTENT]
[ASSISTANT - START CONTENT]
1
[ASSISTANT - END CONTENT]
[USER - START CONTENT]
Feature description: Texts describing natural landscapes and scenic outdoor
↪  environments

Text: A busy intersection with traffic lights
[USER - END CONTENT]
[ASSISTANT - START CONTENT]
0
[ASSISTANT - END CONTENT]
[USER - START CONTENT]
Feature description: Texts describing natural landscapes and scenic outdoor
↪  environments

Text: Snow-capped peaks reflected in a still lake
[USER - END CONTENT]
[ASSISTANT - START CONTENT]
1
[ASSISTANT - END CONTENT]
[USER - START CONTENT]
Feature description: Texts describing natural landscapes and scenic outdoor
↪  environments

Text: A warehouse full of shipping containers
[USER - END CONTENT]
[ASSISTANT - START CONTENT]
0
[ASSISTANT - END CONTENT]
[USER - START CONTENT]
Feature description: Texts describing natural landscapes and scenic outdoor
↪  environments

Text: An autumn forest with golden and red leaves
[USER - END CONTENT]
[ASSISTANT - START CONTENT]
1
[ASSISTANT - END CONTENT]
[USER - START CONTENT]
Feature description: Texts describing natural landscapes and scenic outdoor
↪  environments

Text: A dentist office waiting room with magazines
[USER - END CONTENT]
[ASSISTANT - START CONTENT]
0
[ASSISTANT - END CONTENT]
[USER - START CONTENT]
Feature description: Texts describing people engaged in outdoor physical activities and
↪  recreational sports

Text: Kayakers navigating through river rapids
[USER - END CONTENT]
[ASSISTANT - START CONTENT]
1
[ASSISTANT - END CONTENT]
[USER - START CONTENT]
Feature description: Texts describing people engaged in outdoor physical activities and
↪  recreational sports

Text: A barista making coffee at a cafe counter
[USER - END CONTENT]
[ASSISTANT - START CONTENT]
0
[ASSISTANT - END CONTENT]
[USER - START CONTENT]
Feature description: Texts describing people engaged in outdoor physical activities and
↪  recreational sports

Text: Rock climbers scaling a cliff face
[USER - END CONTENT]
[ASSISTANT - START CONTENT]
1
[ASSISTANT - END CONTENT]
[USER - START CONTENT]
Feature description: Texts describing people engaged in outdoor physical activities and
↪  recreational sports

Text: An accountant reviewing spreadsheets
```

```
[USER - END CONTENT]
[ASSISTANT - START CONTENT]
0
[ASSISTANT - END CONTENT]
[USER - START CONTENT]
Feature description: Texts describing people engaged in outdoor physical activities and
↪  recreational sports

Text: Joggers running along a coastal path at sunrise
[USER - END CONTENT]
[ASSISTANT - START CONTENT]
1
[ASSISTANT - END CONTENT]
[USER - START CONTENT]
Feature description: Texts describing people engaged in outdoor physical activities and
↪  recreational sports

Text: A musician tuning a guitar backstage
[USER - END CONTENT]
[ASSISTANT - START CONTENT]
0
[ASSISTANT - END CONTENT]
```