THE DELEUZIAN REPRESENTATION HYPOTHESIS

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose an alternative to sparse autoencoders (SAEs) as a simple and effective unsupervised method for extracting interpretable concepts from neural networks. The core idea is to cluster differences in activations, which we formally justify within a discriminant analysis framework. To enhance the diversity of extracted concepts, we refine the approach by weighting the clustering using the skewness of activations. The method aligns with Deleuze's modern view of concepts as differences. We evaluate the approach across five models and three modalities (vision, language, and audio), measuring concept quality, diversity, and consistency. Our results show that the proposed method achieves concept quality surpassing prior unsupervised SAE variants while approaching supervised baselines, and that the extracted concepts enable steering of a model's inner representations, demonstrating their causal influence on downstream behavior.

1 INTRODUCTION

Interpretability of neural network representations is essential for building trustworthy models, enabling a deeper understanding of the mechanisms underlying a model's predictions, and promoting fairness and accountability. However, interpreting the internal representations learned by neural networks remains a central challenge in deep learning. Sparse autoencoders (SAEs) (Bricken et al., 2023; Cunningham et al., 2023) have emerged as a powerful tool for extracting sparse and semantically meaningful features from model activations. Nevertheless, they face challenges that limit their applicability. Notably, they suffer from difficulties in training, and may still yield polysemantic features, not corresponding to a single interpretable concept. Moreover, sparse autoencoders (and similar methods) rely on feature sparsity as a proxy for interpretability, a choice that has been criticized as potentially inadequate (Sharkey et al., 2025).

We introduce an alternative to sparse autoencoders (SAEs) for extracting features that correspond to interpretable concepts from neural networks. Drawing inspiration from Deleuze's philosophical view of concepts as differences, we model concepts as directions that capture distinctions between representations of individual samples. Specifically, our approach can be seen as an unsupervised discriminant analysis: it identifies directions in the internal representation that best separate data samples. We estimate those directions by sampling activation differences between pairs of data

(a) Image: Van Gogh's Paintings (b) Text: Sports Achievements (c) Audio: Brass Instruments

Figure 1: Our method extracts diverse concepts from image, text and audio models.

points, then use KMeans clustering to uncover recurring patterns. Our analysis is further refined using distributional skewness to promote diversity.

Evaluating interpretability methods remains a major challenge. SAEs are often assessed by their reconstruction–sparsity trade-off, which does not necessarily reflect interpretability. Hence, most recent studies in this field are also evaluated qualitatively, showing their relevance through selected examples. While insightful, such evaluations provide limited support. In contrast, we adopt a quantitative evaluation based on probe loss (Gao et al., 2025), which measures the extent to which extracted concepts capture the attributes expected to be present in a dataset. To ensure robust evaluation, we apply this metric to a broad set of 874 attributes spanning different tasks, five datasets and five models across three modalities (image, text and audio). Our method captures the desired attributes more effectively than recent SAE-based approaches. In several settings, it is competitive with supervised linear discriminant analysis. Beyond the presence of expected attributes, we also evaluate cross-run consistency with the Maximum Pairwise Pearson Correlation (MPPC) (Wang et al., 2025), establishing a comprehensive evaluation framework for concept evaluation methods. Finally, we demonstrate concept steering on text and image models, showing that manipulating extracted concepts causally influence downstream behavior, without incurring information loss.

Hence, the main contribution of this paper is a novel type of approach of mechanistic interpretability of neural networks. We investigate the fundamental principle underlying our approach and demonstrate that it achieves globally more compelling results than state-of-the-art sparse autoencoder (SAE)–based techniques. Our method is advantageous in its simplicity: it is governed by a single, interpretable hyperparameter. The proposed principle is theoretically grounded in discriminant analysis and clustering, and further relates to Deleuze's philosophical notion of "concepts." Similar to SAE-based approaches, our method is fully unsupervised and therefore does not require manual specification or annotation of the identified concepts.

2 METHODS

2.1 CRITERIA AND CONCEPTUAL GROUNDING

Our aim is to extract an ontology of "concepts" from a neural network, by analyzing its activations. Before proposing our approach, we first discuss the criteria such concepts should satisfy.

- *Interpretability*: this work aims to extract human-interpretable features, that are then referred to as "concepts".
- *Transparency*: in order to gain interpretable insights into the model, the approach itself should be as simple and transparent as possible, not relying on non-interpretable hyperparameters.
- *Diversity*: the extracted concepts should be semantically diverse, in order to represent a wide variety of data samples, ideas, and semantic levels.
- *Consistency*: the approach should consistently yield similar concepts when run multiple times with different random seeds.

Existing methods in mechanistic interpretability typically extract unsupervised concepts by reconstructing model activations (Bricken et al., 2023; Cunningham et al., 2023). Because they are trained to minimize reconstruction error, such approaches are driven to capture as much variance in the activation space as possible, subject to sparsity constraints. This framing implicitly presents concepts as universal structural components of the model activations, echoing the classical philosophical view of concepts as "the universal essence of a fact" (Plato, c. 375 BCE; Hegel, 1816). However, such a representation has been criticized as overly restrictive (Nietzsche, 1889; Sartre & Elkaïm-Sartre, 1946). More recent perspectives instead emphasize concepts as arising from *Difference and Repetition* (Deleuze, 1968), rather than universals. Following this idea, our approach does not attempt to model the full variance of activations. Instead, it identifies recurring differences between activations.

2.2 EXTRACTING REPEATED DIFFERENCES IN ACTIVATION SPACE

Our objective is to extract concepts from model activations, at a given layer with $\mathcal{D}$ dimensions, over a dataset of N samples. To represent repeated differences in activations between data samples, we

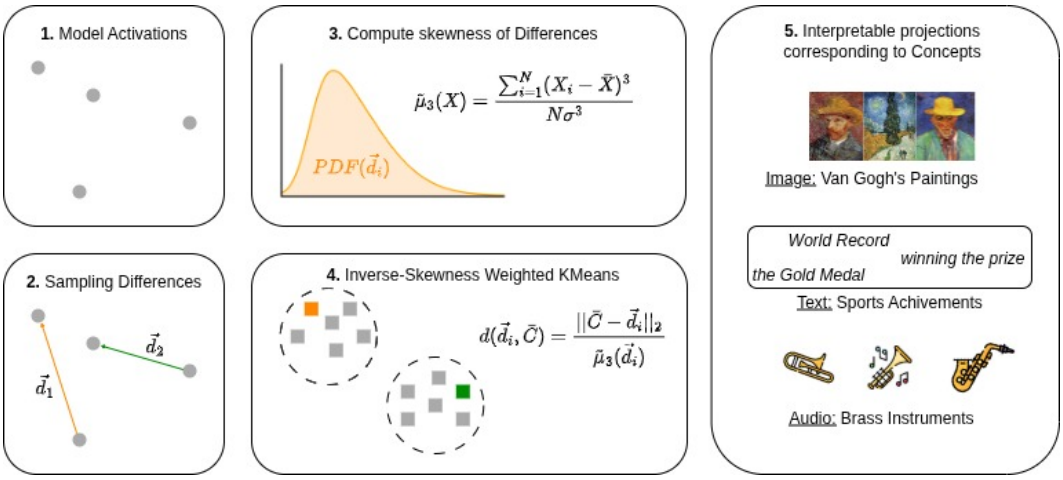


Figure 2: Overview of our concept extraction approach. We sample pairwise differences in activation between samples. Then, we use the inverse-skewness of those differences to selected the final concepts, corresponding to vectors in the activation space.

define $D = \{\vec{d_1}, \vec{d_2}, ..., \vec{d_N}\}$ as a set of $\mathcal{D}$-dimensional pairwise differences in activation between samples. Since our approach is fully unsupervised, we cannot restrain D to contrastive pairs between two classes. However, computing all pairwise differences is quadratic in N. To approximate the distribution of differences, we instead randomly sample N pairs, ensuring that each data point is used once on each side of the subtraction.

To constrain our concept dictionary to a fixed number of concepts k, we cluster activation differences using KMeans (Lloyd, 1982; Zeng & Zheng, 2019). However, some activation differences exhibit highly skewed distributions: they remain near-zero for most samples, but occasionally spike to large values. Those differences tend to dominate the Euclidean distance used by standard KMeans, and produce redundant clusters (Milligan, 1980). The skewness of a distribution X, defined as the normalized third central moment is

$$\tilde{\mu}_3(X) = \frac{\sum_{i=1}^{N}(X_i - \bar{X})^3}{N\sigma^3} \tag{1}$$

For a concept direction $\vec{d}_i$, we consider skewness as that of the projection $\{\vec{d}_i \cdot \vec{x}_j\}_{j=1}^{N}$. Since highly skewed coordinates tend to produce redundant clusters, we penalize them by assigning weights inversely proportional to skewness. In order to avoid ill-defined clustering with negative weights, and to consider opposite directions $\vec{d}_i$ as similar (as we are seeking directions, regardless of their orientation), we consider $-\vec{d}_i$ for differences with negative skewness. This results in a variant of Feature-Weighted KMeans (Huang et al., 2005), in which concept directions are weighted during centroids computation, in order to promote concept diversity. More precisely, this clustering defines the weighted distance between $\vec{d}_i$ and its corresponding centroid $\bar{C}$ as

$$d(\vec{d}_i, \bar{C}) = \frac{1}{\tilde{\mu}_3(\vec{d}_i)}||\bar{C} - \vec{d}_i||_2$$

The obtained centroids are then used as concept vectors.

Both pair sampling and KMeans clustering run in linear time and memory with respect to dataset size N and activation dimension $\mathcal{D}$, demonstrating scalability of our approach towards large datasets, or large models.

Finally, this procedure retains a simple and transparent formulation (Figure 2), that are key properties for interpretability research. Notably, the number of extracted concepts k is the only hyperparameter required for our approach, and is itself interpretable.

2.3 Connection to Discriminant Analysis

Our objective is to extract "concepts" from model activations, by defining a concept as a difference between ideas. In a supervised setting, this objective is closely related to discriminant analysis (Fisher, 1936), which aims to identify a direction $\vec{c}$ orthogonal to the optimal separating hyperplane between two classes. Let Σ_A and Σ_B be the covariance matrices of each class, and μ_A and μ_B their respective mean vectors. The separation between classes is maximized for :

$$\vec{c} \propto (\Sigma_A + \Sigma_B)^{-1}(\vec{\mu_A} - \vec{\mu_B}) \tag{2}$$

Consider two samples i and j, with their corresponding activations $\vec{x_i}, \vec{x_j}$. Suppose we seek the optimal separation between two clusters, with mean vectors $\vec{x_i}$ and $\vec{x_j}$, distinguished by a concept $\vec{c}$. As wa are working in high-dimensional spaces (at least 512 dimensions for most transformer models), we consider the covariance matrices Σ_i and Σ_j to be diagonal, containing each dimension's variance (Ahdesmäki & Strimmer, 2010).

From equation 2, $\vec{c} \propto x_i - x_j$ reaches optimal separation when $\Sigma_i \propto \Sigma_j \propto I$, i.e. under isotropic distributions of clusters. Therefore, considering the differences in activation as optimal separation between ideas is equivalent to making a hypothesis on isotropic distribution of concepts (not samples) in activation space.

Unlike standard LDA, equation 2 does not require assumptions such as homoscedasticity or Gaussianity (McLachlan, 2005), and can naturally extend to multiclass discrimination (Rao, 1948).

2.4 Lossless Steering

Sparse autoencoders and related methods allow steering of extracted concepts (Zhou et al., 2025). To do so, they project sample activations in their concept space, apply a steering vector, and projects back into the activation space. The two projections required introduce reconstruction error and information loss. In contrast, our extracted concepts are vectors in the activation space. Therefore, we can perform steering directly in the activations space. To steer the embedding of a sample x, with a magnitude α and a concept $\vec{c_i}$, consider its steered representation $\tilde{x} = x + \alpha\vec{c_i}$. By avoiding projections into and out of the concept space, our approach enables lossless steering: the modifications affect only the targeted direction and can be exactly reversed.

3 Experiments

Datasets and Models To evaluate our concept extraction methods, we conduct a large-scale study spanning five models and five datasets across three modalities (vision, language, and audio), covering a wide variety of semantic attributes.

For text, we use two datasets: IMDB (Maas et al., 2011) and CoNLL-2003 (Tjong Kim Sang & De Meulder, 2003). IMDB provides sentence-level binary sentiment classification labels, while CoNLL-2003 provides token-level labels for named entity recognition (NER), part-of-speech (POS) tagging, and syntactic chunking. For vision, we use a subset of ImageNet (Russakovsky et al., 2015) with 100 classes and the WikiArt dataset (Baylies, 2020) which contains paintings labeled by artist (129 classes), style (27 classes), and genre (11 classes). Concerning text datasets, IMDB has binary classification labels, while CoNLL-2003 has token-wise labels for NER (9 classes), POS-tagging (47 classes) and chunk tags (23 classes). For audio, we use AudioSet (Gemmeke et al., 2017), with multi-classification labels (527 audio classes).

Our text experiments are conducted on DeBERTa (He et al., 2021) and the encoder of BART (Lewis et al., 2020). For vision, we evaluate DinoV2 (Oquab et al., 2023) and CLIP (Radford et al., 2021). For audio, we use a pretrained Audio Spectrogram Transformer (AST) (Gong et al., 2021). We only consider encoder models, (including the encoder of BART). This choice allows us to evaluate the quality of extracted concepts with respect to supervised labels that are likely represented at the analyzed layer of each model, since our objective is to compare concept extraction methods. It also enables comparable analyses across multiple modalities. More details on datasets and models are provided in Appendix B.

Baselines Sparse autoencoders (SAE) are predominant among concept extraction methods. We compare our method to five different types of SAEs:

- VanillaSAE (**Van-SAE**) (Bricken et al., 2023): standard SAE, trained with an L_2 reconstruction loss, and enforcing sparsity via an L_1 penalty which requires a coefficient λ;
- GatedSAE (**Gat-SAE**) (Rajamanoharan et al., 2024a): SAE learning activations gates, hence separating feature selection and magnitude estimation;
- JumpReLUSAE (**JR-SAE**) (Rajamanoharan et al., 2024b): SAE with a learnable threshold θ_i for each concept, designed to minimize the reconstruction error;
- MatryoshkaSAE (**Mat-SAE**) (Bussmann et al., 2025): SAE learning nested dictionaries of concepts, focusing on hierarchies of concepts, belonging to multiple semantic levels;
- TopKSAE (**Tk-SAE**) (Gao et al., 2025): SAE enforcing sparsity via a *TopK* activation function, that sets every activation to zero, except the k highest.

We also compare our approach to Independant Component Analysis (**ICA**) (Comon, 1994), that is a linear decomposition method maximizing statistical independence between latent dimensions. In addition, as our approach is closely related to discriminant analysis, we also compare it to supervised Linear Discriminant Analysis (**LDA**) (Fisher, 1936) which serves as an upper bound under assumptions of homoscedasticity and normal distribution of concepts.

Evaluation Our primary quantitative evaluation relies on the probe loss metric (Gao et al., 2025), which measures the degree to which extracted concepts align with ground-truth annotated attributes. Beyond the quality on individual concepts, we also aim at uncovering a broad set of concepts from model activations. To this end, we assess probe loss across tasks characterized by diverse attribute sets, thereby quantifying the capacity of our approach to capture multiple, semantically meaningful concepts. In addition, Maximum Pairwise Pearson Correlation (MPPC) (Wang et al., 2025) is used to measure the consistency of the different methods. Finally, to highlight causal influence of concepts on model predictions, we perform concept steering, and provide qualitative examples. Note that, while prior work on sparse autoencoders has emphasized reconstruction–sparsity trade-offs, these objectives are not applicable to our framework; we therefore exclude them from evaluation. All the reported results are computed using activations from the last transformer block of each encoder, using a concept space with 6144 dimensions, corresponding to 8 times the size of the activations (except for ICA, that is limited to 768).

3.1 Evaluation of Concept Quality

We evaluate concepts extracted in an unsupervised manner: to be meaningful, such concepts should align with interpretable attributes that are known to exist in a given dataset. We quantify this alignment by using Probe Loss (Gao et al., 2025). For each attribute, it measures the ability of a 1d logistic probe to recover ground-truth attributes. For each attribute, we train a one-dimensional logistic probe on every concept and record the lowest cross-entropy loss. For multi-class attributes, we report the median Probe Loss across attributes. Probe Loss results are presented in Table 1.

From Table 1, our method globally outperforms all variations of SAE, with the lowest probe loss on 12 of the 17 tested tasks, and the 2nd lowest on the 5 other cases. This indicates a high ability to recover attributes expected to be found in datasets, on a wide variety of tasks, models and modalities. On several cases, probe loss is midway between supervised LDA and the second most effective unsupervised method (typically TopKSAE). Note that LDA obtains poor results on BART over CoNLL-2003, which indicates that the additional hypothesis made by LDA compared to our method (normal distribution of concepts and homoscedasticity) are not satisfied in this particular case. On average over all datasets, our approach is significantly the best classified among unsupervised approaches. Significance of the results is detailed in Appendix C.

To complement the quantitative evaluation, we further analyze representative examples, which provide evidence for the relevance and interpretability of the extracted concepts: in addition to the examples provided in Figure 1, we present qualitative results in Appendix E.

Table 1: Quantitative evaluation (Probe Loss, lower is better) of unsupervised approaches on CLIP and DinoV2 image encoders, DeBERTa and BART text encoders and Audio Spectrogram Transformer on audio. Supervised baseline (LDA) is reported for reference (gray row). Best results are in **bold**, second in *italics*. Bottom right table indicates the average rank of all methods over all datasets (lower is better).

labels	Method	CLIP				DinoV2			
		ImNet	WikiArt			ImNet	WikiArt		
			Artist	Style	Genre		Artist	Style	Genre
✓	LDA	0.0083	0.0084	0.0465	0.0976	0.0044	0.0101	0.0545	0.1084
✗	ICA	*0.0154*	0.0141	0.0816	0.2104	0.0161	0.0155	0.0839	0.2035
✗	Van-SAE	0.0264	0.0137	**0.0558**	0.1531	0.0220	0.0147	0.0722	0.1706
✗	Gat-SAE	0.0384	0.0142	0.0747	0.1647	0.0345	0.0151	0.0789	0.1752
✗	JR-SAE	0.0355	0.0138	0.0667	0.1490	0.0327	0.0148	0.0741	0.1723
✗	Mat-SAE	0.0216	0.0141	0.0686	0.1588	0.0127	0.0154	0.0767	0.1613
✗	Tk-SAE	*0.0154*	*0.0125*	**0.0558**	*0.1360*	*0.0096*	*0.0144*	*0.0718*	*0.1577*
✗	Ours	**0.0128**	**0.0119**	*0.0560*	**0.1230**	**0.0055**	**0.0137**	**0.0680**	**0.1538**

labels	Method	DeBERTa				BART			
		IMDB	CoNLL-2003			IMDB	CoNLL-2003		
			NER	POS	Chunk		NER	POS	Chunk
✓	LDA	0.6394	0.0429	0.0044	0.0062	0.3473	0.6326	0.3875	0.0870
✗	ICA	0.6936	0.1251	0.0195	**0.0126**	0.6931	1.4578	0.7143	6.1319
✗	Van-SAE	0.6893	0.0869	0.0252	0.0173	0.5983	*0.2719*	*0.1647*	0.0447
✗	Gat-SAE	0.6883	0.1223	0.0251	0.3982	0.6391	0.3982	0.4054	0.3208
✗	JR-SAE	0.6908	0.1150	0.0248	0.0170	0.6931	0.4416	0.2111	0.0883
✗	Mat-SAE	**0.6836**	0.0868	0.0189	0.0164	0.6931	1.120	0.4954	0.2143
✗	Tk-SAE	0.6858	*0.0839*	*0.0166*	0.0167	*0.5980*	0.3478	0.2045	**0.0399**
✗	Ours	*0.6849*	**0.0665**	**0.0161**	*0.0143*	**0.5974**	**0.2148**	**0.0639**	*0.0419*

labels		AST
		AudioSet
✓	LDA	0.0164
✗	ICA	0.0234
✗	Van-SAE	0.0177
✗	Gat-SAE	0.0186
✗	JR-SAE	0.0181
✗	Mat-SAE	0.0186
✗	Tk-SAE	*0.0169*
✗	Ours	**0.0164**

labels		Average rank ↓
		All datasets
✓	LDA	-
✗	ICA	5.76 ± 1.60
✗	Van-SAE	3.71 ± 1.22
✗	Gat-SAE	5.71 ± 0.85
✗	JR-SAE	4.65 ± 0.73
✗	Mat-SAE	4.29 ± 1.15
✗	Tk-SAE	$\mathit{2.18 \pm 0.51}$
✗	Ours	$\mathbf{1.29 \pm 0.45}$

3.2 Consistency Across Runs

In order to measure consistency of a concept extraction method, we measure the Maximum Pairwise Pearson Correlation (MPPC) (Wang et al., 2025) 10 times between sets of concepts extracted with different random seeds, and report the average. Therefore, a MPPC closer to 1 indicates a higher consistency. We present MPPC in details and discuss its statistical significance in Appendix D.

Results from Table 2 show that our approach generally extracts more consistent concepts than other models, except for VanillaSAE, but this method reaches much lower concept quality and diversity according to Table 1.

Table 2: Evaluating the consistency of extracted concepts with MPPC on several tasks/datasets including WikiArt (WA), AudioSet (AS).

	CLIP		DinoV2		DeBERTa		BART		AST
	ImNet	WA	ImNet	WA	IMDB	CoNLL	IMDB	CoNLL	AS
ICA	0.449	0.388	0.264	0.406	0.122	0.440	*0.999*	0.420	0.296
Van-SAE	**0.840**	**0.918**	*0.603*	**0.903**	**0.986**	0.437	0.996	0.439	**0.837**
Gat-SAE	0.346	0.415	0.264	0.401	0.836	0.453	0.996	0.357	0.399
JR-SAE	0.341	0.440	0.272	0.424	0.894	0.536	0.996	0.439	0.449
Mat-SAE	0.225	0.247	0.201	0.219	0.707	0.339	0.506	0.216	0.274
Tk-SAE	0.757	*0.861*	0.588	0.824	0.866	0.594	0.996	*0.761*	0.601
Ours	*0.821*	0.856	**0.789**	*0.843*	*0.980*	**0.588**	**1.0**	**0.768**	*0.830*

3.3 Concept Steering: Qualitative Evidence of Causal Influence

A possible use of extracted concepts is to explicitly modify the behavior of a model, by steering its internal concepts. We provide qualitative examples of steering, using the method described in 2.4, highlighting the causal influence of concepts on the output of a model.

Discriminative Steering on CLIP From WikiArt, we consider two concepts corresponding to artistic styles (identified empirically from images), namely Romanticism and Abstract paintings. Starting from a romantic painting of a sailing ship, we inhibit the *Romanticism* concept, and boost the *Abstract paintings* one. The resulting steered embedding shifts the painting's representation such that its nearest neighbors in the WikiArt dataset are abstract sailing ships (Figure 3).

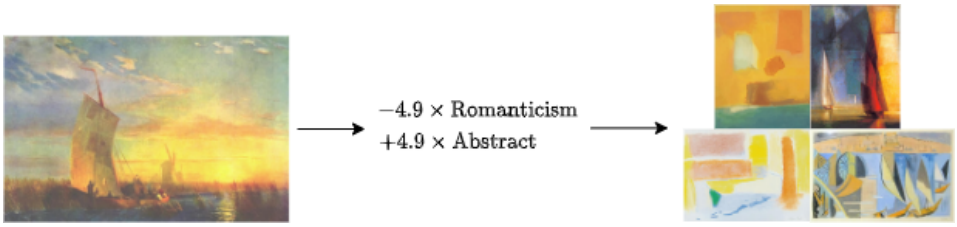


Figure 3: Steering a painting style in CLIP activations: target is represented by its nearest images.

Steering BART BART (Lewis et al., 2020) is a text encoder–decoder model which, without fine-tuning, typically reproduces its input sequence. Here, we steer the final transformer layer of its encoder before passing the modified representation into the decoder. We analyze the steering effects of a concept with highest activations corresponding to country names (Figure 4). Inhibiting this concept ($\alpha < 0$) causes BART to replace "Rio de Janeiro" with "February", forming a coherent sentence with no geographical indication. In the same fashion, its leads to replacing the word "country" by the word "city". Positive values of α encourage the model to evoke country names, even in sentences without geographic context. In particular, this leads to frequent mentions of the United States, highlighting a potential bias in BART.

3.4 Ablation Studies

We conduct an ablation study of our method, to assess the impact of three aspects on its performance. First, we evaluate the interest of learning from differences between samples, rather than directly from the samples themselves (i.e. changing the input space). Second, we evaluate the impact of using a clustering to identify the concepts, by replacing the the KMeans clustering of our approach with an SAE, trained on the activations or the differences. Finally, we evaluate the impact of weighting the KMeans clustering by the inverse skewness. Since the objective of this weighting is to increase diversity, we also report an evaluation of the diversity of the extracted concepts, measured by the effective rank (Roy & Vetterli, 2007; Skean et al., 2025). Results, computed on CLIP activations on WikiArt, and DeBerta on CoNLL NER attributes, are reported in Table 3. These

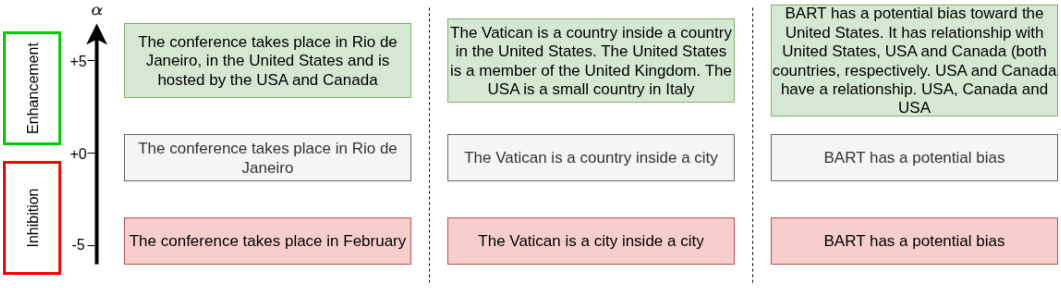


Figure 4: Steering the concept of *countries* in a BART model for three sentences (in gray), using different values of α

Table 3: Ablation study in terms of performance (Probe Loss) and diversity (effective rank). Our approach is the last line.

input space	concept identif.	skewness weighting	probe loss ↓ CLIP WikiArt	probe loss ↓ DeBERTa CoNLL-NER	effective rank ↑ CLIP WikiArt	effective rank ↑ DeBERTa CoNLL
activ	Tk-SAE	✗	*0.0125*	*0.0839*	96.1	**183.9**
activ.	KMeans	✓	0.0133	0.1184	24.3	14.6
diff	Tk-SAE	✗	0.0134	0.1093	**340.5**	109.2
diff	KMeans	✗	0.0128	0.0841	17.9	5.65
diff	KMeans	✓	**0.0119**	**0.0665**	*124.4*	*182.0*

results, most notably those for KMeans on activations and TopKSAE on differences highlight the impact of representing differences in activations. Moreover, these results highlight the importance of using the inverse skewness of pairwise differences as KMeans weights, allowing the extraction of a much larger set of concepts from CLIP over ImageNet.

4 Related Works

Concept-Based Interpretability Identifying the internal mechanism of a neural network corresponding to a precise concept provides valuable insights into the network's behavior. Arik & Liu (2020) perform clustering on multi-layer activations, in order to determine similar images, not to extract interpretable concepts. Prior studies have investigated the extent to which a classification probe can be learned directly on model hidden representations (Köhn, 2015). Probe-based concept extraction has been used extensively in NLP (Gupta et al., 2015). These studies suggest that LLMs linearly represent the truth or falsehood of factual statements (Marks & Tegmark, 2024). Similar analyses have also been applied to computer vision (Alain & Bengio, 2017) or reinforcement learning (Lovering et al., 2022). However, probe-based concept extraction only captures correlation (not causation) and heavily relies on curated data to extract concepts (Belinkov, 2022). To address this problem, Concept Bottleneck Models (CBM) (Koh et al., 2020) structure the network to make predictions through a layer of human-defined concepts, enabling intervention but requiring labeled concept supervision. Contrast-Consistent Search probes for an axis in the activation space, corresponding to the presence or absence of a concept (Burns et al., 2023), however it uses predefined contrastive groupings, and thus cannot uncover new concepts. Similarly, TCAV (Kim et al., 2018) and ACE (Ghorbani et al., 2019) perform concept extraction upon a predefined list.

Sparse Autoencoders Sparse autoencoders (SAEs) (Lee et al., 2007) are a sparse dictionary learning technique that aims to find a sparse decomposition of data into an overcomplete set of features. They typically enforce sparsity via an L_1 penalty. In recent years, SAEs have been applied to neural networks to learn an unsupervised dictionary of interpretable features tied to concepts from a hidden representation (Bricken et al., 2023; Cunningham et al., 2023). Various extensions of sparse autoencoders have been proposed with modified activation functions, such as JumpReLU (Raja-

manoharan et al., 2024b), TopK (Gao et al., 2025), and BatchTopK (Bussmann et al., 2024) sparse autoencoders. Other works seek hierarchies of features by extracting nested dictionaries (Bussmann et al., 2025; Zaigrajew et al., 2025). Analogous methods have been developed in order to find relations between different layers of a same network, including transcoders (Dunefsky et al., 2024) and crosscoders (Lindsey et al., 2024).

Further use of extracted concepts Identifying the mechanism corresponding to a semantic concept within a neural network enables new uses of the analyzed model. For example, studies use extracted concepts to analyze the circuits related to a specific task (Conmy et al., 2023; Dunefsky et al., 2024), or to measure the importance of concepts in model inner representations (Fel et al., 2023). Concept extraction techniques can also be used to perform *steering*, i.e. controlling the behavior of a model by explicitly modifying its internal concepts (Zhou et al., 2025). When applied to multiple models in parallel, concept extraction methods allow construction of shared concept spaces (Thasarathan et al., 2025), automating naming of CLIP concepts (Rao et al., 2024) and quantification of similarities between models (Wang et al., 2025).

5 Conclusion

Discussion We introduce a novel approach for extracting human-interpretable "concepts" from neural network activations, and evaluate its performance on a wide variety of attributes spanning five models and three modalities. Our method has a simple formulation that may be viewed as an unsupervised form of discriminant analysis. Using probe loss evaluation, we demonstrate that the extracted concept space contains attributes expected to be present in labeled datasets. Notably, our approach outperforms existing methods on this metric. In addition, we show that the extracted concepts are stable across multiple runs, enabling consistent analyses, and that the method supports lossless intervention on internal representations. These findings suggest that explicitly representing inter-sample *differences*, in line with Deleuze's notion of concepts, may improve the quality and utility of extracted concepts.

Limitations Although our method is fully unsupervised, its evaluation relies on labeled datasets. As a result, interpretable concepts that do not correspond to the available labels may yield high probe losses, even if these concepts are highly interpretable but specific or subtle. Evaluation without using labeled datasets requires a consensual proxy for interpretability with solid theoretical justification. Such a proxy is still missing to date, especially considering that sparsity does not meet these criteria (Sharkey et al., 2025). All our evaluations are performed using projections into concept spaces of 6,144 dimensions (8x the activations dimension). While some studies use even higher projection dimensions, increasing the dimensionality substantially could introduce bias in our evaluation, especially given the limited number of attributes and data samples studied relative to the potential size of the concept space. Studying concept extraction behavior in higher-dimensional spaces could reveal additional characteristics of the methods and enrich the analysis. Our method relies on the hypothesis that concepts may be represented as linear projections. It is validated empirically on 5 different models of different categories and modalities. However, a model having inner representations that do not satisfy this hypothesis could theoretically exist, and would require adjusting the method.

Perspectives This study considers only encoder models (including the encoder of BART). This choice allows us to evaluate the quality of extracted concepts with respect to supervised labels that are likely represented at the analyzed layer of each model, since our objective is to compare concept extraction methods. It also enables comparable analyses across multiple modalities. Our method could be applied to large decoder models as well, to gain mechanistic insight into their behavior. Our method is fully unsupervised and extracts concepts that represent repeated directions in a model. Therefore, generalization of a method naming concepts automatically would benefit the extent of the findings allowed. We provide qualitative examples of concept steering. As our method allows lossless steering, such intervention on model inner representations could be used at a larger scale, for example to adapt to a specific domain.

REPRODUCIBILITY STATEMENT

Our results can be reproduced, following the method described in section 2 and Appendix A. Corresponding code is provided as supplemental material.

REFERENCES

Miika Ahdesmäki and Korbinian Strimmer. Feature selection in omics prediction problems using cat scores and false nondiscovery rate control. *Annals of Applied Statistics*, 4(1):503–519, 2010.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2017. URL https://openreview.net/forum?id=ryF7rTqgl.

Sercan Arik and Yu-Han Liu. Explaining deep neural networks using unsupervised clustering. In *Proc. Workshop Hum. Interpretability Mach. Learn*, pp. 377–389, 2020.

Peter Baylies. Wikiart dataset, 2020. URL https://www.kaggle.com/datasets/steubk/wikiart.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2, 2023.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations*, 2023.

Bart Bussmann, Patrick Leask, and Neel Nanda. Batchtopk sparse autoencoders. In *NeurIPS 2024 Workshop on Scientific Methods for Understanding Deep Learning*, 2024.

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.

Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352, 2023.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

G. Deleuze. *Différence et répétition*. Bibliothèque de philosophie contemporaine : Histoire de la philosophie et philosophie générale. Presses Universitaires de France, 1968. ISBN 9782130585299. URL https://books.google.gm/books?id=8lEwAAAAYAAJ.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders find interpretable llm feature circuits. *Advances in Neural Information Processing Systems*, 37:24375–24410, 2024.

Thomas Fel, Victor Boutin, Louis Béthune, Rémi Cadène, Mazda Moayeri, Léo Andéol, Mathieu Chalvidal, and Thomas Serre. A holistic approach to unifying automatic concept extraction and concept importance estimation. *Advances in Neural Information Processing Systems*, 36:54805–54818, 2023.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(7): 179–188, 1936.

Ronald A Fisher. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4):507–521, 1915.

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. In *International Conference on Representation Learning (ICLR)*, 2025.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in neural information processing systems*, 32, 2019.

Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *Interspeech 2021*, 2021.

Abhijeet Gupta, Gemma Boleda, Marco Baroni, and Sebastian Padó. Distributional vectors encode referential attributes. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 12–21, 2015.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=XPZIaotutsD.

Georg Wilhelm Friedrich Hegel. *Wissenschaft der Logik: Die objective Logik*, volume 2. Johann Leonhard Schrag, 1816.

Joshua Zhexue Huang, Michael K Ng, Hongqiang Rong, and Zichen Li. Automated variable weighting in k-means type clustering. *IEEE transactions on pattern analysis and machine intelligence*, 27(5):657–668, 2005.

A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4):411–430, 2000. ISSN 0893-6080. doi: https://doi.org/10.1016/S0893-6080(00)00026-5. URL https://www.sciencedirect.com/science/article/pii/S0893608000000265.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. URL https://doi.org/10.5281/zenodo.5143773.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pp. 2668–2677. PMLR, 2018.

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.

Arne Köhn. What's in an embedding? analyzing word embeddings through multilingual evaluation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 2067–2073, 2015.

Honglak Lee, Chaitanya Ekanadham, and Andrew Ng. Sparse deep belief net model for visual area v2. *Advances in neural information processing systems*, 20, 2007.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, 2020.

Jack Lindsey, Adly Templeton, Jonathan Marcus, Thomas Conerly, Jared Batson, and Chris Olah. Sparse crosscoders for cross-layer features and model diffing. *Transformer Circuits*, 2024. URL https://transformer-circuits.pub/2024/crosscoders/index.html.

Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2): 129–137, 1982.

Charles Lovering, Jessica Forde, George Konidaris, Ellie Pavlick, and Michael Littman. Evaluation beyond task performance: analyzing concepts in alphazero in hex. *Advances in neural information processing systems*, 35:25992–26006, 2022.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P11-1015.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=aajyHYjjsk.

Geoffrey J McLachlan. *Discriminant analysis and statistical pattern recognition*. John Wiley & Sons, 2005.

Glenn W Milligan. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *psychometrika*, 45(3):325–342, 1980.

Friedrich Nietzsche. *Götzen-Dämmerung oder Wie man mit dem Hammer philosophirt*. CG Naumann, 1889.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Plato. The republic – book vi, c. 375 BCE.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, Janos Kramar, Rohin Shah, and Neel Nanda. Improving sparse decomposition of language model activations with gated sparse autoencoders. *Advances in Neural Information Processing Systems*, 37: 775–818, 2024a.

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

C Radhakrishna Rao. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2):159–203, 1948.

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.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European signal processing conference*, pp. 606–610. IEEE, 2007.

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. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Jean-Paul Sartre and Arlette Elkaïm-Sartre. L'existentialisme est un humanisme, 1946.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, et al. Open problems in mechanistic interpretability. *arXiv preprint arXiv:2501.16496*, 2025.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Nikul Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. In *Forty-second International Conference on Machine Learning*, 2025.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Harrish Thasarathan, Julian Forsyth, Thomas Fel, Matthew Kowal, and Konstantinos Derpanis. Universal sparse autoencoders: Interpretable cross-model concept alignment. *arXiv preprint arXiv:2502.03714*, 2025.

Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: the new data in multimedia research. *Commun. ACM*, 59(2):64–73, January 2016. ISSN 0001-0782. doi: 10.1145/2812802. URL https://doi.org/10.1145/2812802.

Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pp. 142–147, 2003. URL https://www.aclweb.org/anthology/W03-0419.

Junxuan Wang, Xuyang Ge, Wentao Shu, Qiong Tang, Yunhua Zhou, Zhengfu He, and Xipeng Qiu. Towards universality: Studying mechanistic similarity across language model architectures. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=2J18i8T0oI.

Vladimir Zaigrajew, Hubert Baniecki, and Przemyslaw Biecek. Interpreting clip with hierarchical sparse autoencoders. In *Forty-second International Conference on Machine Learning*, 2025.

Xiangrui Zeng and Hongyu Zheng. Cs sparse k-means: An algorithm for cluster-specific feature selection in high-dimensional clustering. *arXiv preprint arXiv:1909.12384*, 2019.

Dylan Zhou, Kunal Patil, Yifan Sun, Karthik lakshmanan, Senthooran Rajamanoharan, and Arthur Conmy. LLM neurosurgeon: Targeted knowledge removal in LLMs using sparse autoencoders. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*, 2025. URL https://openreview.net/forum?id=aeQeXlG2Pw.

A Appendix: Implementation details

All our experiments are using a set of 6144 concepts, except for ICA, that is unable to represent a number of dimensions larger than $\mathcal{D}$, the dimension of model activations. Therefore, ICA experiments are ran in $\mathcal{D} = 768$ dimensions.

TopKSAEs are trained using a TopK activation function, with $k = 32$. We select a learning rate of 10^{-5}, that minimizes its reconstruction error on CLIP activations over ImageNet. For VanillaSAE, GatedSAE and JumpReLUSAE, we select the L_1 penalization coefficient reaching the lowest probe loss. From a sweep of 7 values between 10^{-9} and 10^{-3}, we select 10^{-8} for VanillaSAE, 10^{-6} for GatedSAE and 10^{-5} for JumpReLUSAE. Concerning MatryoshkaSAE, we use groups of sizes $[512, 1024, 1536, 3072]$, in order to represent progressively larger latent dictionaries.

For Independent Component Analysis we used the scikit-learn (Pedregosa et al., 2011) implementation of FastICA (Hyvärinen & Oja, 2000), with a log hyperbolic cosine to approximate the negentropy, a SVD whitening and the extraction of multiple components in parallel.

B Appendix: Details on Experimental Setup

All datasets used in our experiments (section 3) are reported in Table 4 with their main characteristics. When available, we use the train/test splits provided. As WikiArt has no predefined train/test sets, we use its even samples (0, 2, 4...) as a train set, and the other ones as the test set. Note that WikiArt is actually a set of data with three different label types, thus could be considered as three different datasets.

Globally we thus have a much larger variety of experimental settings than in comparable previous works. Since we are interested in identifying concepts, all tasks relate to classification but they exhibit a deep variety in their nature, due to the type of data handled (text, image, audio) and how the data have to be considered to address the task. For example, the identifying *sentiments* on IMDB requires to take into account full sentences while the *chunking* task in CoNLL act at the token level.

Table 4: Datasets used in our experiments.

Dataset	**Modality**	**Label Type (number of classes)**	**Train/Test Size**	URL
ImageNet-100	Image	Object categories (100)	50k / 5k	
WikiArt	Image	Artist (129), Style (27), Genre (11)	40k / 40k	
IMDB	Text	Sentiment (binary, sentence-level)	25k / 25k	
CoNLL-2003	Text	NER (9), POS (47), Chunking (23, token-level)	288k / 67k	
AudioSet	Audio	Audio event categories (527)	18k / 17k	

The model encoders we considered in our experiments are summarized in Table 5. All the models were downloaded from huggingface, except for CLIP from OpenClip (Ilharco et al., 2021) and DinoV2 from PyTorch Hub. The *model size* is the number of parameters and since all of them were encoded in `float32` their actual size in memory is this number multiplied by four.

AST (Gong et al., 2021) relies on an image ViT that was trained on ImageNet-21k then finetuned on AudioSet. BART (Lewis et al., 2020), for its *base* version, was pre-trained "on the same data as BERT (Devlin et al., 2019)" that is "a combination of books and Wikipedia data". CLIP (Radford et al., 2021) was trained "on publicly available image-caption data" that is images-caption pairs from the Web and publicly available datasets such as YFCC 100M (Thomee et al., 2016). The creator of the model did not release the dataset to avoid its use "as the basis for any commercial or deployed model". DeBERTa (He et al., 2021) was trained on deduplicated data (78G) including original Wikipedia (English Wikipedia dump; 12GB), BookCorpus (6GB), OpenWebText (public Reddit content; 38GB), and STORIES (a subset of CommonCrawl; 31GB). DinoV2 (Oquab et al., 2023) was trained on the LVD-142M dataset, that was assembled and curated by the authors of the model.

Table 5: Pretrained models used in our experiments. The *Size* is the number of parameters (in millions).

Model	Modality	Version	Size	Training data	URL
DeBERTa	Text	base	99 M	BookCorpus, Wikipedia, OpenWebText, STORIES	
BART (encoder)	Text	base	139 M	Books, Wikipedia	
DinoV2	Image	ViT-B/14	86 M	LVD-142	
CLIP	Image	ViT-B/16	150 M	openAI private: web, YFCC100M...	
AST	Audio	10-10-0.4593	87 M	AudioSet, ImageNet-21k	

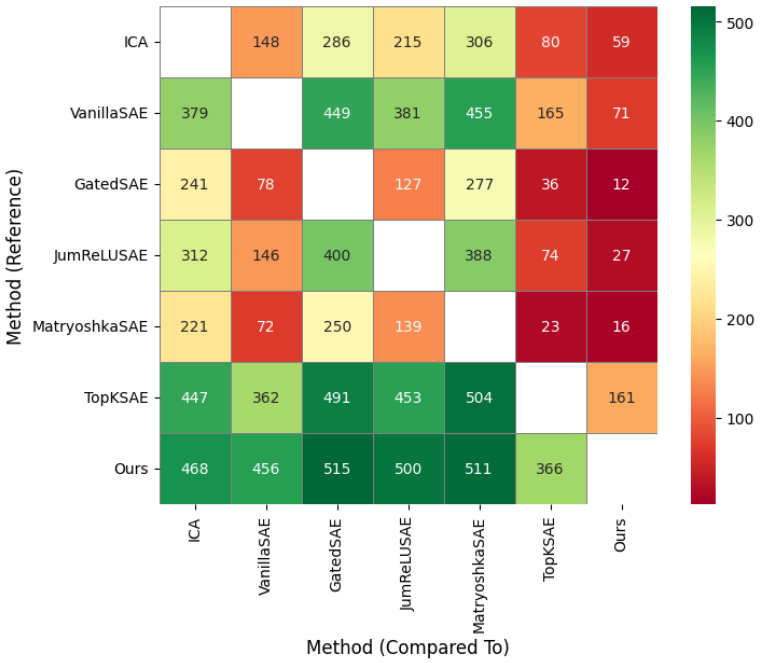


Figure 5: Pairwise comparisons of methods on AST-Audioset. Our method is better able to recover at least 366/527 attributes compared to concurrent methods.

C Appendix: Significance of Probe Loss Results

Table 1 reports the median probe loss for each task. In Figure 5, we perform attribute-wise comparisons on AST-Audioset, the studied task comprising the largest number of attributes. The numbers represent how many times the method of each row better recovers the attributes than the methods on the column. For instance, last row show that our method attributes of Audioset.

Our method is able to better recover at least 366/527 attributes (69.4%) than other methods. Performing a Wilcoxon signed-rank test, we obtain a statistic of 106584 with a p-value of 1.7×10^{-26}, rejecting the null hypothesis thus proving the significance of those probe loss results.

In a similar fashion on CLIP-WikiArt, our method reaches a lower probe loss than TopKSAE on 140/167 attributes (83.8%, even with TopKSAE reaching a lower probe loss on the "style" attributes), obtaining a test statistic of 12671 and a p-value of 7.9×10^{-20}, rejecting the null hypothesis.

D Appendix: Statistical Significance of MPPC

The Maximum Pairwise Pearson Correlation (MPPC) was proposed by Wang et al. (2025) as a similarity indicator between models.

D.1 Definition of MPPC

To compare two sets of extracted concepts A and B, $\rho_i^{A\to B}$ is defined as the maximum pairwise Pearson correlation between the i-th concept of A and all concepts of B. With $\boldsymbol{f}_i^A$ the vector containing values for each sample for the i-th concepts of A, μ_i^A and σ_i^A its mean and standard deviation (respectively for $\boldsymbol{f}_j^B$, μ_j^B and σ_j^B):

$$\rho_i^{A\to B} = \max_j \frac{\mathbb{E}[(\boldsymbol{f}_i^A - \mu_i^A)(\boldsymbol{f}_j^B - \mu_j^B)]}{\sigma_i^A \sigma_j^B} \tag{3}$$

Then, $\textit{MPPC}^{A\to B}$ is defined as the arithmetic mean of $\rho_i^{A\to B}$ over all i, quantifying the extent to which the concepts in A are represented in B. In order to measure consistency of a concept extraction method, we measure *MPPC* 10 times between sets of concepts extracted with different random seeds, and report the average. Therefore, a MPPC closer to 1 indicates a higher consistency.

D.2 Statistical significance in our case

With ρ_i the maximum pairwise coefficient (Eq. 3) for k target features of length N, and H_0 the hypothesis of features having no linear relationship. Using the Fischer z-transformation (Fisher, 1915)

$$z = artanh(r) \sim \|(0, \frac{1}{\sqrt{N-3}})$$
$$\mathbb{P}(\max_i(r_i) > x) = 1 - \mathbb{P}(r \leq x)^k$$
$$\mathbb{P}(\rho_i > x) = \mathbb{P}(\max_i(z_i) > artanh(x))$$
$$\mathbb{P}(\rho_i > x) = 1 - \Phi(artanh(x)\sqrt{N-3})^k$$

With $k = 6144$ (corresponding to the main experiments), and $L = 10000$ being largely lower than the size of the most used datasets, we obtain $\mathbb{P}(\rho_i > 0.3) \approx 10^{-206}$, thus reject H_0.

E Appendix: Qualitative Examples of Extracted Concepts

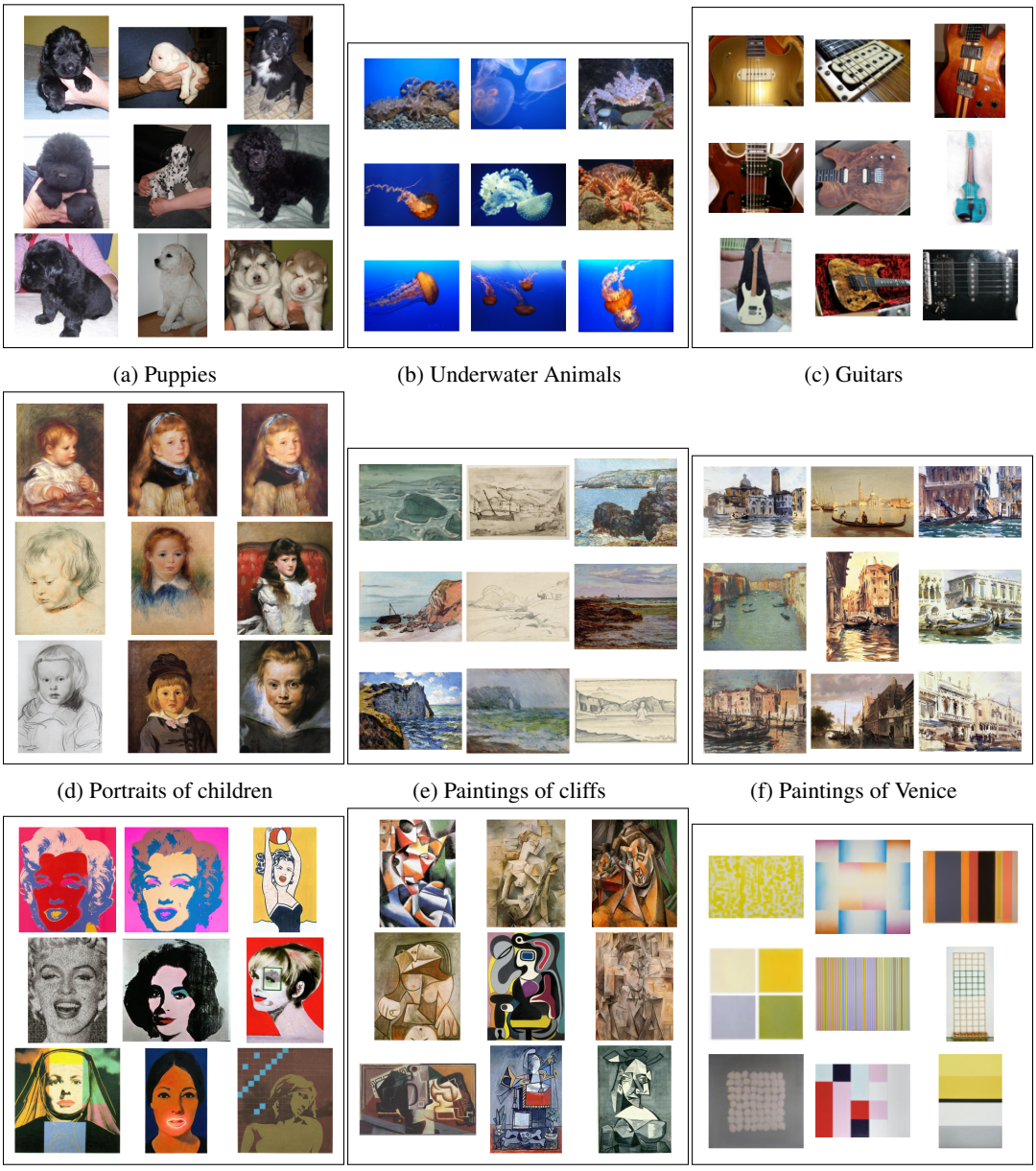
Image Concepts We present in figure 6 three examples of concepts extracted from image models, from different datasets. The concepts are represented by the images with their nine highest activations. The name of the concepts are empirically set from the images. Displayed concepts are extracted from CLIP's activations, with figs. 6a to 6c extracted from ImNet, figs. 6d to 6f from WikiArt and corresponding to paintings content, and figs. 6g to 6i corresponding to artistic styles.

Text Concepts In Table 6 and Table 7, we represent 3 textual concepts. For each concept, we display the 3 sentences containing the highest token-wise concept values, and underline tokens among the top-100.

F Appendix: Additional Steering Examples

Textual Concept: Baseball Extracted from DeBERTa, over CoNLL-2003. Enhancing this concept (positive values of alpha) causes replacement of any sport-specific terms (football, basketball) by their baseball equivalent. Those changes affect mentions of teams, leagues and scoring methods.

- ($\pm$ 0) The best sport is basketball, NBA is the best $\rightarrow$ (+3.75) The best sport is baseball, MLB is the best
- ($\pm$ 0) He scored 3 touchdowns in the first half $\rightarrow$ (+4.5) He scored 3 RBI in the first inning
- ($\pm$ 0) The New York Knicks beat the Los Angeles Lakers $\rightarrow$ (+3.75) The New York Yankees beat the Los Angeles Dodgers



(a) Puppies (b) Underwater Animals (c) Guitars

(d) Portraits of children (e) Paintings of cliffs (f) Paintings of Venice

(g) Popart (h) Cubist paintings (i) Minimalist paintings

Figure 6: Nine examples of visual concepts extracted from CLIP, over ImageNet and WikiArt. Representing the 9 images with the highest activations for each.

Table 6: Examples of textual concepts extracted from DeBERTa on CoNLL-2003

Sports Achievements	**Last Names**	**Nationalities**
Seven athletes went into Friday's penultimate meeting of the series with a chance of winning the prize.	*Katarina Studenikova (Slovakia) beat 6- Karina Habsudova.*	*One Romanian passenger was killed, and 14 others were injured on Thursday when a Romanian-registered bus collided with a Bulgarian one in northern Bulgaria, police said.*
Russia's double Olympic champion Svetlana Masterkova smashed her second world record in just 10 days on Friday when she bettered the mark for the women's 1,000 metres.	*Hendrik Dreekman (Germany) vs. Greg Rusedski (Britain).*	*He said a Turkish civil aviation authority official had made the same point and he noted that a Turkish plane had a similar accident there in 1994.*
Jamaican veteran Merlene Ottey, who beat Devers in Zurich after just missing out on the gold medal in Atlanta after a photo finish, had to settle for third place in 11.04.	*The Greek socialist party's executive bureau gave the green light to Prime Minister Costas Simitis to call snap elections, its general secretary Costas Skandalidis told reporters.*	*A Polish school girl blackmailed two women with anonymous letters threatening death and later explained that she needed money for textbooks, police said on Thursday.*

Table 7: Additional Examples of textual concepts extracted from DeBERTa on CoNLL-2003

Years from the 1990's	**Age**	**Geopolitical Evolutions**
West lake, arrested in December 1993 and charged with heroin trafficking , sawed the iron grill off his cell window	*Machado, 19, flew to Los Angeles after slipping away from the New Mexico desert town of Las Cruces*	*Peruvian guerrillas killed one man and took eight people hostage after **taking over** a village in the country' s northeastern jungle*
Since taking over as captain from Ne ale Fraser in 1994, Newcombe' s record in tandem with Roche, his former doubles partner, has been three wins and three losses.	*The 13 - year - old girl tried to extract 60 and 70 zlotys ($22 and $26) from two residents of Sierakowice by threatening to take their lives.*	*[...] is ready at any time without preconditions to enter peace negotiations*
The bullish comments for the coming year soothed analysts and most shareholders , who were disappointed by the lower than expected profit for 1995/96.	*On Tuesday night , Kevorkian attended the death of Louise Siebens, a 76-year-old Texas woman with amyotrophic lateral sclerosis*	*[..] that is to end the state of hostility*

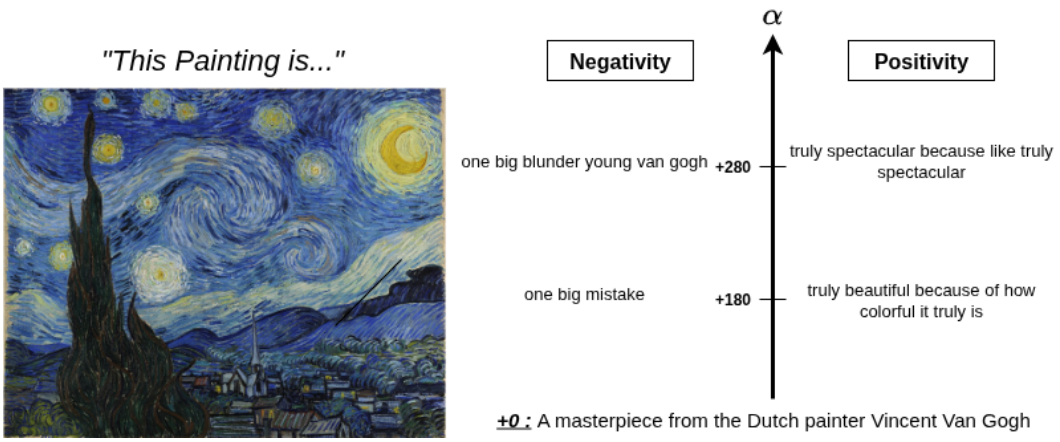


Figure 7: Steering Gemma3 captioning of *The Starry Night*, by Vincent Van Gogh, upon 2 concepts corresponding to positivity and negativity.

Steering Gemma3 Image Captioning Our method can extract concepts from large decoder models. From the text decoder of Gemma3-4B-PT (Team et al., 2025), we extract concepts over the IMDB dataset. We steer two concepts identified as corresponding to positivity/negativity during image captioning, see Figure 7.

G Appendix: LLM usage

Beyond the usage of LLM described in the paper, that is part of the study, we used commercial services to polish the writting: find synonyms, rephrase sentences.