NAMED FROM THE START: CONCEPT-ANCHORED SPARSE AUTOENCODERS FOR INTERPRETABILITY OF VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse autoencoders (SAEs) decompose model activations into interpretable features, but learn those features without semantic identities. Names are assigned only afterward, through inspection or similarity matching, and this step often fails even when the desired concept is present in the learned dictionary. We hypothesize that the problem lies not in feature discovery but in semantic assignment: post-hoc methods must align an already learned dictionary with language, although nothing during training required their geometries to agree. We introduce NAMED-SAE, a concept-anchored SAE that makes semantic identity part of dictionary learning. Text-derived concept embeddings define the named latents, a shared orthogonal adapter aligns them with the activation space, and bounded concept-specific residuals permit adaptation to visual evidence without unrestricted semantic drift. A separately budgeted bank of unnamed latents captures information outside the vocabulary. The adapter is estimated once and frozen; it can be fitted using limited labelled examples or inferred without annotations from unsupervised pseudo-means. Across CLIP and SigLIP, NAMED-SAE provides complete coverage of the target concepts, produces names that are more faithful to activating images and more predictive of intervention effects than post-hoc assignments, and retains most of the reconstruction quality of an unconstrained SAE. These identities also persist under distribution shift without retraining or renaming, while changes in latent activity describe the semantic content of the shift itself. Our results show that semantic identity is more reliably imposed as a constraint during feature learning than recovered through analysis afterward.

1 INTRODUCTION

Interpreting a vision–language model requires expressing, in words, what its representations contain. Standard sparse autoencoders (SAEs) (Huben et al., 2024; Chanin et al., 2026; Haque et al., 2026; Rao et al., 2024; Hua et al., 2025) provide part of this interface by decomposing model activations into a dictionary of sparsely activated directions. Applied to CLIP (Radford et al., 2021) and SigLIP (Zhai et al., 2023), they recover directions that respond coherently to recognizable visual content (Pach et al., 2026). What standard SAEs do not provide is the corresponding language. They are trained for sparse reconstruction, with no constraint that a direction should mean *goldfish*, *fog*, or any other human-defined concept. Meaning is assigned only after training by inspecting highly activating images, captioning them, or matching decoder rows to text embeddings. A latent therefore has no intrinsic identity: its assigned name may be inaccurate, a desired concept may not be represented by any latent, and a single concept may be fragmented across several features.

Existing approaches expose this tension between open-ended feature discovery and guaranteed semantic specification. Post-hoc naming methods, including Network Dissection (Bau et al., 2017), MILAN (Hernandez et al., 2021), CLIP-Dissect (Oikarinen & Weng, 2022), and text matching for VLM SAEs (Pach et al., 2026), preserve representations learned from the activations but assign their semantics only after training. Discover-then-Name (Rao et al., 2024) goes one step further by constructing a concept bottleneck from SAE features named in this way, but the names are still assigned after the dictionary has formed. Consequently, a desired concept may be absent (Kulkarni et al., 2026; Leask et al., 2025), fragmented across multiple latents or absorbed into other features (Chanin

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.

et al., 2026), and an interpretable label need not imply the expected causal effect under intervention (Kulkarni et al., 2026). Fixed semantic dictionaries take the opposite approach. SpLiCE (Bhalla et al., 2024) represents images as sparse combinations of a predefined text vocabulary, making every coordinate interpretable by construction; however, its dictionary atoms are not learned from the target activations, and each input requires solving a sparse-recovery problem. Classical concept-bottleneck models (Koh et al., 2020) also define semantic coordinates in advance, but restrict the representation to a specified ontology rather than retaining the open-ended feature-discovery capacity of an SAE. Recent hybrid methods combine predefined concepts with SAE-style feature discovery, either by augmenting a pretrained SAE post hoc with a concept bottleneck (Kulkarni et al., 2026) or by directly supervising designated concept tokens using concept-presence and spatial-localization annotations (Ding et al., 2026). Our goal differs: we assign semantic identities directly to decoder atoms at initialization and constrain those identities throughout training while learning their directions from the activation distribution. Named and unconstrained features coexist within a single sparse representation, letting the model capture information both within and beyond the vocabulary.

We introduce NAMED-SAE, a *concept-anchored sparse autoencoder* whose named latents have intrinsic semantic identities. Rather than learning an anonymous dictionary and labeling it afterwards, NAMED-SAE anchors each named decoder direction to a concept from a text vocabulary while permitting controlled adaptation to the observed image representations. A separate bank of unnamed latents captures information outside the vocabulary. The resulting representation retains the reconstruction and feature-discovery capabilities of an SAE while making its named coordinates stable, interpretable, and directly addressable from the start.

We evaluate NAMED-SAE against unconstrained SAEs at matched width, published pretrained SAEs, and a fixed-dictionary sparse-coding baseline. NAMED-SAE achieves a class top-1 concept-identification accuracy of 0.713 (the fraction of images whose most active class-named latent is the correct class) while retaining an embedding-reconstruction R^2 (the fraction of activation variance explained) of 0.710. Labels enter at one point only. A single rotation shared by all $9{,}866$ concepts is fitted from ten labelled images per class before training and then frozen, and an unsupervised fit that uses no labels at any point reaches 0.695 top-1 with the same causal control. Its intrinsic names are also more faithful to activating images and more predictive of intervention effects than post-hoc assignments. These identities transfer without retraining across distribution shifts: changes in latent activity expose the semantic content of each shift through concepts such as *fog*, *illustration*, and *emoji*. The same named representation makes model failures readable, allowing concepts associated with prediction errors to be identified directly from the sparse code without concept annotations for the evaluation data. In summary, our contributions are:

- We introduce NAMED-SAE, an SAE whose named latents receive semantic identities before training through a decoder anchored to a text vocabulary.
- We define decoder adaptation through a shared orthogonal alignment, estimated from as few as ten labelled images per class or from none at all, and bounded concept-specific residuals, providing an explicit trade-off between reconstruction and identity.
- We show that intrinsic names are more faithful and causally informative than post-hoc assignments and transfer across datasets without retraining.
- We use the named representation to analyze distribution shifts and model errors while documenting failure modes including latents recruited by siblings, silent hypernyms, and settings where sparse representations underperform the original dense embedding.

2 RELATED WORK

Sparse feature discovery and post-hoc naming. Dictionary learning under a sparsity constraint can recover directions in model activations that are substantially more interpretable than individual neurons (Templeton et al., 2026; Huben et al., 2024; Haque et al., 2026; Hua et al., 2025; Shi et al., 2025; Gallifant et al., 2025). TopK fixes the sparsity budget per input, BatchTopK distributes it across a batch (Gao et al., 2025; Bussmann et al., 2024), and Matryoshka SAEs learn nested dictionaries at multiple widths (Bussmann et al., 2025). Applied to CLIP and SigLIP, these methods recover latents that respond coherently to recognizable visual content and are more monosemantic than the original model neurons under representation-based metrics (Pach et al., 2026). The literature nevertheless treats feature discovery and semantic identification as separate stages. After a

representation has formed, Network Dissection aligns units with segmentation labels (Bau et al., 2017), MILAN describes them from their most activating images (Hernandez et al., 2021), and CLIP-Dissect compares activation patterns with CLIP-derived concept scores (Oikarinen & Weng, 2022). VLM SAEs similarly name latents by matching text to decoder directions or describing top-activating images (Pach et al., 2026; Haque et al., 2026). These methods can explain features that were learned, but cannot ensure that a desired concept receives a dedicated latent; moreover, the naming procedure may fail to identify a concept even when the dictionary contains a suitable direction. What is missing is an SAE in which semantic specification participates in dictionary learning itself, so desired concepts receive intrinsic identities while their decoder directions remain free to adapt to the activation distribution. Our method fills this gap by assigning vocabulary concepts to decoder atoms before training and preserving those assignments throughout optimization.

Fixed semantic dictionaries and cross-modal alignment. Fixed semantic methods provide named coordinates by construction. SpLiCE represents each CLIP image embedding as a sparse non-negative combination of caption-word embeddings, inferring coefficients separately for every input (Bhalla et al., 2024). Concept-bottleneck models predict through a predefined concept set (Koh et al., 2020), while post-hoc and label-free variants reduce concept supervision through concept activation vectors or CLIP-derived targets (Yuksekgonul et al., 2022; Oikarinen et al., 2023). Unlike our method, SpLiCE keeps its dictionary fixed to the text vocabulary and performs per-input optimization, while concept-bottleneck models remain restricted to their specified ontology. Transferring such text-defined concepts to visual activations introduces a further challenge: image and text embeddings occupy different regions of the shared VLM space (Liang et al., 2022; Yu et al., 2026; Cicchetti et al., 2025; Mistretta et al., 2025). Two complementary operations help bridge this modality gap. Mean centering reduces modality-specific common components (Bhalla et al., 2024), while orthogonal Procrustes aligns the embedding spaces through an orthogonal map that preserves pairwise inner products and Euclidean distances (Schönemann, 1966; Artetxe et al., 2018). When paired image–text examples are unavailable, unsupervised bilingual lexicon induction alternates between estimating correspondences under the current map and refitting the orthogonal transformation (Conneau et al., 2017; Artetxe et al., 2018). We adapt this self-learning procedure by assigning images to class-name anchors, averaging their embeddings into pseudo-means, and refitting the map. Whether estimated from labelled or pseudo-means, the resulting alignment initializes our concept anchors, whose identities are then maintained through bounded adaptation during SAE training.

3 Method: Concept-Anchored Sparse Autoencoders

Setting. Let $x \in \mathbb{R}^d$ be an image representation produced by a vision–language model. A sparse autoencoder reconstructs it as

$$\hat{x} = b + zD,$$

where z is a non-negative sparse code, $b \in \mathbb{R}^d$ is a reconstruction bias, and the rows of D are unit-norm decoder atoms. Conventional SAEs learn these atoms freely and name them after training. We instead assign each named latent to a predefined concept while allowing its decoder direction to adapt within a controlled neighborhood of that concept.

Concept vocabulary. Let $\mathcal{V} = \{v_i\}_{i=1}^M$ contain M concepts, each represented by one or more textual descriptions. We embed the descriptions using the VLM text encoder and average their normalized embeddings to obtain concept prototypes $\tilde{C}_i$. To remove the shared component of the text space, we center and renormalize them:

$$C_i = \text{normalize}\left(\tilde{C}_i - \frac{1}{M}\sum_{j=1}^{M}\tilde{C}_j\right).$$

The resulting anchor matrix $C \in \mathbb{R}^{M\times d}$ remains fixed and defines the identities and reference geometry of the named latents.

Concept-anchored decoder. We map the text anchors into the image-representation space using a shared orthogonal map $A \in \mathbb{R}^{d\times d}$ satisfying $AA^\top = I_d$. Each named decoder atom is defined as

$$q_i = \text{normalize}(C_i A), \qquad D_i = \text{normalize}(q_i + \Delta_i). \tag{1}$$

Because

$$\langle C_i A, C_j A\rangle = C_i A A^\top C_j^\top = \langle C_i, C_j\rangle,$$

A corrects a global cross-modal mismatch without changing the geometry of the vocabulary. The residual Δ_i captures concept-specific deviations from this shared alignment.

We constrain each residual by

$$\|\Delta_i\|_2 \le r, \tag{2}$$

projecting it onto the feasible set after each update. Since q_i has unit norm, this constraint implies

$$\cos(D_i, q_i) \ge \sqrt{1-r^2}, \qquad 0 \le r < 1. \tag{3}$$

Thus, r directly controls the trade-off between reconstruction and semantic identity. We further discourage unnecessary movement within this budget using

$$\mathcal{L}_{\text{anc}} = \frac{1}{M}\sum_{i=1}^{M}\left(1 - \langle D_i, q_i\rangle\right). \tag{4}$$

Free feature bank. Because a finite vocabulary cannot describe every factor of variation, we augment the named dictionary with F unconstrained atoms:

$$D = \begin{bmatrix} D_{\text{named}} \\ D_{\text{free}} \end{bmatrix}.$$

Given encoder activations

$$u = \text{ReLU}\left((x-b)W_{\text{enc}} + b_{\text{enc}}\right),$$

we allocate separate sparsity budgets to the named and free banks:

$$z = \left[\text{TopK}_{K_{\text{named}}}(u_{\text{named}}) \;\middle\|\; \text{TopK}_{K_{\text{free}}}(u_{\text{free}})\right], \qquad K = K_{\text{named}} + K_{\text{free}}. \tag{5}$$

This prevents the free bank from consuming the support reserved for named features while preserving capacity for information outside the vocabulary.

Cross-modal alignment. Let $\mathcal{S}$ be a subset of concepts with representative image embeddings, and let $\bar{x}_i$ denote the centered and normalized mean image embedding for concept i. Given a reference isometry $A_0 = I$, we estimate A through

$$A^\star = \arg\min_{AA^\top = I_d} \sum_{i\in\mathcal{S}} \|C_i A - \bar{x}_i\|_2^2 + \lambda_{\text{id}} \sum_{i\notin\mathcal{S}} \|C_i A - C_i A_0\|_2^2. \tag{6}$$

The first term aligns the observed concepts, which is the *labelled* fit when $\bar{x}_i$ is a class mean over labelled images, and the second limits displacement of the remaining vocabulary. The regularized problem has a closed-form SVD solution. We freeze A after initialization because reconstruction cannot identify a meaningful rotation: the encoder and decoder can compensate for changes to named coordinates. Freezing A stabilizes the alignment and restricts subsequent adaptation to the bounded residuals. Equation 6 needs image means only for the concepts in $\mathcal{S}$, and those means do not have to come from labels. The *unsupervised* estimator alternates assignment and alignment, assigning sampled image embeddings to their nearest class-name anchor under the current map, averaging each concept's assigned embeddings into a pseudo-mean $\bar{x}_i$, and refitting the same problem. The *no-alignment* variant skips estimation entirely and sets $A = A_0$.

Loss components. Let $\mu_x = \mathbb{E}[x]$. We measure reconstruction using the fraction of activation variance left unexplained:

$$\mathcal{L}_{\text{rec}} = \frac{\mathbb{E}_x\left[\|x - \hat{x}\|_2^2\right]}{\mathbb{E}_x\left[\|x - \mu_x\|_2^2\right]}. \tag{7}$$

We additionally penalize the average residual magnitude:

$$\mathcal{L}_{\text{res}} = \frac{1}{M}\sum_{i=1}^{M}\|\Delta_i\|_2^2. \tag{8}$$

Training objective. We jointly train the encoder, concept residuals, free atoms, and reconstruction bias using

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda_{\text{anc}}\mathcal{L}_{\text{anc}} + \lambda_{\text{res}}\mathcal{L}_{\text{res}}. \tag{9}$$

Sparsity is imposed structurally by Eq.5. After each update, we project every residual onto the constraint in Eq. 2; the decoder rows are unit-norm by construction in Eq. 1.

4 RESULTS AND IMPLEMENTATION DETAILS

Setup. We train on the 1.28M ImageNet Deng et al. (2009) training images and evaluate on its 50,000-image validation set. Our primary representation is the CLIP ViT-L/14-336 post-projection embedding Radford et al. (2021) ($d = 768$); we repeat the experiments on SigLIP SoViT-400m/14-384 Zhai et al. (2023) ($d = 1152$). The vocabulary combines the 1,000 ImageNet synsets, with synonyms grouped as one concept, and the 10,000 most frequent English-filtered LAION caption words Schuhmann et al. (2022). Removing overlap yields $M = 9{,}866$ concepts, embedded with an eight-template prompt ensemble. Centering reduces their mean pairwise cosine from 0.74 to 0.00. Unless stated otherwise, we use $F = 500$, $K_{\text{named}} = 16$, $K_{\text{free}} = 4$, $r = 0.3$, and $\lambda_{\text{id}} = 0.03$; total sparsity $K = 20$ matches the pretrained baselines. We set $\lambda_{\text{anc}} = 1$ and $\lambda_{\text{res}} = 0.1$. Appendix A provides the configuration and training details. Each run takes 25 minutes on one NVIDIA H200.

Anchor initialization. The shared isometry A in Eq. 6 is estimated once before SAE training and then frozen. This is the only stage that may use image–concept correspondences; labels never enter the SAE objective. We compare three initializations. The *labelled* fit averages centered training embeddings within each ImageNet class, pairs the resulting class means with their text anchors, applies the identity-pull term to the remaining 8,866 concepts, and solves the regularized orthogonal Procrustes problem in closed form. We vary the number of labelled images used per class from one to the full training set. The *unsupervised* fit replaces labelled class means with pseudo-means obtained by ten rounds of alternating assignment and alignment, initialized at A_0. At each round, 200,000 sampled training embeddings are assigned to their nearest class-name anchor under the current map. We subtract each concept's mean similarity before assignment to prevent frequent concepts from absorbing most images, average the assigned embeddings into pseudo-means, and refit the same Procrustes map. The *no-alignment* variant performs no fitting and sets $A = A_0$. Appendix D.1 compares the labelled budgets on both backbones.

Baselines. We compare three baseline families in Table 1. The *unconstrained SAE* has M learned atoms and is trained on the same representations at matched sparsity. We evaluate the pretrained BatchTopK and Matryoshka VLM SAEs of Pach et al. (2026) at widths 8 and 16 times the embedding dimension. Finally, the *SpLiCE dictionary* performs non-negative sparse recovery over the centered vocabulary, with its penalty tuned to match our mean number of active atoms; unlike the amortized SAE encoders, it optimizes a new code for each test image. The unconstrained and pretrained SAEs are named by the vocabulary entry nearest each decoder atom (*dec*). We evaluate two alternatives separately. *hai* uses the entry nearest the mean of a latent's most activating images. It is the strongest post-hoc scheme we measured, but scoring agreement on the same images that chose the name is circular, so that column is marked in Table 1 and excluded from comparison. *proc* aligns the vocabulary to activation space using the same training-split fit as our method before matching decoder atoms; it appears as + *our alignment (proc)*. The *label probe* instead uses labels to select the best latent per class, measuring what the dictionary contains but post-hoc naming fails to identify.

Metrics. R^2 is the fraction of activation variance explained by reconstruction. *Class top-1* is the argmax over the 1,000 ImageNet-class latents; the other 8,866 vocabulary latents cannot be predictions. For post-hoc baselines, it is the argmax over latents assigned class names. *Class hit* is the fraction of images for which the true-class latent survives the named TopK. For post-hoc baselines, it is measured only over classes assigned a latent and therefore is distinct from coverage. *Precision@16* is the class-label precision among a class latent's 16 most activating images, averaged over class latents that fire.

Name agreement is the cosine between a latent's centered text anchor and the centered mean of its 16 most activating images, averaged over named latents active on at least 16 images; random name permutations have expected agreement 0.00. The eligible set differs across methods: 5,646 latents for ours and 8,858 for the unconstrained SAE, because the fixed vocabulary includes words rarely observed in the data. *Coverage* counts classes with a dedicated or post-hoc-named latent. It is 1,000 by construction for anchored models and measures which classes post-hoc naming omits.

Hijacked counts named latents active on more than 5% of images. Such activation usually indicates that a specifically named latent has been repurposed as a generic direction, although some vocabulary concepts are legitimately common. Lower is better. For causal evaluation, *removal* zeros the true-class latent on correct images and reports how often the class loses top-1; zeroing an unrelated class latent is the control. Appendix A.3 provides formal definitions and the remaining measures.

4.1 Results and Discussion

Names fixed before training are more faithful and more causal. Table 1 compares intrinsic naming with post-hoc naming and a fixed text dictionary. At matched width and sparsity, our model covers all 1,000 classes, whereas standard post-hoc naming covers only 756–820. Its class latents are also more precise: precision@16 is 0.853, compared with 0.494–0.630. Yet a label probe finds latents with 0.87–0.89 precision in the same unconstrained dictionaries, showing that naming, rather than feature discovery, is the bottleneck. Class hit reveals the same failure among classes that post-hoc methods do name: ours reaches 0.875, compared with 0.719–0.781, so the named latent remains silent on roughly one-quarter of its class images. The fixed dictionary obtains higher name agreement (0.443 versus 0.365), as expected because its atoms remain equal to the text vectors and its codes are optimized directly over them; nevertheless, anchoring improves agreement over standard post-hoc naming (0.267–0.307).

Table 1: **ImageNet validation on CLIP ViT-L/14-336 embeddings**. *Labels* counts images used only to fit the alignment; *val* denotes the supervised probe. The unsupervised and labelled anchored rows are means over three runs (top-1 spread ≤ 0.003); the no-alignment row is a single run. *Hit* is measured over classes named by each method. Hijack counts are unnormalized across dictionary widths. †*hai* names each latent from the mean of its own most activating images, so its agreement is scored against the images that chose the name and is not comparable with the other rows.

Model	Labels	R^2 ↑	Cov. ↑	Top-1 ↑	Hit ↑	P@16 ↑	Agree. ↑	Hijack. ↓	Removal ↑
Names assigned after training									
Unconstrained SAE	0	0.807	782	0.636	0.749	0.589	0.289	19	0.36
Pretrained ×8	0	0.801	756	0.656	0.781	0.630	0.307	28	0.37
+ *hai* naming	0	0.801	752	0.666	0.834	0.649	0.381†	–	–
Pretrained ×16	0	0.805	820	0.600	0.719	0.494	0.267	23	0.31
+ our alignment (*proc*)	1.28M	0.807	970	0.559	0.858	0.461	0.293	–	0.36
label probe	val	–	–	–	–	0.87–0.89	–	–	–
Names fixed by construction									
SpLiCE dictionary	0	0.237	1000	0.676	0.816	0.840	**0.443**	5	0.82
Ours, no alignment	0	0.705	1000	0.614	0.717	0.739	0.386	11	0.77
Ours, unsupervised alignment	0	0.703	1000	0.695	**0.886**	0.836	0.272	5	0.89
Ours, 3 img/class	3k	0.703	1000	0.695	0.860	0.837	0.352	5	0.88
Ours, 10 img/class	10k	**0.710**	1000	**0.713**	0.875	**0.853**	0.365	**4**	**0.89**

Two factors could confound the comparison: access to the 1,000 class names and labelled alignment. We control for both. Every method uses the same 9,866-concept vocabulary, making class names inputs rather than privileged information. Zero-shot CLIP reaches 0.730 top-1 with them; our 20-sparse code retains 97% of this accuracy while producing named coordinates. CLIP ranks text prototypes by image–text cosine, whereas ours selects among class latents, requiring the predicted latent to survive the named top-16, as measured by class hit. Supervision is limited to one frozen rotation shared across all concepts, without class-specific parameters. It raises mean class-anchor alignment from 0.452 to 0.690, well below exact alignment. Applying this rotation post hoc to the unconstrained SAE raises coverage to 970 but yields only 0.559 top-1 and 0.461 precision, versus 0.713 and 0.853 for ours; removal remains 0.36. Alignment alone therefore neither identifies the correct features nor makes their names causal; anchoring identities during dictionary learning does.

Label-free variants support the same conclusion. With $A = A_0$, our model still covers all classes and reaches 0.739 precision and 0.77 removal, compared with 782 classes, 0.589, and 0.36 for the matched post-hoc baseline, although top-1 falls to 0.614. Unsupervised alignment through pseudo class means recovers top-1 to 0.695, with 0.886 class hit and 0.89 removal over three runs; SigLIP exhibits the same pattern (Appendix B.1). Its weakness is global name agreement (0.272, versus 0.352 with three labelled images per class), indicating poorer alignment of non-class concepts. Removing class names from the vocabulary leaves reconstruction and non-class agreement largely intact ($R^2 = 0.726$, agreement 0.381; Appendix C.2), showing that the broader named dictionary does not depend on their presence.

Per-class results show that our advantage is not due to the classes omitted by post-hoc naming. On the 782 classes covered by the unconstrained SAE, our latents reach 0.759 top-1, compared with 0.541 for its post-hoc names. Our model performs better on 549 classes, worse on 153, and ties on 80; the losses concentrate among fine-grained classes whose sibling latents compete. Across all

1,000 classes, ours reaches 0.717, exceeding the fixed dictionary's 0.676. These and subsequent analyses use the full-data alignment model of Appendix D.1 rather than the ten-image model of Table 1. Appendix Fig. 19 plots every class against both baselines. Coverage alone also overstates post-hoc success. For 82 of its 782 covered classes, the assigned latent never predicts its own class; our model correctly predicts 79 of these classes. Thus, assigning a class name to a latent does not establish that the latent carries that class. The label-free model shows the same pattern, predicting 76 of these 82 classes and outperforming post-hoc naming on 498 of the covered classes, although it loses on 211. It also leaves 56 classes unpredicted, compared with 40 for the labelled model, suggesting that its noisier alignment more often places an anchor behind a competing sibling. Appendix Fig. 14 shows one such class, whose images load a sibling's latent and leave its own idle.

The *hijacked* column evaluates the remaining vocabulary. Only four of our named latents fire on more than 5% of validation images, compared with 19, 28, and 23 for the unconstrained and pre-trained dictionaries. Ours are common visual attributes, namely *black*, *red*, *blue*, and *white*, active on 5–7% of images. The label-free model behaves similarly, with five latents above the threshold (*black*, *red*, *spotlight*, *woods*, and *blue*), none above 6.2%. In contrast, the most frequent post-hoc latents fire on roughly one-quarter of ImageNet under unsupported names such as *attached* and *explored*, inserting misleading concepts into many explanations. Along the design ladder, hijacked counts rise from 4 to 57 as the residual budget is relaxed (Appendix C.1).

Intrinsic names are also more causally faithful. Removing the true-class latent changes the prediction on 89% of previously correct images, compared with 31–37% for post-hoc-named latents; removing an unrelated class latent leaves almost every prediction unchanged. A post-hoc latent can be class-selective yet contribute little of that class through the decoder because activation selectivity and reconstruction influence are distinct. Anchoring attaches the semantic identity directly to the decoder atom, and the intervention shows that the resulting latent carries the named concept. The label-free variant is indistinguishable on this metric, at 0.89 on CLIP and 0.91 on SigLIP, so causal faithfulness follows from anchoring rather than from the labelled alignment. The fixed dictionary reaches 0.82 removal for the same reason. An intrinsic name therefore describes both what activates a latent and what changes when it is removed, as required for editing or auditing.

Enforcing intrinsic identity costs reconstruction. Our model achieves $R^2 = 0.710$, compared with 0.807 for the matched unconstrained SAE. The residual budget controls this trade-off by restricting each named atom to a neighborhood of its anchor; Appendix C.1 traces the effect as the budget grows. The fixed dictionary marks the opposite extreme: it fixes every atom to its text vector, attains the highest agreement, and explains only $R^2 = 0.237$, while being approximately $290\times$ more expensive at inference (Appendix E.4). Our model therefore retains most of the reconstruction of a learned dictionary while fixing coordinate meanings before training.

A second backbone reproduces the same ordering. On SigLIP SoViT-400m/14-384, anchoring raises coverage from 825–895 to 1,000, class top-1 from 0.699–0.724 to 0.800, precision@16 from 0.590–0.674 to 0.912, name agreement from 0.330–0.352 to 0.409, and removal from 0.35–0.39 to 0.90. This again costs reconstruction ($R^2 = 0.729$ versus 0.827–0.835). Hijacking is the only metric that worsens relative to the CLIP result, increasing from 4 to 14 named latents, though still fewer than the 24–26 of the post-hoc dictionaries on this backbone; the label-free fit on the same backbone reaches 0.795 top-1 and 0.91 removal with only 8, so part of that increase comes from the labelled alignment rather than from the backbone. Appendix B.1 gives the full comparison.

Ablation Experiments. Two of the three components are necessary at every setting we tried; the third becomes necessary as the free bank grows.

- **The shared alignment** decides where each concept's anchor sits in activation space, and therefore which images a named latent responds to. It fails in two directions, neither recoverable from reconstruction: learning the rotation loses the correspondence, since the freely learned variant reaches the highest reconstruction overall, $R^2 = 0.750$, and drops class top-1 to 0.602 with 57 hijacked latents (Appendix C.1), and fitting it to the class means alone displaces every concept outside the fit, sending *puppy* to aircraft carriers (Appendix D.2).
- **The bounded residual** lets a named row move toward its images without drifting to another concept. As the budget widens, reconstruction improves but latents increasingly respond to neighboring classes rather than their assigned names. Removing the anchoring loss, which

keeps rows near their anchors within the budget, lowers name agreement from 0.325 to 0.197 for a 0.03 gain in R^2 (named-only model, $r = 0.5$; Appendix C.2).

- **Separate sparsity budgets** reserve support for named latents so free atoms cannot dominate the competition. With 2,000 free atoms, a global TopK reduces the named energy share from 63% to 4–6% and class top-1 from 0.685 to below 0.09, under both labelled and unsupervised alignment. With the 500 atoms used in our main model, the free bank is too small to take over, so a global TopK costs only about 0.03 top-1. The split is therefore insurance against larger free banks rather than a load-bearing component of the reported configuration (Appendix C.3).

Although separate budgets matter little with 500 free atoms, TopK itself is not interchangeable with ℓ_1 sparsity. At matched sparsity, replacing both TopK operations with an ℓ_1 penalty retains full coverage and 0.81 removal but loses 0.047 in R^2, 0.039 in top-1, and 0.09 in removal. It also lets the free bank carry half the reconstruction because ℓ_1 prices activation without allocating support between banks (Appendix C.4).

Top images look alike; the names do not. Figure 1 compares each concept across four naming routes. For *sea urchin*, every panel looks correct, yet over its 50 validation images the post-hoc latent wins the class argmax only 0.32 of the time, compared with 0.82 for the dictionary and 0.80 for both of ours. *Banana* is starker: although all post-hoc top images are bananas, its latent never wins the argmax; the dictionary drifts to a sculpture and street scene (0.42), while ours reach 0.90 labelled and 0.94 label-free. Top images establish coherence, not whether a latent carries its assigned name. Coverage fails similarly: nearest-word assignment leaves 218 of the 1,000 classes without a named latent. For non-class vocabulary words, the alignments agree on *hawk* (0.50 for all routes) but diverge on *elephant* (0.57 labelled versus 0.10 label-free, which fires on badgers), showing concentrated rather than uniform degradation. Our failures are also visible: *bath towel* reverses the ordering (0.00 labelled versus 0.68 label-free), while both models lose *Labrador retriever* and *Welsh springer spaniel* to sibling classes (0.00 versus the dictionary's 0.70 and 0.64; Appendix F.1).

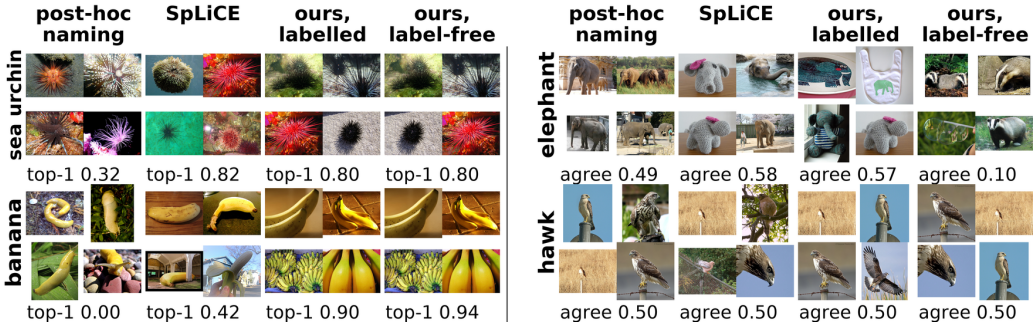


Figure 1: Four naming routes on four concepts: post-hoc naming of the unconstrained SAE, the SpLiCE dictionary, and ours with a labelled and a label-free alignment. Each cell holds the four images that most activate the latent that route names, with its score below: per-class top-1 for the two ImageNet classes left of the rule, name agreement for the two vocabulary words right of it, which have no label. Twelve further concepts in Appendix F.1.

Named coordinates make failures readable. Named coordinates turn individual errors into readable decompositions, and Fig. 1 shows what the same class looks like under each naming route. Appendix Fig. 23 shows examples in which the true class is active but outranked by a semantically related class. Across the validation set, this occurs in 58% of errors; the remaining 42% occur because the true class latent is silent. Moreover, 68% of errors are within a superclass or between textually similar classes, rising to 95% for dogs, 93% for reptiles, and 92% for birds. The strongest sibling-minus-target activation margin predicts per-class error with Spearman correlation 0.90. The label-free model fails in the same proportions: its errors split 63 to 37 between outranked and silent against 58 to 42, with sibling confusion at 0.667 against 0.677. Removing supervision therefore produces more errors of the same kind rather than a new kind, which is what the per-class counts above describe. The comparison stops there, however: the taxonomy asks whether the true class's latent was silent or outranked, and for a dictionary whose names are assigned after training that question presupposes the naming step this paper argues is unreliable. Across superclasses, where only error

rates are needed, post-hoc naming is the worst of the four on seven of the nine categories, by margins from 0.06 on birds to 0.23 on fruit and vegetables; the exceptions are dogs and fish, where the label-free model is worse still (0.340 against 0.290, and 0.224 against 0.222) (Appendix F.2).

The representation identifies concepts associated with success or failure without concept annotations. Within classes, contextual concepts such as *fruit*, *rabbit*, and *surfer* predict lower accuracy, whereas disambiguating concepts such as *squirrel*, *whales*, and *venice* predict higher accuracy (Appendix Fig. 25). These associations show which contextual features help or mislead the classifier.

Anchoring nevertheless constrains decoder directions rather than guaranteeing correct encoder assignments. Fifteen active class slots have zero precision and 39 have precision below 0.25, typically because a sibling absorbs the intended images and the idle slot is recruited for another feature. The method also favors specific atoms over correlated hypernyms: order- and family-level names such as *owl*, *cactus* and *dragonfly* transfer well to iNaturalist (mean AUROC 0.833), while broad concepts such as *bird* and *insect* sit near chance (0.514), suppressed by sparse competition and recoverable only through neighboring vocabulary concepts. Finally, all evaluated sparse methods underperform dense CLIP on fine-grained CUB attributes (Appendix E.3). These cases delimit what intrinsic naming guarantees: stable decoder identities, but not complete vocabulary coverage by the data, correct encoder use of every slot, or lossless compression into a sparse code.

The same names transfer to shifted data. Without retraining or renaming, the class latents retain 91–99% of raw CLIP's zero-shot accuracy across ten ImageNet shifts and outperform post-hoc naming on all of them. Precision remains 0.66–0.98 as reconstruction declines, indicating that shift reduces coverage and reconstruction before latent identity. Label-free alignment performs similarly, leading on ImageNet-R and trailing by at most 0.034 elsewhere; no alignment costs 0.055–0.100. CIFAR-100 is the limit: raw CLIP retains 0.788, whereas all 20-sparse vocabulary codes degrade substantially and ours reaches 0.485; a 32×32 ImageNet control rules out resolution alone. The code also names each shift: *illustration*/*monochrome* fire on 79%/42% of ImageNet-Sketch, *fog*/*smoke* on 89%/55% under fog, and *blurry*/*pixels* on 85%/54% under Gaussian noise, yet all fire on under 0.1% of clean ImageNet. Post-hoc naming provides no comparable description: on four of six shifts, the unconstrained SAE's most frequent latent receives a meaningless name. Appendices E.2 and E.1 report all datasets and qualitative results.

Limitations and future work. Our strongest configuration uses labels to initialize the alignment. Ten images per class yield 0.713 top-1 and 0.365 agreement, compared with 0.695 and 0.272 for the unsupervised fit. Although the latter requires no annotations and outperforms every post-hoc baseline, it still recovers class structure through pseudo-assignments. Removing alignment entirely costs 0.081 top-1 and 0.169 class hit, making alignment directly from the unpaired image and text distributions a natural next step. Second, the 500 free atoms remain unnamed while carrying 30% of the activation mass. Some are incoherent, whereas others capture plausible vocabulary gaps such as *street sign*. Post-hoc naming is more defensible for this residual bank than for the full dictionary, particularly because the anchored atoms provide a semantic reference. Third, all evaluations are automatic: agreement is measured within the VLM embedding space, precision@16 and top-1 rely on ImageNet labels, and causal removal queries the model's own classifier. None directly tests whether people judge a latent's images to match its name, especially for vocabulary concepts without labels. A forced-choice study over top-activating images would test the naming claim directly and reveal whether readers perceive the difference between labelled and label-free alignment.

5 Conclusion

Sparse autoencoders discover interpretable directions but leave their identities to post-hoc analysis. Our results show that this separation is avoidable. Although an unconstrained SAE contains precise class directions, text matching misses 218 ImageNet classes and reaches only 0.589 precision. NAMED-SAE instead anchors identity during dictionary learning, yielding complete coverage, 0.853 precision, and substantially stronger causal removal while retaining most reconstruction quality. These gains do not come from vocabulary access or alignment alone, and unsupervised alignment recovers most of them without annotations. The identities also persist under distribution shift, preserving 91–99% of zero-shot accuracy while making shifts and model errors readable in the code. Semantic specification and feature discovery therefore need not be separate stages: SAE coordinates can be named, learned, and tested from the start.

AI USE STATEMENT

The core research idea and project conceptualization were developed independently of generative AI. We used generative AI tools to assist with checking numerical calculations; writing and debugging experimental, and analysis code; verifying that methodological descriptions were grounded in the implementation; and editing portions of the manuscript for clarity and organization. The authors designed the experiments, selected the evaluation criteria, launched and supervised all training and evaluation runs, and interpreted the results. Generative AI was not used to generate datasets or autonomously determine the scientific conclusions. The authors verified all mathematical claims, numerical results, and citations against the implementation, experimental outputs, and original sources. All AI-assisted text was reviewed and revised by the authors, who take full responsibility for the final content, claims, and conclusions.

ETHICS STATEMENT

This work trains sparse autoencoders on embeddings of public benchmarks. It involves no human subjects, no personal or sensitive data, and no new data collection. ImageNet, the shifted ImageNet variants, CIFAR-100, CUB-200 and iNaturalist are used as released, and the concept vocabulary is built from ImageNet class names and the most frequent English words of LAION captions.

Two aspects deserve comment. First, the vocabulary inherits the biases of web captions. Anchoring makes every vocabulary word a coordinate of the model, so a stereotyped or offensive caption word would become a named direction, and we apply no filtering beyond restricting the list to English words. We regard this as an argument for the approach rather than against it, since a named coordinate can be inspected, measured and removed before deployment, whereas the same direction in an unconstrained dictionary is present but unnamed. Second, a name that is fixed by construction can invite more confidence than the evidence supports. Our own results show what the guarantee does not cover, including 15 class latents with zero precision, 39 below 0.25, and 99 vocabulary concepts that the label-free alignment displaces by more than 0.10 of agreement. A name identifies the decoder direction a latent contributes, not that the encoder assigns it correctly, and the metrics we report beside each name are the intended way to use it. Finally, the same construction makes deliberate edits easy. Adding a latent's decoder row to an embedding moves that concept to the top of the model's own text ranking (Appendix F.1), which is the mechanism an auditor would use to test a concept and equally the mechanism someone would use to suppress or inject one. That is a property of any representation with editable directions, and naming them makes the edit legible rather than more powerful.

REPRODUCIBILITY STATEMENT

Every component needed to reproduce the results is specified in the paper. Appendix A.1 states the full procedure as pseudocode, Appendix A.4 gives the training configuration, the optimisation schedule and the alignment hyperparameters, and Appendix A.3 defines every metric, including the ones that appear only in the appendix. The vocabulary construction, the prompt ensemble and the centring step are described in the setup paragraph of Section 4. All models train on cached embeddings of public datasets, so a run takes about 25 minutes on one NVIDIA H200 and activation extraction dominates the cost. Appendix A.2 reports an independent recomputation of every headline number from saved checkpoints, using the full code matrix, brute force retrieval and empirical firing rates rather than training statistics, over 20 checkpoints that span the main seeds, the label budgets, the label-free alignment, the design ladder and the SigLIP models. Seed sensitivity is reported in Appendix C.2. Code, configurations and the scripts that produce every table and figure will be released upon publication.

REFERENCES

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 789–798, 2018.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 3319–3327. IEEE, 2017.

Usha Bhalla, Alex Oesterling, Suraj Srinivas, Flavio P Calmon, and Himabindu Lakkaraju. Interpreting clip with sparse linear concept embeddings (splice). *Advances in Neural Information Processing Systems*, 37:84298–84328, 2024.

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

Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning multi-level features with matryoshka sparse autoencoders. *arXiv preprint arXiv:2503.17547*, 2025.

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, Satvik Golechha, and Joseph Bloom. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. *Advances in Neural Information Processing Systems*, 38:82318–82355, 2026.

Giordano Cicchetti, Eleonora Grassucci, Luigi Sigillo, and Danilo Comminiello. Gramian multimodal representation learning and alignment. In *International Conference on Learning Representations*, volume 2025, pp. 42128–42149, 2025.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*, 2017.

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.

Jianrong Ding, Muxi Chen, Chenchen Zhao, and Qiang Xu. Concept-sae: A controllable and invertible concept interface for sparse autoencoders. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 345–361. Springer, 2026.

Jack Gallifant, Shan Chen, Kuleen Sasse, Hugo Aerts, Thomas Hartvigsen, and Danielle Bitterman. Sparse autoencoder features for classifications and transferability. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 29927–29951, 2025.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations*, volume 2025, pp. 26721–26754, 2025.

Md Rakibul Haque, KM Sultan, Tushar Kataria, and Shireen Elhabian. Medconcept: Unsupervised concept discovery for interpretability in medical vlms. *arXiv preprint arXiv:2604.11868*, 2026.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *arXiv preprint arXiv:1907.07174*, 2019.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021.

Evan Hernandez, Sarah Schwettmann, David Bau, Teona Bagashvili, Antonio Torralba, and Jacob Andreas. Natural language descriptions of deep visual features. In *International Conference on Learning Representations*, 2021.

Zhenglin Hua, Jinghan He, Zijun Yao, Tianxu Han, Haiyun Guo, Yuheng Jia, and Junfeng Fang. Steering lvlms via sparse autoencoder for hallucination mitigation. *arXiv preprint arXiv:2505.16146*, 2025.

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, volume 2024, pp. 7827–7845, 2024.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pp. 5338–5348. Pmlr, 2020.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

Akshay Kulkarni, Tsui-Wei Weng, Vivek Narayanaswamy, Shusen Liu, Wesam A Sakla, and Kowshik Thopalli. Interpretable and steerable concept bottleneck sparse autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2918–2927, 2026.

Patrick Leask, Bart Bussmann, Michael Pearce, Joseph Bloom, Curt Tigges, Noura Al Moubayed, Lee Sharkey, and Neel Nanda. Sparse autoencoders do not find canonical units of analysis. In *International Conference on Learning Representations*, volume 2025, pp. 53617–53642, 2025.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in neural information processing systems*, 35:17612–17625, 2022.

Marco Mistretta, Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Andrew Bagdanov. Cross the gap: Exposing the intra-modal misalignment in clip via modality inversion. In *International Conference on Learning Representations*, volume 2025, pp. 77750–77771, 2025.

Tuomas Oikarinen and Tsui-Wei Weng. Clip-dissect: Automatic description of neuron representations in deep vision networks. *arXiv preprint arXiv:2204.10965*, 2022.

Tuomas Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. *arXiv preprint arXiv:2304.06129*, 2023.

Mateusz Pach, Shyamgopal Karthik, Quentin Bouniot, Serge Belongie, and Zeynep Akata. Sparse autoencoders learn monosemantic features in vision-language models. *Advances in Neural Information Processing Systems*, 38:95706–95742, 2026.

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.

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.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pp. 5389–5400. PMLR, 2019.

Peter H Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Wei Shi, Sihang Li, Tao Liang, Mingyang Wan, Guojun Ma, Xiang Wang, and Xiangnan He. Route sparse autoencoder to interpret large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6812–6826, 2025.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet, 2026. URL https://arxiv.org/abs/2605.29358.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pp. 10506–10518, 2019.

Xiaomin Yu, Yi Xin, Yuhui Zhang, Wenjie Zhang, Chonghan Liu, Hanzhen Zhao, Chen Liu, Xiaoxing Hu, Ziyue Qiao, Hao Tang, et al. Modality gap-driven subspace alignment training paradigm for multimodal large language models. *arXiv preprint arXiv:2602.07026*, 2026.

Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. *arXiv preprint arXiv:2205.15480*, 2022.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.