Coarse-to-Fine: A Reproducibility Study of Hierarchical Sparse Representations for Multimodal Interpretation

Anonymous authors
Paper under double-blind review

Abstract

Sparse Autoencoders (SAEs) aim to improve interpretability through disentangled features in its latent code. Zaigrajew et al. (2025) build upon SAEs by proposing Matryoshka SAEs (MSAEs), which learn hierarchical features at increasing granularities and show improved performance over existing SAE models. Furthermore, the authors show that MSAEs can be applied to concrete problems by interpreting CLIP representations. Though our work validates their claims, our results slightly differ from those of the original work. We also identify a population of always-on neurons specific to MSAE and show that a targeted L_1 penalty removes them at a small cost to reconstruction. We further introduce a concept saliency method that localizes MSAE features within an image, complementing the model's image-level concept decomposition with a spatial, pixel-level account.

1 Introduction

Multimodal models, in particular Contrastive Language-Image Pre-Training (CLIP, Radford et al. (2021)), have seen widespread adoption. CLIP learns to represent text and images in a single embedding space. These embeddings have proven useful in modern applications such as visual search (e.g. Khainas et al., 2025) and medical image classification (e.g. Li et al., 2026). Despite success, interpreting CLIP embeddings remains a significant challenge. The high dimensionality of the embedding space and the complexity of contrastive learning make it difficult to understand how features are encoded (Zaigrajew et al., 2025). Sparse Autoencoders (SAEs) have aimed to tackle this challenge by representing features in a disentangled manner through self-supervised learning (Bricken et al., 2023; Sharkey et al., 2022; Cunningham et al., 2023; Rao et al., 2024). Since individual neurons often represent a mixture of concepts, training SAEs on CLIP embeddings poses a trade-off between sparsity (with the goal of disentangling concepts) and faithful reconstruction (Bricken et al., 2023; Gao et al., 2024). This trade-off is further complicated by dead neurons that increase sparsity and reduce representation bandwidth without increased disentanglement (Bricken et al., 2023).

Previous efforts have enforced sparsity through L_1 regularization (ReLU SAE, Bricken et al. (2023)), which may cause activation shrinkage. Alternatively, rigid sparsity constraints can be imposed on the number of active features (TopK SAE, Makhzani & Frey (2013); Gao et al. (2024)), which could be relaxed from sample to batch level for more flexible sparsity (BatchTopK SAE, Bussmann et al. (2024)). Matryoshka SAE (MSAE, Bussmann et al. (2025); Zaigrajew et al. (2025)) addresses these limitations by learning hierarchical sparse representations inspired by Matryoshka representation learning (Kusupati et al., 2022).

Learning features at multiple granular levels allows MSAEs to relax the rigid constraints of TopK selection while mitigating activation shrinkage.

This study evaluates the work of Zaigrajew et al. (2025). Particular attention is given to performance, hierarchical sparsity structure and its applications. Beyond reproduction, we investigate the presence of always-on neurons in MSAEs through a UMAP-based analysis (McInnes et al., 2020) and show that a targeted L_1 regularization strategy removes the corresponding high-activation-frequency cluster. In addition, we adapt the Concept Visualization framework of Giulivi & Boracchi (2024) to the MSAE, producing per-concept saliency maps that show *where* in an image a learned feature is active.

2 Scope of reproducibility

We identified and investigated the following three claims from Zaigrajew et al. (2025). For further reference, we label them **C1**, **C2** and **C3** respectively.

- **C1: Performance.** MSAE efficiently achieves flexible sparsity, improving balance between reconstruction and sparsity while maintaining representation bandwidth and alignment with the input across CLIP modalities.
- **C2: Hierarchy.** MSAE learns hierarchical features from coarse-to-fine as representation sparsity decreases.
- **C3: Application.** MSAE can be applied to concrete practical problems, in particular identification and validation of interpretable concepts in CLIP multimodal embedding space. Applications include concept-based similarity search and bias analysis in downstream tasks.

3 Method

This section describes the models, datasets and hyperparameters used in the experimental setup combined with details of the experiments. In addition, we propose extensions that build upon the original work.

3.1 Model descriptions

Matryoshka SAEs. Zaigrajew et al. (2025) propose MSAE, a model learning a hierarchy of sparse representations simultaneously. Let $\mathcal{D} = \{\mathrm{CLIP}(\mathbf{u}_j)\}_{j=1}^{N}$ be a set of n-dimensional embeddings extracted by a frozen CLIP encoder over N inputs (Section 3.2). MSAEs decompose CLIP activations $\mathbf{x} \sim \mathcal{D}$, with $\mathbf{x} \in \mathbb{R}^n$, across multiple TopK operations with increasing k values up to latent dimensionality d. The set of granularities is defined as: $\{k_i\}_{i=1}^{h}$ where h is the number of granularity levels, $k_i = 2^{\ell_i}$, $\ell_1 < \ldots < \ell_h$, and $k_h \leq d$. For each granularity level i the latent code $\mathbf{z}_i \in \mathbb{R}^d$ and pre-activation $\mathbf{a} \in \mathbb{R}^d$ are defined as:

$$\mathbf{a} \doteq \boldsymbol{W}_{\mathrm{enc}}^{\top}(\mathbf{x} - \boldsymbol{b}_{\mathrm{pre}}) + \boldsymbol{b}_{\mathrm{enc}}, \qquad \mathbf{z}_i \doteq \mathrm{ReLU}\big(\mathrm{TopK}_i(\mathbf{a})\big). \tag{1}$$

The encoder matrix $\boldsymbol{W}_{\mathrm{enc}} \in \mathbb{R}^{n \times d}$, encoder bias $\boldsymbol{b}_{\mathrm{enc}} \in \mathbb{R}^d$ and preprocessing bias $\boldsymbol{b}_{\mathrm{pre}} \in \mathbb{R}^n$ are the learnable parameters for the encoder-part of this architecture. After $\mathrm{TopK}_i(\cdot)$ operation $d - k_i$ entries are zero-valued (masked), sparsifying $\mathbf{z}_i$.

Hierarchical property. Let $\mathcal{I}_i(\mathbf{a}) \subseteq \{1, \ldots, d\}$ denote the index set of the k_i largest coordinates of the pre-activation $\mathbf{a}$. The Matryoshka constraint requires these selected index sets to be nested across granularities: $\mathcal{I}_1(\mathbf{a}) \subseteq \mathcal{I}_2(\mathbf{a}) \subseteq \cdots \subseteq \mathcal{I}_h(\mathbf{a})$. Zaigrajew et al. (2025) argue that this nesting enforces consistency of the active feature dimensions as the sparsity moves from coarse to fine granularities: small k_i yields a coarse code, while larger k_i introduces finer-grained features, analogous to nested Matryoshka dolls.

Reconstruction and objective. Given a learnable decoder matrix $\boldsymbol{W}_{\mathrm{dec}} \in \mathbb{R}^{d \times n}$, the reconstructed activation $\hat{\mathbf{x}}_i \in \mathbb{R}^n$ at granularity i is:

$$\hat{\mathbf{x}}_i \doteq \boldsymbol{W}_{\mathrm{dec}}^{\top}\, \mathbf{z}_i + \boldsymbol{b}_{\mathrm{pre}}. \tag{2}$$

Note that the reconstruction is a linear combination of the rows of $\boldsymbol{W}_{\mathrm{dec}}$ weighted by the entries of $\mathbf{z}_i$, thus each row $\mathbf{p}_c = [\boldsymbol{W}_{\mathrm{dec}}]_{c,:} \in \mathbb{R}^n$ represents a learned feature direction in CLIP space.

Let $\boldsymbol{\Theta} = \{\boldsymbol{W}_{\mathrm{enc}}, \boldsymbol{b}_{\mathrm{enc}}, \boldsymbol{W}_{\mathrm{dec}}, \boldsymbol{b}_{\mathrm{pre}}\}$ denote the total set of learnable parameters and let $\bar{\mathbf{x}} = \frac{1}{|\mathcal{D}|} \sum_{\mathbf{x} \in \mathcal{D}} \mathbf{x}$ be the mean activation precomputed over the training set $\mathcal{D}$. The per-sample loss normalizes the reconstruction error at each granularity by the input variance, yielding the normalized mean-squared error (nMSE):

$$\mathcal{L}_{\mathrm{MSAE}}(\mathbf{x};\, \boldsymbol{\Theta}) = \sum_{i=1}^{h} \alpha_i \, \frac{\|\mathbf{x} - \hat{\mathbf{x}}_i\|_2^2}{\|\mathbf{x} - \bar{\mathbf{x}}\,\|_2^2}\,, \tag{3}$$

and the model is trained by minimizing $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\big[\mathcal{L}_{\text{MSAE}}(\mathbf{x};\,\Theta)\big]$. This normalization ensures that the loss is scale-invariant across different CLIP architectures and modalities, so that the weighting coefficients α_i have consistent effects regardless of the embedding magnitude.

Weighting strategies. Zaigrajew et al. (2025) present two weighting strategies. Uniform weighting (UW) sets $\alpha_i = 1$ for all i, effectively treating each hierarchy equally. Reverse weighting (RW) uses $\alpha_i = h - i + 1$, which penalizes reconstruction error at lower k values more heavily. The RW strategy encourages the MSAE model to learn representations that maintain reconstruction quality across all granularities.

Concept naming. Each decoder direction $\mathbf{p}_c$ corresponds to a learned feature in CLIP space but does not carry an inherent semantic label. To obtain human-readable names, Zaigrajew et al. (2025) follow the CLIP-based concept matching procedure of Rao et al. (2024), avoiding the computational overhead of LLM-based naming (e.g. Bills et al., 2023). Given a vocabulary $\mathcal{V} = \{w_1, \ldots, w_{|\mathcal{V}|}\}$ of candidate concept words, each entry is embedded with the CLIP text encoder, and the concept assigned to neuron c is the vocabulary element whose embedding is most aligned with $\mathbf{p}_c$:

$$s_c = \arg\max_{w\in\mathcal{V}} \cos\big(\mathbf{p}_c,\, \text{CLIP}(w)\big) = \arg\max_{w\in\mathcal{V}} \frac{\mathbf{p}_c \cdot \text{CLIP}(w)}{\|\mathbf{p}_c\|\,\|\text{CLIP}(w)\|}. \tag{4}$$

Bias validation via concept-level interventions. To investigate whether the MSAE encodes concept-level biases in downstream classifiers, we follow the causal intervention framework of Zaigrajew et al. (2025). Let $g_\psi\colon \mathbb{R}^n \to [0,1]$ be a single-layer binary classifier trained on CLIP embeddings $\mathbf{x}$ with gender labels $y \in \{0,1\}$ using cross-entropy loss. For a target concept word w, we identify the most aligned code dimension c^* using Equation 4. Given an input $\mathbf{x}$ and its corresponding latent code $\mathbf{z}$, we define the *intervened* code:

$$\mathbf{z}^{(\delta)} = \mathbf{z} + \delta\,\mathbf{e}_{c^*}. \tag{5}$$

Where $\mathbf{e}_{c^*}$ is the standard basis vector and $\delta \geq 0$ the intervention magnitude. The intervened reconstruction $\hat{\mathbf{x}}^{(\delta)} = \boldsymbol{W}_{\text{dec}}^\top \mathbf{z}^{(\delta)} + \boldsymbol{b}_{\text{pre}}$ is passed through the classifier to obtain the response curve $\phi(\delta) = g_\psi\big(\hat{\mathbf{x}}^{(\delta)}\big)$. Sweeping δ over a range reveals whether increasing the magnitude of a single concept dimension causally shifts the predicted gender probability, and at what magnitude this effect saturates.

3.2 Datasets

We use three datasets in our experiments. The CC3M training subset (denoted $\mathcal{D}^{\text{train}}_{\text{CC3M}}$) is used to train the SAEs on frozen CLIP embeddings (Radford et al., 2021). Subsequently, the ImageNet-1K training ($\mathcal{D}^{\text{train}}_{\text{IN1K}}$) and the CC3M validation ($\mathcal{D}^{\text{val}}_{\text{CC3M}}$) sets serve as evaluation benchmarks for image and text modalities respectively. Finally, ImageNet-100 ($\mathcal{D}^{\text{val}}_{\text{IN100}}$), ImageNet-1k ($\mathcal{D}^{\text{val}}_{\text{IN1K}}$) validation sets and CelebA ($\mathcal{D}_{\text{CelebA}}$) are used for downstream interpretation tasks.

CC3M. CC3M contains approximately **3 300 000** images annotated with captions harvested from the web (Sharma et al., 2018) 88% of which were available in the final dataset due to link rot.

ImageNet. To replicate the experiments, we used two versions of ImageNet. ImageNet-1K contains approximately **1 300 000** images across **1 000** object categories, and is used as a standard benchmark for image classification (Russakovsky et al., 2015). ImageNet-100 is a subset of ImageNet-1K, and contains approximately **130 000** images across **100** object categories.

CelebA. CelebA consists of over **200 000** images of celebrities, with each image containing 40 binary attribute annotations (Liu et al., 2015). We use the official train / validation / test splits (approximately **80% / 10% / 10%**). This dataset is used for systematic analysis of gender biases in downstream classification models.

Concept vocabularies. Two vocabularies instantiate $\mathcal{V}$ (Equation 4): a subset of CLIP captions containing **20 000** words ($\mathcal{V}_{\text{C20}}$) and the approximately **37 000** most frequent unigrams from LAION-400m captions ($\mathcal{V}_{\text{L4M}}$) (Bhalla et al., 2024; Schuhmann et al., 2021).

Preprocessing. All datasets, except the concept vocabularies, are preprocessed in a consistent manner by extracting the frozen CLIP embeddings and applying per-modality centering (i.e., subtracting the modality-specific mean $\bar{\mathbf{x}}$ like Equation 3). Furthermore, the datasets are scaled so that $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\|\mathbf{x}\|] = \sqrt{n}$ where n is the CLIP embedding dimension. This scaling ensures that the L_1 sparsity regularization coefficient (λ) has consistent effects across setups, which is relevant for the ReLU SAE baselines.

3.3 Hyperparameters

Experiments with SAE models are conducted using the hyperparameters defined in Table 1. These hyperparameters are consistent with those used by Zaigrajew et al. (2025). Unless stated otherwise, CLIP ViT-L/14 embeddings are used ($n = 768$) and an expansion factor of **8** ($d = 8 \times n = 6144$). We use batch sizes **4 096** and **1 024** for training and inference respectively. For all SAEs, the encoder ($\boldsymbol{b}_{\text{enc}}$) and preprocessing ($\boldsymbol{b}_{\text{pre}}$) biases are zero-initialized. The decoder matrix ($\boldsymbol{W}_{\text{dec}}$) is initialized using Kaiming-uniform initialization (He et al., 2015), then scaled to L_2-norm **0.1** (i.e., $\|\boldsymbol{W}_{\text{dec}}\| \approx 0.1$). The encoder matrix ($\boldsymbol{W}_{\text{enc}}$) is initialized as the transpose of the decoder matrix. All models are trained for **30** epochs, optimized with AdamW and use a learning rate scheduler with linear decay **0.8** (Loshchilov & Hutter, 2019).

Table 1: Hyperparameters for SAE-based models used for experimentation.

SAE family	**Variant**	**Sparsity hyperparam**	**LR**	**Granularities / k**	**Notes**
ReLU SAE	$\lambda = 0.03$	L_1 coeff. λ	5×10^{-5}	–	Standard ReLU SAE.
ReLU SAE	$\lambda = 0.003$	L_1 coeff. λ	5×10^{-5}	–	
ReLU SAE	$\lambda = 0.001$	L_1 coeff. λ	5×10^{-5}	–	
TopK SAE	$k = 32$	TopK budget k	5×10^{-4}	$k = 32$	Rigid sparsity constraint.
TopK SAE	$k = 64$	TopK budget k	5×10^{-4}	$k = 64$	
TopK SAE	$k = 128$	TopK budget k	5×10^{-4}	$k = 128$	
TopK SAE	$k = 256$	TopK budget k	5×10^{-4}	$k = 256$	
TopK SAE	$k = 512$	TopK budget k	5×10^{-4}	$k = 512$	
BatchTopK SAE	$k = 16$	BatchTopK budget k	5×10^{-4}	$k = 16$	Trained with the same hyperparameters as TopK.
BatchTopK SAE	$k = 32$	BatchTopK budget k	5×10^{-4}	$k = 32$	
BatchTopK SAE	$k = 64$	BatchTopK budget k	5×10^{-4}	$k = 64$	
BatchTopK SAE	$k = 128$	BatchTopK budget k	5×10^{-4}	$k = 128$	
BatchTopK SAE	$k = 256$	BatchTopK budget k	5×10^{-4}	$k = 256$	
MSAE	UW	TopK hierarchy + weights α_i	1×10^{-4}	$k_i \in \{2^6, 2^7, \ldots, 6144\}$	UW: $\alpha_i = 1$ for all i.
MSAE	RW	TopK hierarchy + weights α_i	1×10^{-4}	$k_i \in \{2^6, 2^7, \ldots, 6144\}$	RW: $\alpha_i \in \{8, 7, 6, 5, 4, 3, 2, 1\}$.

3.4 Experimental setup and code

Our reproduction consists of two main experiments. First, we focus on representation learning by training and evaluating SAE-based models, which we label **E1**. Second, we address representation interpretation by performing inference with the MSAE model on CLIP activations, labeled **E2**. The code used to reproduce this experiment can be found in this GitHub repository by Zaigrajew et al. (2025). The code to our extension can be found in two (train, inference) repositories. Compute has been provided by the Snellius supercomputer through the use of NVIDIA A100 and H100 GPUs.

3.4.1 Training and evaluating SAE-based models (E1)

We train and evaluate a variety of models to investigate **C1** and **C2**: ReLU SAEs, TopK SAEs, BatchTopK SAEs and MSAEs. Evaluation sets are as defined in Section 3.2. We evaluate **E1** on several metrics, an overview is documented in Appendix A.

Sparsity-fidelity trade-off. To assess **C1**, we analyze the sparsity-fidelity trade-off. The sparsity-fidelity trade-off implies that as sparsity in codes ($\mathbf{z}$) grows, the reconstruction quality (fidelity) may decrease.

Relating to **C1**, this is our objective: keep reconstructions ($\hat{\mathbf{x}}$) explainable while increasing sparsity. Following Zaigrajew et al. (2025), we measure sparsity (L_0) as the mean proportion of masked dimensions of pre-activations ($\mathbf{a}$). Fraction of Variance Unexplained (FVU), defined as nMSE centered over a held-out set, measures fidelity. Cosine Similarity (CS) between activations ($\mathbf{x}$) and reconstructions is used as a secondary metric for fidelity. We count the Number of Dead Neurons (NDN) to quantify sparsity that did not contribute to disentanglement, while reducing the bandwidth available for reconstruction.

Semantic alignment. Zaigrajew et al. (2025) propose to investigate alignment with Linear Probes (LP) and Centered Kernel Nearest-Neighbor Alignment (CKNNA). Considering their experimentation showed high correlation between LP metrics, FVU and CS, we deem it interesting to only reproduce CKNNA to evaluate local geometric alignment between activation space and code space on a set of k-Nearest Neighbors (kNNs) in each respective space.

Modality transfer. To assess the ability of the models to generalize towards a different modality, we evaluate our models on sparsity-fidelity and semantic alignment across both the vision and language modalities.

Progressive recovery. Regarding **C2**, we perform a progressive recovery task on each model listed in Table 1. However, following Zaigrajew et al. (2025), we exclude the BatchTopK models. During inference, we increase the number of active neurons (k) while measuring FVU and CKNNA. A hierarchical representation should recover coarse features at low k, while FVU decreases and CKNNA increases as k increases, indicating that finer features are recovered.

3.4.2 Applying SAE-based models for concept interpretation (E2)

To evaluate **C3**, we qualitatively investigate three aspects of the claim: concept naming of latent codes, concept-based similarity search and bias validation in downstream tasks.

Concept naming. To investigate whether individual latent codes correspond to coherent semantic concepts, we assign labels to decoder directions using the CLIP-based similarity framework (Section 3.1). We select a concept—*face*—and embed each candidate word $w \in \mathcal{V}$ using the CLIP text encoder. For each neuron c, we identify the vocabulary element whose embedding is most aligned with $\mathbf{p}_c$ using Equation 4. The resulting concept assignments are evaluated on $\mathcal{D}^{\text{val}}_{\text{IN1K}}$ and $\mathcal{V}_{\text{L4M}}$ by inspecting top-activating examples for the concept *face* across both modalities. In practice, poor concept assignment can occur when the best found similarity is low or when coarse-to-fine concepts compete for the same feature. Following Zaigrajew et al. (2025) we address this by filtering matches, requiring the top match (CS $>$**0.42**) to be at least twice as strong as the second best match and the target concept to be the strongest match to a feature.

Similarity search. We evaluate whether targeted interventions on individual concept magnitudes in the learned code space produce semantically consistent changes in CLIP-space nearest-neighbor retrieval. Given a query image with CLIP embedding $\mathbf{x}$ and its corresponding latent code $\mathbf{z}$ obtained via Equation 1, we identify the code dimension c^* most aligned with a target concept—*red*—from $\mathcal{V}_{\text{C20}}$ using precomputed concept match scores (Equation 4). We then construct an intervened code $\tilde{\mathbf{z}}$ by setting $z_{c^*} = \bar{z}_{c^*} + \alpha$, where $\bar{z}_{c^*}$ denotes the mean activation and $\alpha > 0$ controls the intervention magnitude. The intervened code is decoded and post-processed to obtain a reconstructed vector $\hat{\mathbf{x}}^{(\alpha)}$ in CLIP embedding space (Equation 2), which is then L_2-normalized. We retrieve the top kNNs of $\hat{\mathbf{x}}^{(\alpha)}$ under cosine distance over the CLIP embeddings and select the closest non-self match. We then examine whether the retrieved image reflects increased presence of the target concept relative to the unmodified query.

Bias validation in downstream tasks. To inspect whether the MSAE encodes concept-level biases for learned features, we train a single-layer binary classifier $g_\psi(\cdot)$ (Section 3.1) for **11** epochs on $\mathcal{D}_{\text{CelebA}}$ using cross-entropy loss. We select three semantic attributes—*bearded*, *glasses*, and *blondes*—and identify their corresponding code dimensions c^* via Equation 4. For each concept, the activation magnitude of z_{c^*} is increased over **50** iterations with varying δ (Equation 5). The intervened codes are decoded and passed

through $g_\psi(\cdot)$ to obtain the predicted gender class probability $\phi(\delta)$, revealing concept-gender associations and their saturation behaviour.

3.5 Extensions

Always-on neurons. After reading a post-publishing remark by Zaigrajew et al., a concern about neurons that are always activated was raised. We question whether the coarsest granularities are filled by true concept detectors or by features that fire on (almost) every input to reduce reconstruction error.

To probe this, we embed each *non-dead latent neuron* as its dense (eval-mode) activation profile over a random subset of **4 000** images, and run UMAP (McInnes et al., 2020) (cosine metric, $n_{\text{neighbors}} = 15$, min_dist $= 0.1$) on the resulting $[d, 4000]$ matrix, coloring each point by the fraction of images on which it is active. We label a neuron *always-on* when it fires (activation $> 10^{-6}$) on at least **95%** of images. In their post publishing remark, Zaigrajew et al. note that always-on neurons could be caused by the finest granularity, $k_i = 6144$, which performs an unconstrained dense reconstruction with no sparsity penalty. We hypothesize that applying L_1-regularization to these fine granularities during training could reduce the number of always-on neurons and thus improve the interpretability of the features. To test this, we add a penalty on the latent code beyond a per-image rank cutoff k^*, leaving the high-magnitude head (top-k^*) untouched:

$$\mathcal{L}_{\text{MSAE}'} = \mathcal{L}_{\text{MSAE}} + \lambda \left\| \mathbf{z} - \text{TopK}_{k^*}(\mathbf{z}) \right\|_1, \tag{6}$$

which suppresses the tail while avoiding L_1's classic activation-shrinkage on the important head features. We train UW and RW MSAE models with, $k^* = 512$ and $\lambda = 0.003$.

Concept saliency maps. Concept naming and similarity search establish that individual MSAE features correspond to coherent concepts, but only at the whole-image level: they reveal *which* images activate a feature, not *where* in an image it is active. As a further extension we add an inference-only, task-agnostic spatial probe that turns a single MSAE feature into a per-concept saliency map, giving a pixel-level counterpart to the image-level decomposition above in support of **C3**.

We adapt Concept Visualization (ConVis, Giulivi & Boracchi (2024)), which slides two-scale square patches across an image, scores each patch, and aggregates the scores back onto pixels. We keep this spatial scaffold but replace its WordNet-definition scorer with the MSAE itself. Recalling that each decoder row $\boldsymbol{p}_c = [\boldsymbol{W}_{dec}]_{c,:} \in \mathbb{R}^n$ is a feature direction in CLIP space, we match a concept word w to a feature using the criterion of our concept naming setup (Section 3.4.2) without the fixed vocabulary $\mathcal{V}$:

$$c^* = \underset{c \in \{1,\ldots,d\}}{\arg\max} \cos\big(\text{CLIP}_T(w),\, \boldsymbol{p}_c\big). \tag{7}$$

For each patch location (i, j) we form the mean embedding $\tilde{\boldsymbol{e}}_{i,j}$ of a small (δ_S) and large (δ_L) crop exactly as in ConVis, encode it with the trained MSAE (Section 3.1, Equation 1), and take the matched feature's activation $s_{i,j} = [\mathbf{z}(\tilde{\boldsymbol{e}}_{i,j})]_{c^*}$ as the patch score. We convert scores to per-concept ranks across the P patches of the image, thereby removing the dependence on raw activation scale and on patch detail, and, unlike ConVis's ranking across concepts, suiting maps shown one at a time. Then we aggregate onto pixels following ConVis:

$$r_{i,j} = \tfrac{1}{P} \left| \{(l,m) : s_{l,m} < s_{i,j}\} \right|, \qquad \boldsymbol{y}_{\boldsymbol{x}}^{w}[i,j] = \underset{|l-i|<\delta_L,\, |m-j|<\delta_L}{\text{avg}} r_{l,m}. \tag{8}$$

We use the frozen CLIP ViT-L/14 backbone ($n = 768$) and the RW image-trained MSAE on $\mathcal{D}^{\text{train}}_{\text{CC3M}}$ ($d = 6144$, TopK $k = 64$), the same checkpoint as our main experiments (Section 3.4.1), with $\delta_S = 64$, $\delta_L = 128$, $\omega = 16$. Maps are rendered as heatmap overlays in a grid, one panel per concept; a monotone power-law contrast stretch is applied for display only and leaves Equation 8 unchanged.

4 Results

This section presents the results of experiments **E1** and **E2**. Overall our findings are consistent with claims **C1** through **C3**. We first report the reproducibility results, followed by the results of our proposed extensions.

4.1 Results reproducing original paper

Quantitative model comparisons. Figure 1 shows that the Matryoshka variants achieve the lowest FVU and highest CS across all models, while maintaining substantially sparser representations than the best-performing ReLU models. TopK models attain higher sparsity than MSAE, but at the cost of reduced reconstruction fidelity and alignment. Across metrics, MSAE occupies a favorable sparsity–fidelity Pareto frontier, combining near-perfect reconstruction quality with strong neighborhood preservation. CKNNA scores are generally high for ReLU, TopK, and MSAE models, with the Matryoshka UW variant achieving the highest alignment. Finally, MSAE exhibits virtually no dead neurons, comparable to ReLU and most TopK variants, indicating effective utilization of the learned feature space. We defer discussion of the L_1-regularized variants to Section 4.2. These results align with **C1**, demonstrating that MSAE achieves a favorable balance between sparsity, fidelity, representation bandwidth, and alignment.

(a) Quantitative comparison of SAE-based models on ImageNet-1k.

SAE family	L_0 ↑	FVU ↓	CS ↑	CKNNA ↑	NDN ↓
ReLU ($\lambda=0.03$)	$0.920_{\pm 0.012}$	$0.072_{\pm 0.015}$	$0.964_{\pm 0.008}$	$0.737_{\pm 0.008}$	**0**
ReLU ($\lambda=0.003$)	$0.650_{\pm 0.007}$	$0.002_{\pm 0.000}$	$0.999_{\pm 0.000}$	$0.951_{\pm 0.003}$	0
ReLU ($\lambda=0.001$)	$0.545_{\pm 0.006}$	$0.003_{\pm 0.002}$	$0.999_{\pm 0.001}$	$0.970_{\pm 0.001}$	0
TopK ($k=32$)	$\mathbf{0.963}_{\pm 0.010}$	$0.159_{\pm 0.072}$	$0.928_{\pm 0.025}$	$0.654_{\pm 0.003}$	0
TopK ($k=64$)	$0.950_{\pm 0.011}$	$0.104_{\pm 0.039}$	$0.953_{\pm 0.014}$	$0.745_{\pm 0.004}$	0
TopK ($k=128$)	$0.924_{\pm 0.011}$	$0.055_{\pm 0.020}$	$0.975_{\pm 0.008}$	$0.804_{\pm 0.003}$	0
TopK ($k=256$)	$0.899_{\pm 0.005}$	$0.006_{\pm 0.002}$	$0.997_{\pm 0.001}$	$0.936_{\pm 0.002}$	0
TopK ($k=512$)	$0.920_{\pm 0.014}$	$0.072_{\pm 0.060}$	$0.965_{\pm 0.030}$	$0.579_{\pm 0.004}$	0
BatchTopK ($k=16$)	$0.670_{\pm 0.031}$	$0.314_{\pm 0.142}$	$0.866_{\pm 0.046}$	$0.592_{\pm 0.003}$	28
BatchTopK ($k=32$)	$0.784_{\pm 0.025}$	$0.169_{\pm 0.054}$	$0.926_{\pm 0.022}$	$0.724_{\pm 0.002}$	1
BatchTopK ($k=64$)	$0.882_{\pm 0.015}$	$0.097_{\pm 0.026}$	$0.956_{\pm 0.011}$	$0.769_{\pm 0.002}$	0
BatchTopK ($k=128$)	$0.896_{\pm 0.011}$	$0.042_{\pm 0.011}$	$0.980_{\pm 0.005}$	$0.838_{\pm 0.003}$	2
BatchTopK ($k=256$)	$0.885_{\pm 0.005}$	$0.005_{\pm 0.002}$	$0.997_{\pm 0.001}$	$0.930_{\pm 0.001}$	0
Matryoshka (UW)	$0.690_{\pm 0.005}$	$\mathbf{0.000}_{\pm 0.000}$	$\mathbf{1.000}_{\pm 0.000}$	$\mathbf{0.988}_{\pm 0.001}$	0
Matryoshka (RW)	$0.807_{\pm 0.007}$	$0.002_{\pm 0.001}$	$0.999_{\pm 0.001}$	$0.947_{\pm 0.002}$	1
Matryoshka+L_1 (UW)	$0.876_{\pm 0.007}$	$0.004_{\pm 0.001}$	$0.998_{\pm 0.001}$	$0.924_{\pm 0.002}$	0
Matryoshka+L_1 (RW)	$0.901_{\pm 0.007}$	$0.012_{\pm 0.004}$	$0.994_{\pm 0.002}$	$0.886_{\pm 0.003}$	0

(b) Sparsity–fidelity trade-off across SAE families.

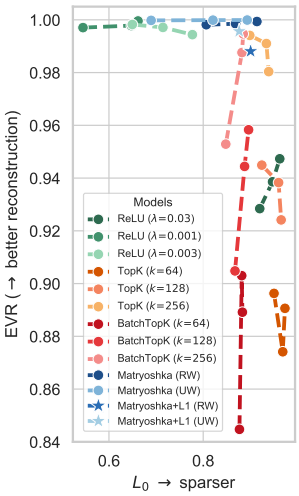

Figure 1: Comparison (mean±std.) of SAE-based models evaluated on the image modality of $\mathcal{D}^{\text{train}}_{\text{IN1K}}$. Quantitative metrics are shown in (a), while (b) illustrates the corresponding sparsity–fidelity trade-off.

Table 2: MSAE models trained on language ($\mathcal{D}^{\text{train}}_{\text{CC3M}}$) and vision ($\mathcal{D}^{\text{train}}_{\text{IN1K}}$) are evaluated (mean±std.) on $\mathcal{D}^{\text{val}}_{\text{CC3M}}$ (language) and $\mathcal{D}^{\text{val}}_{\text{IN1K}}$ (vision).

	Language metrics on CC3M				Vision metrics on ImageNet-1k				
Variant	L_0 ↑	FVU ↓	CS ↑	CKNNA ↑	L_0 ↑	FVU ↓	CS ↑	CKNNA ↑	NDN ↓
Image (RW)	$0.814_{\pm 0.024}$	$0.040_{\pm 0.048}$	$0.981_{\pm 0.022}$	$0.417_{\pm 0.022}$	$0.807_{\pm 0.007}$	$0.002_{\pm 0.001}$	$\mathbf{>.999}_{\pm <.001}$	$0.947_{\pm 0.002}$	1
Image (UW)	$0.730_{\pm 0.014}$	$0.021_{\pm 0.027}$	$0.991_{\pm 0.012}$	$0.385_{\pm 0.008}$	$0.690_{\pm 0.005}$	$\mathbf{<.001}_{\pm <.001}$	$\mathbf{>.999}_{\pm <.001}$	$\mathbf{0.988}_{\pm <.001}$	**0**
Text (RW)	$\mathbf{0.822}_{\pm 0.012}$	$0.003_{\pm 0.002}$	$0.999_{\pm <.001}$	$0.964_{\pm 0.011}$	$\mathbf{0.823}_{\pm 0.010}$	$0.018_{\pm 0.020}$	$0.992_{\pm 0.009}$	$0.681_{\pm 0.006}$	**0**
Text (UW)	$0.744_{\pm 0.013}$	$\mathbf{<.001}_{\pm <.001}$	$\mathbf{>.999}_{\pm <.001}$	$\mathbf{0.981}_{\pm 0.012}$	$0.765_{\pm 0.010}$	$0.008_{\pm 0.011}$	$0.996_{\pm 0.005}$	$0.736_{\pm 0.006}$	1

Modality transfer. Table 2 shows that MSAE models perform best on the modality on which they were trained. On the language modality, text-trained models substantially outperform image-trained models across FVU, CS and CKNNA, with Text (UW) achieving the strongest alignment and reconstruction quality. Conversely, on the vision modality, image-trained models achieve lower FVU and higher CKNNA, indicating better preservation of visual representations. Across both modalities, text-trained models consistently exhibit higher sparsity (L_0) while maintaining negligible numbers of dead neurons. These results suggest that language training encourages sparser representations and strong semantic alignment, whereas image training

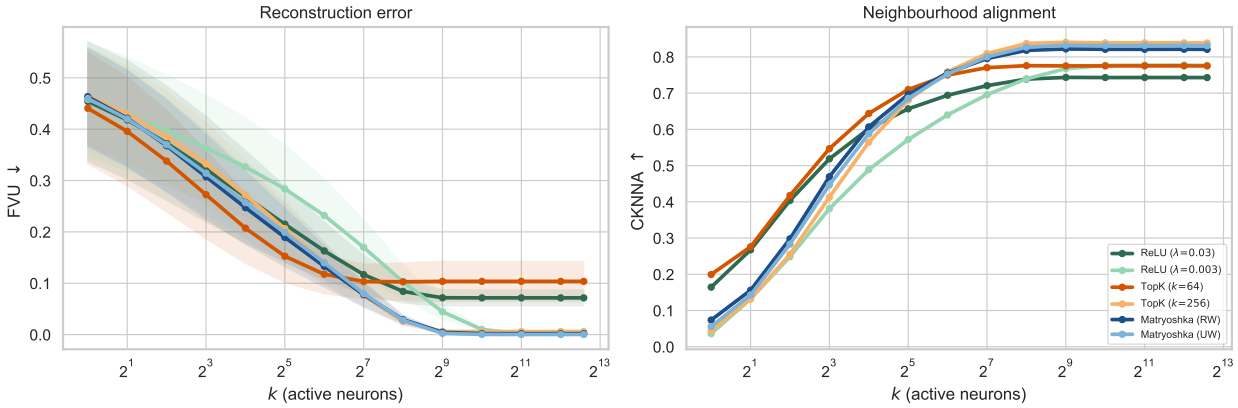

Figure 2: Progressive recovery results on $\mathcal{D}^{\text{train}}_{\text{IN1K}}$ image modality. Results for varying models on FVU (left) and CKNNA (right) with increasing active neurons (k).

Figure 3: Comparison of top-activating examples for concept *face* across ReLU, TopK, and MSAE for sparse (top) and reconstruction-focused (bottom) checkpoints. The images are evaluated on $\mathcal{D}^{\text{val}}_{\text{IN1K}}$ and text $\mathcal{D}^{\text{val}}_{\text{CC3M}}$, with concepts named using $\mathcal{V}_{\text{L4M}}$. 'top-1' refers to when there exists a neuron that has highest similarity with the concept *face*, 'fallback' refers to an alternative when no top-1 match exists.

yields superior reconstruction and neighborhood preservation on visual data. Consistent with **C1**, MSAE maintains high fidelity and alignment across modalities while preserving sparse representations.

Progressive recovery. Figure 2 shows that increasing the number of active neurons (k) generally improves reconstruction fidelity and neighborhood alignment. TopK models plateau near the sparsity level they were trained on, while ReLU models improve as k increases but underperform TopK when k is lower, reflecting the sparsity-fidelity trade-off. MSAEs maintain competitive performance across the full range of k and perform better than TopK and ReLU models at higher k. These results are consistent with **C2** in the sense that

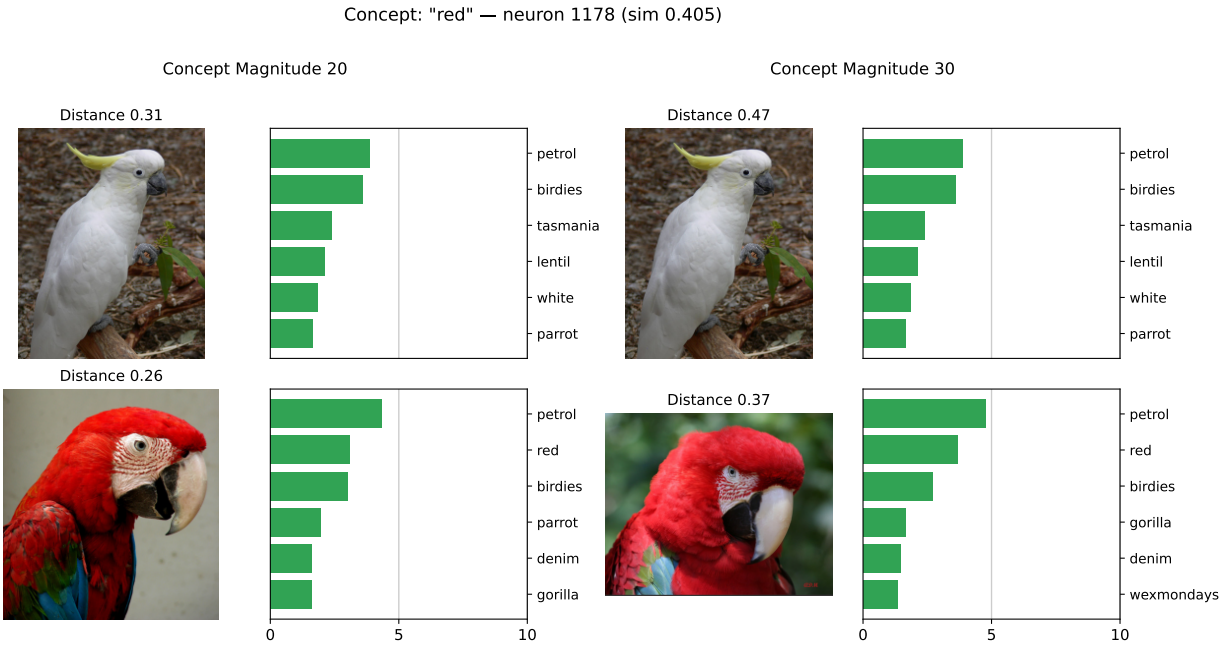

Figure 4: Nearest-neighbor retrieval for a query image from ImageNet-100 in CLIP embedding space after intervening concept *red* at different activation magnitudes, with the corresponding highest activated concept labels. The concept labels correspond to the highest-activated SAE neurons, mapped to a fixed vocabulary derived from $\mathcal{V}_{C20}$.

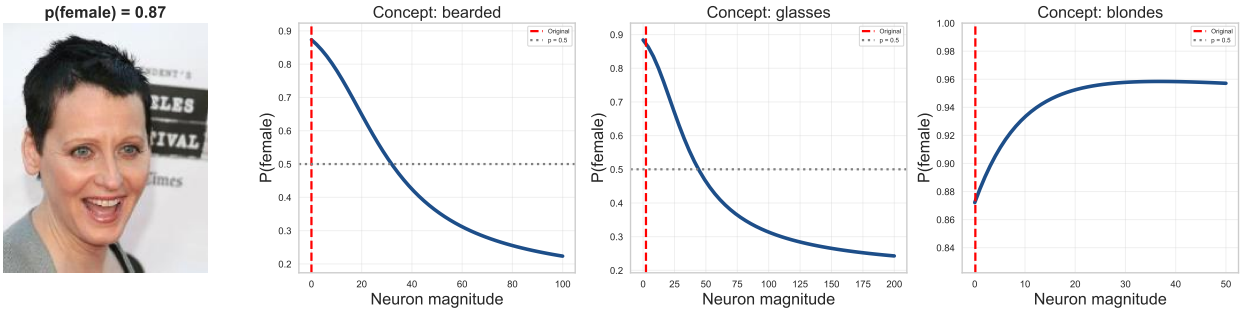

Figure 5: Impact of RW MSAE codes on CLIP gender classification on train images from CelebA dataset after increasing concept magnitudes for concepts *bearded*, *glasses* and *blondes*. In this figure, the y-axis illustrates the class probability (1.0 for female, 0.0 for male).

information can be progressively recovered as additional latent dimensions become active. However, the figure alone does not establish that this recovery corresponds to semantically coarse-to-fine features.

Concept naming. As shown in Figure 3, all three models show a consistent presence of the concept *face* in the image modality. While ReLU and TopK SAEs capture certain visual aspects related to faces, their top-activating images contain objects and faces of animals. For MSAE, the highest-matching neurons correspond to human faces and make-up related features, yielding more faithful concept retrieval than the baseline models.

With regard to the text modality, we observe that all models are able to retrieve captions that are closely related to faces, such as references to people, facial attributes, or close-up descriptions. Compared to Zaigrajew et al. (2025), our ReLU model captures the concept *face* less, while our TopK and MSAE images

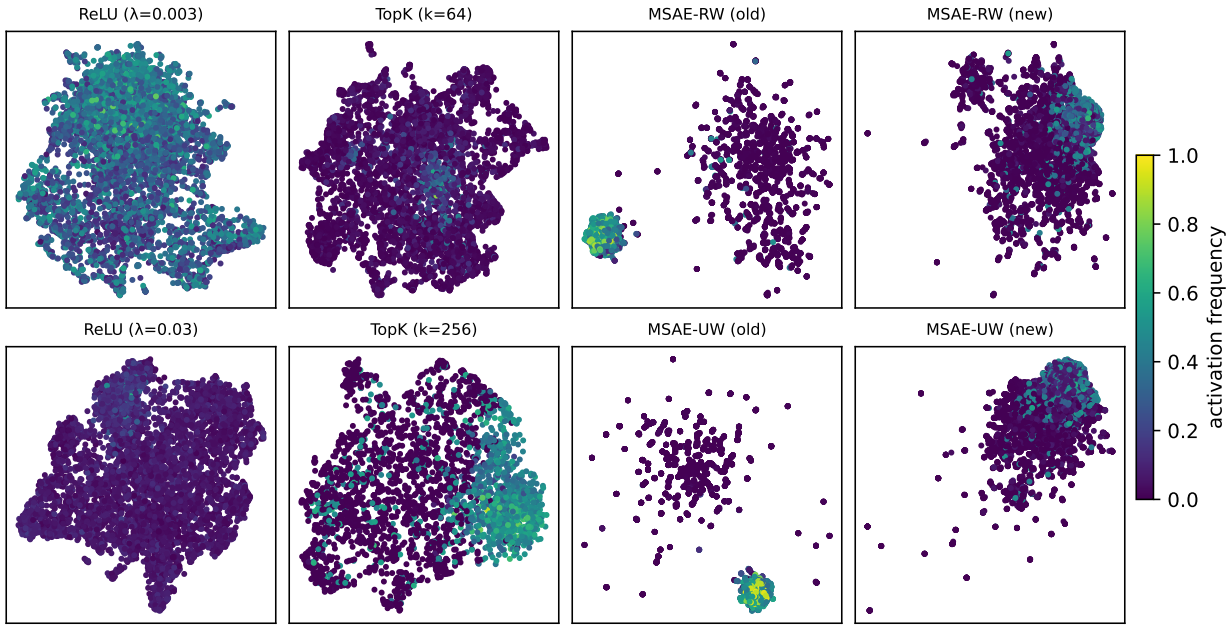

Figure 6: Per-neuron UMAP, where each point is one non-dead neuron positioned by its activations across images and colored by activation frequency. MSAE-UW (new) and MSAE-RW (new) have $k^* = 512$.

capture the concept *face* more. Furthermore, the original paper shares a substantial part of the captions with our results across all three models. These results support **C3**.

Similarity search. Figure 4 shows how increasing the activation magnitude of a concept's neuron results in corresponding changes in both the retrieved image and its associated concepts. The input image is a white cockatoo, for which *white* appears among the highest-activated concepts. After increasing the activation magnitude of the concept *red*, the closest image is that containing a red parrot. Moreover, this image has *red* as a retrieved concept word, demonstrating that increasing the activation magnitude of a neuron affects both visual and text modalities. Furthermore, Figure 4 shows that a higher concept magnitude leads to a mostly similar alignment with the text modality: all images contain concepts *petrol*, *birdies* and *parrot* across both magnitudes. Unlike Zaigrajew et al. (2025), the distances between the query image and the retrieved image decrease after increasing the activation magnitude. The images qualitatively display a clearer difference between the query image and the retrieved image, while the concept labels of the original paper show more semantic consistency. These results are largely consistent with **C3**.

Bias validation in downstream task. As shown in Figure 5, increasing concept magnitudes leads to the corresponding neurons biasing toward either male or female classification. Concepts such as *bearded* and *glasses* bias the model toward male classification, while the concept *blondes* biases the model towards female classification. Furthermore, the influence of each concept saturates beyond a certain activation magnitude, suggesting that stronger magnitudes have little effect on the gender classifier once a plateau in class probability is reached. Different concepts saturate at different activation magnitudes, indicating that their influence on the classifier varies across concepts. Compared to the original paper, *glasses* and *blondes* have a slightly weaker bias, but the general result is similar. These results support **C3** by demonstrating that individual MSAE features correspond to semantically meaningful directions that can causally influence downstream model predictions.

Figure 7: Concept saliency maps from the RW image-trained MSAE, shown as heatmap overlays where brighter regions indicate higher saliency $\boldsymbol{y}_{\boldsymbol{x}}^{w}$. Each row shows the input image (left) followed by one map per concept. *(Top)* Fruit bowl with concepts *fruit*, *banana*, *grapes*, *orange*. *(Bottom)* An animal scene with concepts *cat*, *dog*, *paws*, *parrot*. Maps localize to the region matching each concept: *fruit* covers the whole arrangement, while *cat* responds to both the cat and the tiger.

4.2 Results beyond original paper

Always-on neurons. Figure 6 shows the neuron-UMAP grid across all variants. The ReLU and TopK ($k = 64$) models form a single cloud with no separated population, while TopK ($k = 256$) develops a distinct frequently-on cluster. The unregularized MSAE variants exhibit a compact, high activation frequency cluster detached from the scattered selective neurons. MSAE-UW has a more pronounced high activation frequency cluster than MSAE-RW, likely attributed to the RW strategy providing less weight to the finer granularities. These neurons fire at a roughly constant level on every image, rather than acting like a feature detector. This is consistent with the post-publishing remark that these always-on neurons arise from the unconstrained finest granularity. Applying our L_1-regularization to the finer granularities removes the dense cluster for both UW and RW models, which in the UMAP corresponds to the detached high activation frequency cluster dissolving.

Figure 1a provides quantitative results of the models with L_1-regularization applied to the finer granularities. For both weightings the regularized models become sparser while incurring only a small drop in fidelity and alignment. Dead neurons stay zero. The regularized variants therefore remain among the strongest models in Figure 1a, retaining competitive reconstruction and high neighborhood preservation. This is, however, a trade-off instead of a free improvement. Compared to the unregularized MSAE, the regularized models sit slightly below the sparsity-fidelity Pareto frontier (1b), and truncating the unregularized code to a matched L_0 reconstructs better than the regularized model at the same sparsity. The contribution should thus be seen as removing the uninterpretable always-on cluster at a small and controlled fidelity cost, not as advancing the frontier.

Concept saliency maps. Figure 7 reports per-concept maps for two images: a fruit bowl and a scene containing a tiger, dog, cat and parrot. In each panel the saliency concentrates on the region whose visual content matches the queried concept. For the fruit bowl, *fruit* spreads across the whole arrangement while *banana*, *grapes* and *orange* each localize to the corresponding items; the broader concept covers more of the image than the specific ones, a spatial reflection of concept generality. The animal scene behaves consistently at the object level: *dog* and *parrot* localize cleanly to the dog and the parrot, while *cat* responds to both the domestic cat and the tiger, consistent with their proximity in CLIP space, where the matched feature fires on

felines rather than a single instance. The part concept *paws* responds over the animals' legs and feet rather than any whole body, indicating that localization is not limited to object-level concepts. Because the rank transform assigns every patch a nonzero value, low-saliency background is suppressed but not fully removed, leaving a faint diffuse response outside the matched object. The maps thus give a pixel-level, qualitative view of the feature-concept correspondence already established by concept naming and similarity search, showing *where* a matched MSAE feature is active within an image, in further support of **C3**.

5 Discussion

We reproduce **C1** through **C3** successfully. However, our results do not exactly match those of the original work. This could be due to the authors releasing post-publishing remarks. We show that MSAEs achieve flexible sparsity through their ability to generalize across granularities at inference. Alignment is slightly lower, while representation bandwidth is higher for MSAE, compared to TopK and BatchTopK models. MSAE has higher alignment and similar representation bandwidth, compared to the ReLU models.

While progressive recovery demonstrates that reconstruction fidelity and neighborhood alignment improve as additional neurons become active, these metrics do not directly verify a coarse-to-fine semantic hierarchy. We therefore complement the quantitative analysis with qualitative visualizations, which provide evidence consistent with **C2**. Regarding **C3**, concept naming and similarity search show that MSAEs can be applied for concept interpretation in CLIP multimodal embedding space. Through bias validation in a downstream task, we show that neurons learn biases when neuron magnitudes are increased.

Our concept saliency extension shows that MSAE features are not only nameable but spatially localizable: querying a feature highlights the image regions it responds to, with broader concepts covering larger regions than specific ones. This provides qualitative, pixel-level evidence for **C3** and complements our image-level analyses. Localization is most reliable when the queried concept admits a decoder direction of high cosine similarity, mirroring the validity issues observed in concept naming.

We further identify a population of always-on neurons specific to MSAE, and show that a targeted L_1 penalty on the finer granularities removes them at a small, controlled cost to reconstruction fidelity rather than a free improvement.

Future work could investigate training efficiency in addition to inference-time performance, as our evaluation focuses primarily on reconstruction quality, sparsity, and alignment. The L_1 intervention could also be explored further by varying the penalty strength λ and the rank cutoff k^*, and by evaluating whether removing the always-on neurons improves downstream interpretability rather than reconstruction and sparsity alone. Finally, while our concept saliency maps provide qualitative evidence for interpretability, it remains unclear whether cross-modal alignment leads to more interpretable features, motivating a systematic human evaluation of aligned and non-aligned models.

References

Usha Bhalla, Alex Oesterling, Suraj Srinivas, Flavio P. Calmon, and Himabindu Lakkaraju. Interpreting clip with sparse linear concept embeddings (splice), 2024. URL `https://arxiv.org/abs/2402.10376`.

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.

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.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023. URL https://arxiv.org/abs/2309.08600.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders, 2024. URL https://arxiv.org/abs/2406.04093.

Loris Giulivi and Giacomo Boracchi. Concept visualization: Explaining the clip multi-modal embedding using wordnet, 2024. URL https://arxiv.org/abs/2405.14563.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *CoRR*, abs/1502.01852, 2015. URL http://arxiv.org/abs/1502.01852.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis, 2024. URL https://arxiv.org/abs/2405.07987.

Oleksandr Khainas, Nataliia Melnykova, and Solomia Fedushko. Contrastive language-image pre-training (clip) in e-commerce: applications, methodologies, and performance. *Advances in Cyber-Physical Systems*, 10:100–104, 06 2025. doi: 10.23939/acps2025.01.100.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. Matryoshka representation learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 30233–30249. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/c32319f4868da7613d78af9993100e42-Paper-Conference.pdf.

Zheng Li, Xiangwei Zheng, Dejian Su, and Mingzhe Zhang. Cdclip: An interpretable zero-shot medical image classification framework based on concept decomposition. *Biomedical Signal Processing and Control*, 116: 109585, 2026. ISSN 1746-8094. doi: https://doi.org/10.1016/j.bspc.2026.109585. URL https://www.sciencedirect.com/science/article/pii/S1746809426001394.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild, 2015. URL https://arxiv.org/abs/1411.7766.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

Alireza Makhzani and Brendan Frey. K-sparse autoencoders. *arXiv preprint arXiv:1312.5663*, 2013.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020. URL https://arxiv.org/abs/1802.03426.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Sukrut Rao, Sweta Mahajan, Moritz Böhle, and Bernt Schiele. Discover-then-name: Task-agnostic concept bottlenecks via automated concept discovery, 2024. URL https://arxiv.org/abs/2407.14499.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge, 2015. URL https://arxiv.org/abs/1409.0575.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs, 2021. URL https://arxiv.org/abs/2111.02114.

Lee Sharkey, Dan Braun, and Beren Millidge. Interim research report: Taking features out of superposition with sparse autoencoders. AI Alignment Forum, 2022. URL https://www.alignmentforum.org/posts/z6QQJbtpkEAX3Aojj/interim-research-report-taking-features-out-of-superposition.

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.

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.

A Metrics

Metrics reproduced by us, as used in Zaigrajew et al. (2025), are documented in Table 3. We include formal definitions and an intuition of the metric. Furthermore, we include a direct reference to the original literature.

Table 3: Metrics used to evaluate SAE-based models on activation spaces (e.g., CLIP), with intuition and formal definitions.

Metric	Meaning / intuition	Formal definition	Original reference
L_0	Sparsity of activations (higher ↑ = sparser).	For codes $\mathbf{z} \in \mathbb{R}^d$, $$L_0 = \mathbb{E}\left[\frac{1}{d}\sum_{j=1}^{d}\mathbb{I}\{z_j = 0\}\right].$$	(Gao et al., 2024)
FVU	Reconstruction fidelity relative to input variance.	Let $\hat{\mathbf{x}}$ be the reconstruction, and $\bar{\mathbf{x}} = \mathbb{E}[\mathbf{x}]$ the sample mean over a held-out dataset. $$\mathrm{FVU} = \frac{\mathbb{E}[\lVert\mathbf{x}-\hat{\mathbf{x}}\rVert_2^2]}{\mathbb{E}[\lVert\mathbf{x}-\bar{\mathbf{x}}\rVert_2^2]},$$	(Gao et al., 2024)
CS	Directional similarity between $\mathbf{x}$ and $\hat{\mathbf{x}}$ (often tracks reconstruction quality).	$$\mathrm{CS}(\mathbf{x},\hat{\mathbf{x}}) = \frac{\mathbf{x}^\top\hat{\mathbf{x}}}{\lVert\mathbf{x}\rVert\,\lVert\hat{\mathbf{x}}\rVert}.$$	-
CKNNA	Local-geometry preservation via kNN neighborhood alignment (higher ↑).	Let K^A, K^B be kernels over the same n samples (e.g. in space A: CLIP embeddings; in B: SAE activations). Let $M^A_{ij} = \mathbb{I}\{j \in k\mathrm{NN}_A(i)\}$ and similarly M^B, and define a mutual mask $M = M^A \odot M^B$. With centered kernels $\tilde{K} = HKH$ where $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$, $$\mathrm{CKNNA} = \frac{\langle \tilde{K}^A \odot M,\ \tilde{K}^B \odot M\rangle_F}{\lVert\tilde{K}^A \odot M\rVert_F\ \lVert\tilde{K}^B \odot M\rVert_F}.$$	(Huh et al., 2024; Kornblith et al., 2019)
NDN	Wasted capacity: latents that never fire on a reference set (lower ↓).	Given dataset $\mathcal{D}$ and threshold τ (often $\tau = 0$ for ReLU-gated), $$\mathrm{NDN} = \sum_{j=1}^{m}\mathbb{I}\left\{\max_{x\in\mathcal{D}} a_j(x) \le \tau\right\}.$$	(Gao et al., 2024)