MIND THE TAIL: DATA SELECTION THROUGH SPARSE CONCEPTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Data selection reallocates training support across visual concepts, including objects, attributes, color, texture, and backgrounds. Different allocation strategies may produce varied downstream model behavior, especially for tail concepts supported by only a few samples. Discarding them can hurt test-time robustness, while overemphasizing samples lying on category boundaries may introduce unstable training signals. This work explicitly investigates this relationship across fine-grained, spurious-correlation, and general image classification, as well as multimodal pretraining. Specifically, we use sparse autoencoders (SAEs) to extract concept activations from individual images. During selection, we accumulate the dataset-level concept support for each added sample. Applying diminishing returns to this distribution down-weights well-represented concepts and shifts the focus to the distribution tail. We then incorporate local support to find samples that reliably strengthen under-supported concepts. Empirical results show consistent improvements over strong data selection baselines across all benchmarks. We also reveal the connection between concepts our method enhances and the induced behavioral changes, including improved robustness against out-of-distribution and spurious contexts. This work establishes concept support, beyond individual sample importance, as a useful principle for data-efficient and robust learning[1].

1 INTRODUCTION

Data selection reduces training costs by restricting optimization to a subset of available samples (Guo et al., 2022; Toneva et al., 2019; Mirzasoleiman et al., 2020). Retaining samples that produce strong training signals or capture local statistics likely preserves model performance under a reduced training budget (Sener & Savarese, 2018; Paul et al., 2021; Zheng et al., 2023). Yet data selection also plays a more proactive role: *data shapes model behavior*. Machine learning models continuously pick up bias from training data, which can include representative traits for recognizing a bird species but can also include observation-induced background elements, such as bird feeders, creating potential spurious correlations (Torralba & Efros, 2011; Dharmasiri et al., 2025).

Such a close relationship between the prevalence of visual evidence and downstream behavior motivates us to rethink data selection through the lens of *concepts*. Each image carries various concepts, including objects, attributes, textures, and backgrounds (Choi et al., 2025), and their support is inherently uneven across a dataset (Zhu et al., 2014; Liu et al., 2019). As shown in Figure 1, generic concepts, such as bird and branch, are broadly supported by many samples; toward the long tail, we observe appearance and background concepts associated with fewer categories or incidental events. Consequently, selecting subsets implicitly reallocates training support across all of the concepts the samples carry. More broadly, we argue that data selection *need not preserve the source distribution*; it should allocate training support to evidence most useful for downstream learning. While this intuition has been recognized in targeted data selection (Xia et al., 2024; Engstrom et al., 2024; Xie et al., 2023), long-tailed selection (Vo et al., 2024; Xu et al., 2024), and other specialized settings, we still lack a principled way to determine the concept allocation across different learning scenarios.

We study this question using concepts derived by Sparse autoencoders (SAEs), which decompose pretrained features into sparse coordinates associated with recurring visual patterns (Makhzani & Frey,

[1]We attach source code, selected subsets, and models at the anonymized repository.

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

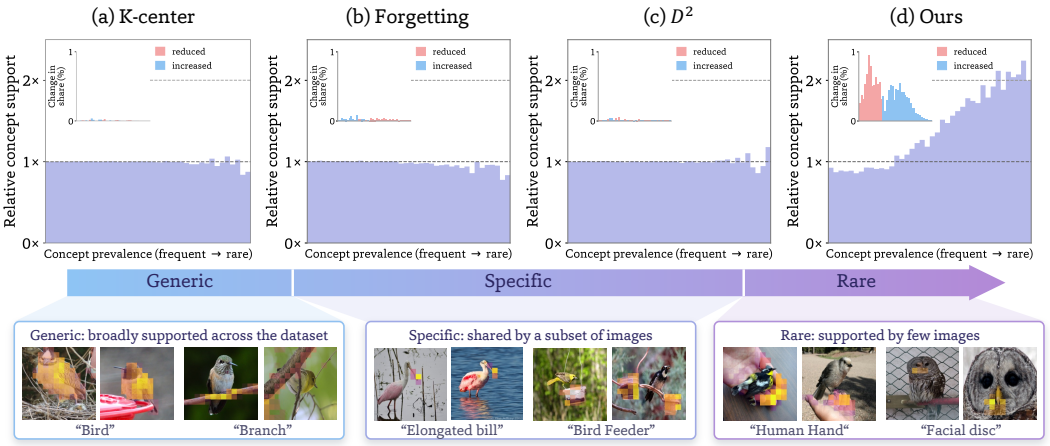


Figure 1: **Concept-support distribution produced by different selectors.** Data selection reallocates the concept distribution from the original dataset. The top bar compares three previous methods and our approach on relative change of concept support in a 10% subset. Our method substantially amplifies the support for tail concepts (insets show the absolute change, where red indicates suppression and blue amplification). In contrast, K-Center, Forgetting, and D^2 Pruning largely preserve the original concept distribution. Bottom: representative SAE concepts across the same prevalence spectrum, from broadly supported **generic** concepts, through **specific** patterns shared by a subset of images, to **rare** concepts supported by only a few images. Highlighted patches indicate high-activation regions.

2013; Gao et al., 2025; Choi et al., 2025; Stevens et al., 2025). The sparse representation characterizes a selected subset by its induced concept support. Unlike previous data-selection literature based on SAEs (Yang et al., 2025; Zhou et al., 2026; Das et al., 2026), we seek a direct relationship between the allocation of training support across concepts and the resulting downstream model behavior. More specifically, different source distributions, including fine-grained, spurious-correlation, and general datasets, can have distinct demands on the concept tail: some settings may favor repeatedly supported discriminative evidence, while others require greater support for less frequent concepts.

To this end, we formulate a concept-based data-selection objective to make concept allocation explicit. Specifically, we accumulate a concept distribution as samples are selected. Imposing diminishing returns on this distribution gradually shifts selection pressure from saturated concepts toward the under-supported tail. After identifying valuable concepts that fill the distribution gap, we use local support to ensure that the selected samples, as well as their neighbors, contribute to those concepts. Therefore, the selection process favors recurring underrepresented concepts over isolated occurrences. Together, these components allow us to vary the balance between local support and tail concepts and study the downstream behaviors induced by different allocation strategies.

As shown in Figure 1, our method significantly alters the concept distribution in the selected subset, particularly amplifying tail concepts. Empirically, our selected subsets consistently outperform strong data-selection baselines across classification and multimodal pretraining tasks. More importantly, varying concept allocation reveals systematic changes in downstream behavior. Concept support enhanced by the proposed strategy improves test accuracy on data with those concepts. Even for concepts not directly tied to specific categories, increasing exposure during training reduces the risk of test samples falling out of distribution. Broader concept coverage can improve robustness to spurious context, while aggressively prioritizing tail concepts can hurt performance when recurring discriminative evidence matters more. These findings establish concept allocation, beyond individual sample importance, as a useful interface for studying and designing data selection.

2 Related Work

Dataset selection. Data selection, or data pruning, seeks to retain a compact training subset while preserving downstream performance (Guo et al., 2022). Existing methods characterize useful data from several perspectives. Geometry-based approaches emphasize representativeness and diversity

through clustering, K-Center, or submodular coverage (Sener & Savarese, 2018; Wei et al., 2015), while training-dynamics and optimization-based methods use forgetting events, early-training scores, gradients, or validation objectives (Toneva et al., 2019; Paul et al., 2021; Mirzasoleiman et al., 2020; Killamsetty et al., 2021a;b). Self-supervised scores further enable pruning without supervised scoring models (Sorscher et al., 2022). Recent methods increasingly combine sample utility with coverage and redundancy, including Moderate Coreset, CCS, D^2-Pruning, and InfoMax (Xia et al., 2023; Zheng et al., 2023; Maharana et al., 2024; Tan et al., 2025), while data-attribution methods estimate the contribution of individual training examples (Koh & Liang, 2017; Ghorbani & Zou, 2019). Pretrained representations have further enabled fine-grained and multimodal data selection, including RAM-APL and large-scale image–text filtering (Wan et al., 2025; Gadre et al., 2023; Fang et al., 2024; Mahmoud et al., 2024; Nicolicioiu et al., 2026). Concept bottlenecks have also been used to estimate sample difficulty for coreset selection (Mehra et al., 2025). Our method performs data selection through sparse concepts across fine-grained, spurious-correlation, and general settings. Importantly, we examine the relationship between concept distributions produced by different selection strategies and downstream model behavior across dataset types.

Concept representations and sparse autoencoders. Concept-based interpretability aims to describe model behavior using human-understandable factors, as explored by methods such as Network Dissection, TCAV, ACE, and concept bottleneck models (Bau et al., 2017; Kim et al., 2018; Ghorbani et al., 2019; Koh et al., 2020). Sparse autoencoders (SAEs) provide a complementary approach by decomposing dense activations into sparse, often interpretable features (Huben et al., 2024; Bricken et al., 2023; Gao et al., 2025). Recent work extends these representations to vision and vision–language models, where SAE features capture localized and hierarchical visual patterns (Stevens et al., 2025; Pach et al., 2025; Bussmann et al., 2025; Zaigrajew et al., 2025). Concept representations have also begun to support data selection, including SDM for target-distribution matching, alongside work on concept-based difficulty estimation, diversity-driven selection, and dynamic concept exposure (Das et al., 2026; Mehra et al., 2025; Yang et al., 2025; Zhou et al., 2026). Our work directly optimizes concept support together with local support, yielding an interpretable selection criterion for classification and multimodal pretraining.

3 METHODOLOGY

Let $\mathcal{D} = \{I_i\}_{i=1}^n$ be a set of training images and $B \ll n$ a budget. Data selection constructs a subset $\mathcal{S} \subseteq \mathcal{D}$ with $|\mathcal{S}| = B$ that preserves the information in the original dataset needed for model training. Each image consists of various concepts, such as objects, attributes, colors, and textures. Therefore, we view the data selection process as reallocating support for these concepts under a limited budget. Specifically, we apply an SAE with C dimensions to the DINOv3 (Siméoni et al., 2025) patch features of each image and max-pool the resulting patch activations, so that each image $\boldsymbol{I}_i$ is represented by a sparse vector $\boldsymbol{x}_i \in \mathbb{R}^C$, where $\boldsymbol{x}_i[c]$ is the corresponding activation for concept c, *i.e.*, the c-th SAE dimension.

SAE-extracted concepts usually demonstrate a long-tailed distribution. A small number of concepts receive support from many images, while other concepts occur in only a limited portion of the dataset. These tail concepts can capture fine-grained attributes and uncommon objects or context variations. Although the latter may not be directly tied to a category, removing the few samples that contain them can narrow the training distribution's support, making similar test cases more likely to appear out of distribution. However, their limited support also makes these concepts particularly vulnerable to data reduction, motivating us to explicitly track the accumulated concept support in the selected subset.

At the same time, the attention to tail concepts does not imply blindly equalizing the concept distribution. Different source distributions place varied demand on concept allocation under a limited budget. Expanding concept coverage too aggressively toward the tail may introduce chaotic training signals from isolated samples. We therefore formulate data selection as a progression from *concept value* to *sample value*: it should identify concepts that are insufficiently represented as selection proceeds, while also determining which samples provide reliable support for those concepts in the source distribution. Our method follows this progression, measuring **concept coverage** to suggest *what evidence* is needed and **local support** to decide *which samples* meaningfully fill that gap. Figure 2 illustrates the proposed selection pipeline.

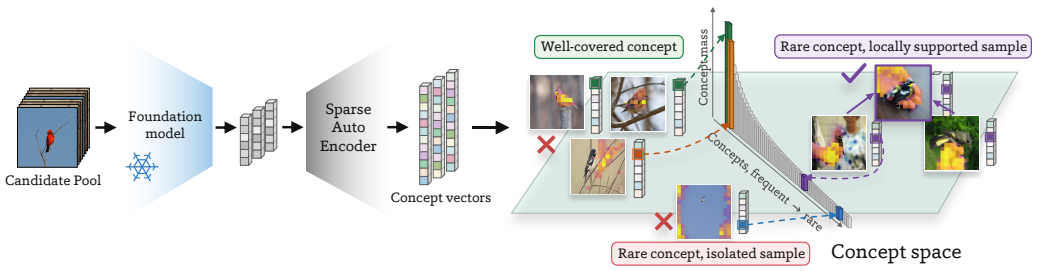


Figure 2: **Overview of concept-based subset selection.** We use SAEs to decompose foundation-model patch features into sparse concept vectors and track the accumulated support of each concept in the selected subset. Concept-level diminishing returns reduce the marginal value of already well-covered evidence, steering the selection toward tail concepts. We then use local support to locate samples that contribute to filling the distributional gap together with their neighbors. The resulting objective favors candidates representing under-supported evidence with local support over redundant head concepts or isolated rare regions.

3.1 Concept Coverage as a Probe for the Distribution Tail

We first apply diminishing returns to accumulated concept support and score a subset by

$$F(\mathcal{S}) = \sum_{c=1}^{C} \log\left(1 + \sum_{\boldsymbol{I}_i \in \mathcal{S}} \boldsymbol{x}_i[c]\right). \tag{1}$$

The logarithm reduces the marginal gain of repeatedly adding evidence for concepts that are already well represented. Consequently, under-supported concepts become more competitive as selection proceeds, without an explicit definition of the tail or a target concept frequency. §D compares alternative concave utilities.

3.2 Local Support to Expose Valuable Samples

Concept coverage determines *where* additional support is needed, but does not by itself indicate *whether* a sample carrying an under-supported concept reflects meaningful structure in the source distribution. Therefore, we augment each candidate with support from its local neighborhood. For each image $\boldsymbol{I}_t$, let $\mathcal{N}_k(t)$ denote its k nearest neighbors in SAE space. We accordingly update the concept contribution in Eq. 1 for candidate $\boldsymbol{I}_i$ to

$$\tilde{\boldsymbol{x}}_i[c] = \boldsymbol{x}_i[c] + \beta \sum_{t:\, \boldsymbol{I}_i \in \mathcal{N}_k(t)} a_{ti} \boldsymbol{x}_t[c], \tag{2}$$

where β controls the strength of local support and a_{ti} is the cosine similarity between $\boldsymbol{I}_t$ and $\boldsymbol{I}_i$, normalized across the neighboring set $\mathcal{N}_k(t)$. The first term captures concepts directly present in I_i, and the second term assigns complementary credit recurring among nearby samples. As a result, repeatedly supported concepts gradually receive diminishing marginal gains. Yet rare concepts do not automatically dominate selection because the corresponding samples may receive limited local support when isolated. We report ablations of β and the row normalization of a_{ti} in §4.2, with additional neighborhood-size and objective-design studies in §D.

Optimization. The objective also admits efficient optimization. Since $\log(1+\cdot)$ is monotone and concave and $\tilde{x}_i \geq 0$, F is monotone submodular, greedy selection attains a $(1-1/e)$ approximation to the optimal subset (Nemhauser et al., 1978). We run lazy greedy (Minoux, 1978) from $\mathcal{S} = \emptyset$, avoiding recomputing the marginal gain of every candidate at each iteration. Because both the original SAE activations and their effective concept contributions are sparse, each marginal-gain computation only involves the nonzero dimensions of $\tilde{\boldsymbol{x}}_i$. The resulting procedure preserves the structure of the concept-based objective while remaining practical for large-scale subset selection.

4 Experiments

Tasks, models, and datasets. We evaluate our method on image classification and multimodal pre-training. For image classification, we use NABirds (Van Horn et al., 2015) and Stanford Cars (Krause

Table 1: **Results on image classification datasets.** Comparison of subset selection performance with existing coreset selection methods using ResNet-50. We report classification accuracy for NABirds and Stanford Cars, and worst-group accuracy (WGA) for Waterbirds. All results are averaged over three runs. **Bold** and underlined entries indicate the best and second-best results, respectively.

Methods	NABirds						Stanford Cars						Waterbirds						Spawrious					
Selection Rate	10%	20%	30%	50%	70%	100%	10%	20%	30%	50%	70%	100%	10%	20%	30%	50%	70%	100%	10%	20%	30%	50%	70%	100%
Random	35.6	55.3	64.9	73.3	77.1	80.3	21.9	45.2	63.1	78.0	83.7	87.5	52.3	66.2	67.5	68.8	79.1	80.1	63.8	67.4	72.6	68.8	69.5	76.3
K-Center	31.6	54.4	64.9	74.1	77.8	-	16.6	39.1	56.3	77.5	83.9	-	57.5	61.3	64.2	**79.4**	69.1	-	71.3	71.7	71.9	69.9	69.1	-
Forgetting	22.4	49.8	63.3	74.3	78.1	-	11.8	37.0	56.7	72.2	81.1	-	31.6	43.6	60.1	73.4	77.0	-	61.1	68.5	67.9	74.3	74.3	-
EL2N	16.8	44.4	61.2	74.4	**78.5**	-	8.9	28.4	49.2	71.0	79.3	-	31.0	52.3	64.2	72.9	76.5	-	**74.6**	**74.3**	75.6	61.0	68.7	-
Moderate	34.6	56.0	64.9	73.7	77.5	-	19.7	45.9	59.8	77.3	83.8	-	60.5	68.3	68.0	71.1	71.9	-	70.4	72.0	72.7	68.4	67.3	-
CCS	39.7	57.7	65.5	73.6	77.6	-	23.6	49.3	62.7	77.8	**84.6**	-	52.8	62.0	67.1	69.9	70.7	-	72.9	73.4	**76.7**	**75.3**	**76.7**	-
D^2-Pruning	25.8	54.8	66.6	74.7	77.9	-	22.6	50.9	63.9	72.4	74.1	-	47.0	49.2	56.4	60.8	66.2	-	54.6	61.0	60.4	64.3	61.3	-
InfoMax	22.5	52.0	64.7	72.7	76.3	-	11.1	37.3	57.3	75.5	81.3	-	22.6	54.0	61.7	70.9	72.7	-	66.7	62.1	61.2	62.3	64.0	-
RAM-APL	39.3	54.7	61.7	69.8	74.8	-	**26.0**	50.9	61.5	73.9	81.0	-	30.2	38.8	51.2	65.0	68.2	-	45.5	46.0	47.2	51.1	53.3	-
SDM	37.5	56.3	64.5	73.7	78.2	-	20.6	45.4	62.5	76.8	82.4	-	61.0	64.3	70.4	75.2	74.4	-	71.4	71.8	70.9	72.0	74.0	-
Ours	**42.9**	**59.4**	**67.4**	**75.2**	78.3	-	23.9	**53.4**	**65.3**	**78.8**	84.0	-	**68.6**	**72.7**	**77.9**	79.3	**79.8**	-	74.3	73.5	74.7	72.3	73.8	-

Table 2: **Results on multimodal pretraining.** Comparison of subset selection performance on CC12M using CLIP ViT-B/16. We report zero-shot and linear-probe classification accuracy on ImageNet-1K, and Recall@1 for image-to-text (I2T) and text-to-image (T2I) retrieval on Flickr30K. **Bold** and underlined entries indicate the best and second-best results, respectively.

Methods	Zero-Shot ImageNet-1K						Linear Probe ImageNet-1K						I2T Flickr30K						T2I Flickr30K					
Selection Rate	10%	20%	30%	50%	70%	100%	10%	20%	30%	50%	70%	100%	10%	20%	30%	50%	70%	100%	10%	20%	30%	50%	70%	100%
Random	3.6	7.7	12.9	20.9	26.4	32.4	21.6	30.1	39.1	48.1	54.2	58.7	3.8	8.5	17.2	31.1	42.5	51.1	**3.1**	6.8	11.4	22.4	30.5	38.1
CLIP Score	3.2	9.9	15.2	23.9	28.6	-	23.0	36.1	40.6	49.4	55.0	-	**3.9**	9.3	14.7	28.7	42.0	-	2.7	7.0	10.6	20.1	31.0	-
K-center	1.7	6.8	10.7	21.5	29.3	-	18.5	31.8	38.5	49.6	55.8	-	1.8	12.6	21.3	**38.4**	46.4	-	1.6	8.3	13.9	26.7	33.8	-
Moderate	3.8	10.1	14.6	22.5	29.0	-	22.3	34.9	41.7	50.1	55.8	-	3.1	11.0	19.1	32.8	43.3	-	2.4	8.3	13.0	22.9	32.4	-
CCS	4.0	10.3	15.4	22.5	25.2	-	23.6	36.3	42.5	50.0	53.5	-	3.7	13.5	22.3	35.4	41.7	-	2.8	10.0	15.4	24.7	29.3	-
D^2-Pruning	2.5	8.8	14.4	24.3	29.6	-	22.3	36.0	42.7	51.7	56.6	-	2.5	9.6	20.0	33.5	44.9	-	2.0	7.6	13.1	23.2	31.4	-
Infomax	2.3	8.9	14.2	24.1	29.2	-	22.1	36.1	41.4	51.2	55.9	-	2.3	9.4	19.3	34.1	43.8	-	2.1	7.3	12.7	23.5	30.8	-
Ours	**6.4**	**12.4**	**16.6**	**25.9**	**30.9**	-	**24.7**	**37.0**	**43.9**	**52.8**	**57.3**	-	3.7	**15.4**	**24.7**	37.2	**48.7**	-	2.6	**10.2**	**16.5**	**27.6**	**35.8**	-

et al., 2013) for fine-grained recognition, Waterbirds (Sagawa et al., 2020) and Spawrious (Lynch et al., 2023) for robustness to spurious correlations. We train ResNet-50 (He et al., 2016) on the selected subsets. For multimodal pretraining, we select subsets from CC12M (Changpinyo et al., 2021) and train CLIP ViT-B/16 (Radford et al., 2021). We evaluate ImageNet-1K zero-shot and linear-probe classification, together with image-to-text and text-to-image retrieval on Flickr30K (Plummer et al., 2015). All methods use the same selection budget and downstream training configuration.

Baselines. In the classification task, we compare with **1) Random**, uniform subset sampling; **2) K-Center** (Sener & Savarese, 2018), geometry-based feature-space coverage; **3) Forgetting** (Toneva et al., 2019), based on forgetting events during training; **4) EL2N** (Paul et al., 2021), based on prediction-error norms; **5) Moderate** (Xia et al., 2023), which selects samples of moderate difficulty; **6) CCS** (Zheng et al., 2023), coverage-aware sampling over difficulty scores; **7) D^2-Pruning** (Maharana et al., 2024), which combines difficulty and diversity; **8) InfoMax** (Tan et al., 2025), which jointly considers informativeness and redundancy; and **9) RAM-APL** (Wan et al., 2025), foundation-model-based selection for fine-grained recognition. **10) SDM** (Das et al., 2026), submodular distribution matching over SAE concepts. We follow the original supervision and implementation protocols of each method. For multimodal pretraining, we use the CLIP score (Hessel et al., 2021) as a strong baseline for CC12M. Based on the CLIP scores, we instantiate **Moderate**, **CCS**, **D^2-Pruning**, and **InfoMax**. We additionally compare with **Random** and **K-Center** as unsupervised baselines.

4.1 Main Results

Image classification. We evaluate subset selection across fine-grained recognition and spurious-correlation benchmarks in Table 1. Our method shows its largest advantages in the low-data regime. On NABirds, it improves over the strongest baseline by 3.2% at the 10% selection rate, with consistent gains as the budget increases. The advantage is even more pronounced on Waterbirds, reaching 7.6% and 7.5% at the 10% and 30% rates, respectively, indicating substantially better robustness. Our method also achieves the best performance at multiple budgets on Stanford Cars while remaining competitive on Spawrious. EL2N and CCS yield better performance on Spawrious, but their performance on other benchmarks is generally weak.

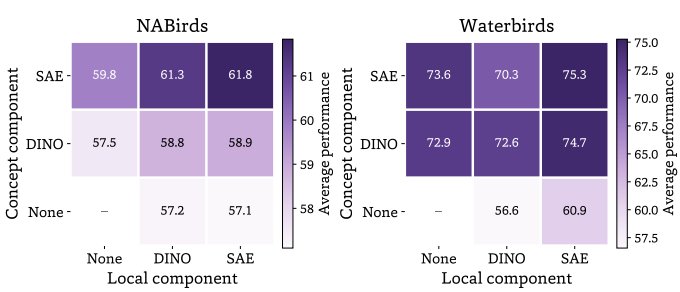


Figure 3: **Ablation of representations used for concept coverage and local support.** Rows indicate the representation used for the concept component, and columns indicate the representation used for the local component; *None* removes the corresponding component. We report NABirds classification accuracy and Waterbirds worst-group accuracy (WGA), averaged across the 10%, 30%, and 50% selection rates.

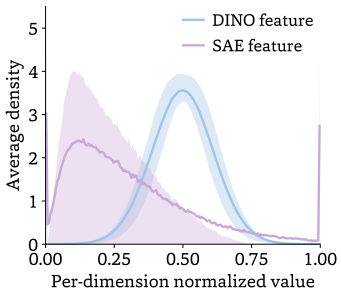


Figure 4: **Activation distributions of DINO and SAE features.** We normalize each DINO feature dimension and SAE feature dimension independently, and report the average activation density across dimensions.

Table 3: **(a) Robustness to different pretrained features** used for SAE training and subset selection. **(b) Effect of normalization** over neighbor similarity. We report accuracy for NABirds and worst-group accuracy (WGA) for Waterbirds averaged across selection ratios. Best results are **bolded**.

(a)

Dataset	Pretrained Architecture				
	DINOv2-B	DINOv3-S	DINOv3-B	CLIP ViT-B/16	SigLIP 2
NABirds Acc	**62.0**	**62.0**	61.8	61.6	60.8
Waterbirds WGA	75.1	**76.8**	75.3	76.0	75.7

(b)

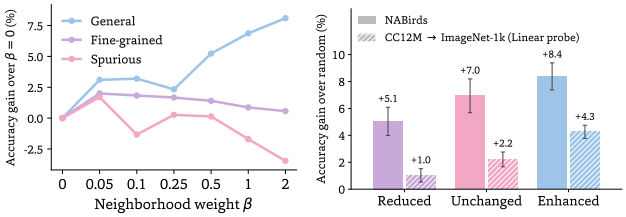
Dataset	Local Support	
	Unnormalized	Normalized
NABirds Acc	61.5	**61.8**
Waterbirds WGA	74.6	**75.3**

Multimodal pretraining. We further evaluate whether the same selection principle transfers to large-scale multimodal pretraining by selecting CC12M subsets and training CLIP ViT-B/16 (Table 2). Our method selects data samples based on the image-level concept support. Even *without information from the text distribution*, our method consistently achieves the best ImageNet-1K zero-shot and linear-probe accuracy across selection rates, with a 2.4% improvement in zero-shot accuracy at the 10% budget. The benefit also extends beyond classification to cross-modal retrieval: at 30%, I2T and T2I Recall@1 improve by 2.4% and 1.1% over the strongest baselines, and the corresponding margins are 2.3% and 2.0% at 70%. These results indicate that the concept-support principle transfers beyond image classification and improves image–text alignment at multimodal pretraining scale.

4.2 Ablation Study

Concept and local representations. Our proposed objective combines concept coverage and local support. Both metrics can be calculated from dense embeddings or sparse concept activation vectors. We ablate this representation usage in Figure 3. On NABirds, using SAE features for concept coverage consistently provides the strongest results. On Waterbirds, we see a stronger performance improvement when moving the calculation of local support from dense DINO embeddings to the SAE space. We further investigate why the sparse concepts identified by SAEs facilitate better selection despite using the same objective by comparing their dimensional distribution across the dataset. As shown in Figure 4, DINO features generally illustrate normal distributions across samples. In contrast, SAEs decompose dense features into monosemantic units and yield multiple peaks per dimension. Such distributions create clear separation between samples with different characteristics. Thus, the data selection objective receives more indicative signals for the accumulated and rare concepts.

Robustness to pretrained visual features. Our method is not tied to a particular pretrained representation. In Table 3a, we construct the concept representation from different pretrained visual encoders, including DINOv2-B (Oquab et al., 2024), DINOv3-S and DINOv3-B, CLIP-B, and SigLIP 2-B (Tschannen et al., 2025). The method remains consistently effective across these substantially different feature spaces, demonstrating that its performance does not depend on a specific backbone or pretraining objective. This robustness aligns with our formulation: the selection objective examines the accumulated distribution to locate tail concepts. Therefore, any encoder that extracts



(a) (b)

Figure 5: **(a) Sensitivity to the neighborhood weight** β, averaged over the 10%, 30%, and 50% selection rates, for general (Tiny-ImageNet), fine-grained (NABirds), and spurious-correlation (Waterbirds, WGA) settings. **(b) Concept support and downstream gain.** Accuracy gain over random selection on images containing concepts with reduced, unchanged, or enhanced training support.

Table 4: **Concept-support profiles of NABirds subsets.** Unique Coverage: the number of concepts receiving nonzero support. Effective Number: exponential of the entropy of the normalized concept-mass distribution. Prevalence Ratio: mass-weighted mean concept prevalence normalized by the full set; values below one indicate greater relative support for tail concepts.

Methods	Unique Coverage	Effective Number	Prevalence Ratio
Random	7,973	865	1.00
K-Center	8,215	893	0.97
Forgetting	8,120	888	0.97
D^2-Pruning	8,099	887	0.97
Ours	**8,951**	**1,124**	**0.74**

diverse concepts can provide sufficient information for our approach. Thus, the concept-based subset selection can serve as a general selection principle on top of different pretrained visual features.

Effect of similarity normalization. We normalize the similarity between each sample and its neighbors to produce the local support component in Eq. 2. We examine its necessity in Table 3b. Specifically, we compare directly using raw cosine similarities with normalized ones. Normalization yields slightly better performance across selection rates. It gives each target sample a bounded contribution budget that is distributed among its neighbors according to their relative similarities, preventing samples in dense neighborhoods from contributing disproportionately large local support.

Sensitivity to neighborhood weight. Finally, we study the neighborhood weight β, which controls the relative contribution of local support to the selection objective, in Figure 5a. Setting $\beta = 0$ removes the local contribution and simplifies the objective to concept coverage alone. Particularly, we test its influence on different types of original data.

Across fine-grained, spurious, and general cases, having $\beta = 0.05$ consistently improves the performance over concept coverage alone. However, their preferred weights differ. Spurious correlation and fine-grained datasets yield the best training performance at $\beta = 0.05$, whereas general-purpose datasets favor a substantially stronger local contribution with $\beta = 2$. This suggests that when the task scale is relatively small and concentrated, broader concept coverage reduces the risk of test-time distribution shift. In contrast, general classification tasks, which usually have larger concept breadth, benefit from emphasizing local support. We therefore use the corresponding optimal value for each dataset group throughout our experiments. See §D for additional ablations and per-rate results.

5 Analysis and Discussion

The concept allocation produced by data selection creates a direct connection to downstream model behavior. We examine the concept distribution under different selection strategies and how model behavior changes accordingly, asking whether the evidence made more available during training is also where downstream recognition improves.

Different samples can produce similar concept support. Selecting images with different criteria need not substantially change the concept distribution in the resulting subset. We summarize the concept coverage produced by K-Center, Forgetting, D^2-Pruning, and our approach in Table 4. **Unique Coverage** counts concepts receiving nonzero support; **Effective Number** measures how broadly normalized activation mass is distributed; and **Prevalence Ratio** measures its mass-weighted mean concept prevalence relative to the full dataset. A prevalence ratio below one indicates greater relative support for tail concepts. Despite distinct selection criteria, their allocations of activation mass across concepts are similar, as confirmed in Figure 1. Our subset supports 8,951 concepts, compared with 7,973 under random selection, and increases the effective number from 865 to 1,124. Its prevalence ratio falls from 1.00 to 0.74, while the three compared selectors remain at 0.97. The enrichment profile in Figure 1 shows our subset allocates a greater share of its concept support to

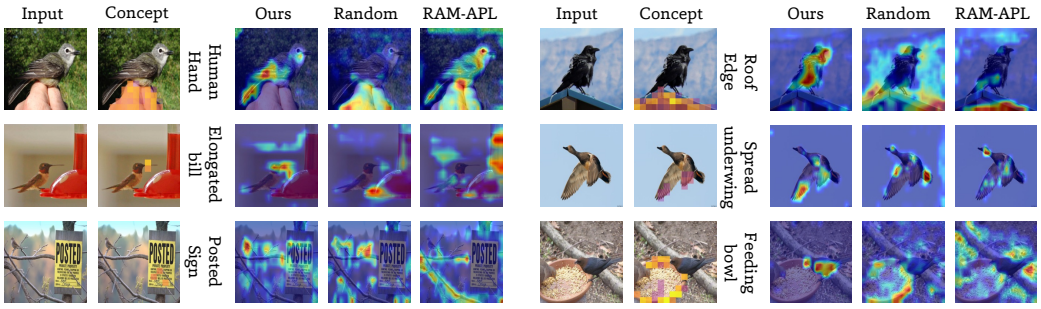


Figure 6: **Qualitative examples of enhanced visual concepts.** Our selection enhances concepts of both fine-grained attributes and contextual evidence. Finer-CAM visualizations compare the regions that models trained on our, random, and RAM-APL subsets rely on for prediction, showing that broader concept exposure during training improves test-time robustness.

tail concepts. The change proves that our approach effectively reallocates the concept distribution compared with the original dataset and other baselines.

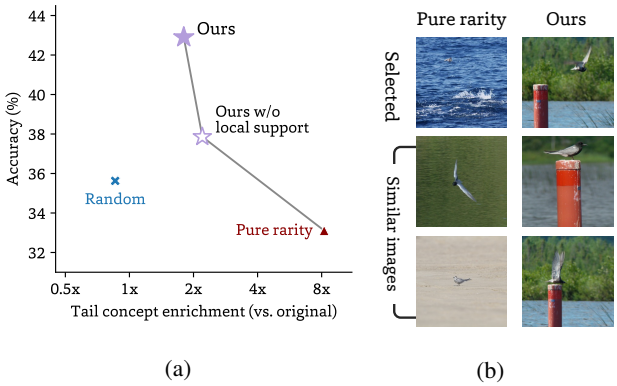
From concept support to better recognition. The change in concept support motivates us to investigate their downstream correspondence. We start with test accuracy and use NABirds and ImageNet-1K linear probing from multimodal pretraining as test tasks. Based on the change in its training support compared with random selection, we label each concept as **enhanced**, **unchanged**, or **reduced**. For each concept, we select the test samples with the highest activation and calculate their average accuracy. Then, we measure how much our model improves over the random-selection model on each concept category. As shown in Figure 5b, while all three categories improve, samples carrying enhanced concepts show substantially larger improvements than the other two categories. The observation proves that concepts enhanced by our approach, while remaining in the distributional tail, actively shape model behavior. It also raises another question: *do these enhanced concepts always relate to specific categories*? We answer this question using two case studies.

5.1 Case Study: NABirds

We first look into fine-grained classification and show example concepts in NABirds with enhanced support produced by our selection approach in Figure 6. Some concepts are activated on discriminative features, such as beaks and wings. At the same time, other concepts that do not directly relate to categorical appearance, such as hands, signs, and feeders, are also enhanced. Test accuracy on samples carrying these concepts also improves.

Beyond category-related concepts. Using the concept *hand* as an example, we examine its support over the dataset. Among 23,929 training images, only 160 activate the dimension, spread across 108 categories. Random selection retains 20 such images across 20 categories, whereas our subset retains 60 across 50 categories, substantially broadening the support to this uncommon context. On strongly hand-activated test images, accuracy rises from 28.6% to 51.4%. We visualize the explanation heatmap using Finer-CAM (Zhang et al., 2025) to understand this behavior, which suggests that the model trained on our selected data remains concentrated on the bird, while data selected by random or RAM-APL biases the model to these contextual concepts. These analyses show that while tail contextual concepts are not category-related discriminative features, they increase the robustness of model training, so these concepts remain in distribution at test time. Broader category coverage also reduces the risk of *hand* spuriously correlating with a small number of categories.

Necessity of local support. While some tail concepts improve training robustness, not all do. In Figure 7a, we show that removing the local support term only slightly increases the amplification on tail concepts, but reduces accuracy from 42.9% to 37.9%. If we rank samples only by their concept rarity, which is the inverse prevalence of their active concepts, the tail concepts are substantially enriched by $8.2\times$, but the resulting model only yields 33.1% accuracy. Aggressively prioritizing tail concepts can discard useful representative structure. Figure 7b partially explains that pure-rarity selection often favors atypical examples with little local support, whereas our objective favors candidates that also represent under-supported evidence within their local neighborhoods.



(a) (b)

Figure 7: **Local support identifies useful tail examples without overemphasizing rarity. (a)** Further increasing tail enrichment beyond the full objective does not improve NABirds accuracy, indicating that rarity alone is insufficient. **(b)** Pure-rarity selection can retain isolated examples, whereas our local-support objective favors underrepresented examples that remain supported by nearby samples.

Table 5: **Subset composition and robustness to background changes on Waterbirds.** Counter-corr. is the fraction of bias-conflicting examples. All-bg correct requires correct prediction under all four background replacements. Group-matched controls use the same label–background counts as Ours. All values are percentages.

Method	Counter-corr.	WGA ↑	All-bg ↑
Random	4.0	52.3	57.6
Forgetting	5.0	31.6	32.4
K-Center	5.4	57.5	72.6
Moderate	9.3	60.5	78.3
Ours	18.8	**68.6**	**86.5**
Group-matched controls			
Random	18.8	63.7	84.2
Difficult	18.8	52.2	73.7
Easy	18.8	64.0	85.8

5.2 Case Study: Waterbirds

After locating category-irrelevant concepts that contribute to model robustness, we run another test to expose how our approach alters support for spurious correlations. Waterbirds provides a controlled setting in which bird identity is spuriously correlated with background.

Selection and background reliance. As shown in Table 5, our approach yields 18.8% bias-conflicting examples (*e.g.*, waterbirds with land background) in the selected set, while the other compared approaches select less than 10%. Although this ratio does not have a strict positive relationship with downstream performance, broader concept coverage weakens the background shortcut by diversifying the object–background combinations.

Group composition examination. Part of this robustness improvement could arise from the larger fraction of bias-conflicting examples alone. We therefore construct random subsets with the same number of samples from four label–background groups. Matching the group composition improves WGA from 52.3% for ordinary random selection to 63.7%, still below our method. Moreover, with the same group quotas, selecting difficult examples yields only 52.2% WGA, while selecting easy examples reaches 64.0%. These results are consistent with prior evidence that group proportions alone do not determine coreset robustness (Dharmasiri et al., 2025). Together, the results suggest that concept-based selection improves robustness through both a more favorable group composition and richer visual evidence within those groups.

6 Conclusion

This paper views data selection as reallocating training support for visual concepts under a limited data budget and studies the relationship between allocation strategy and downstream model behavior. Specifically, we extract visual concepts from each image using sparse autoencoders. Each selected sample contributes its concepts to an accumulated distribution throughout the selection process. We design an objective that balances concept coverage and local support to mitigate the gap in tail concepts while avoiding isolated samples. This flexibility also allows us to vary allocation strategies across fine-grained, spurious-correlation, and generation selection scenarios. Experiments on classification and multimodal pretraining tasks suggest the effectiveness of the proposed concept-based selection. Importantly, we use the concepts enhanced by our method to probe model properties trained on the selected dataset. Analyses show that context concepts, while not directly contributing category-related information, improve model robustness when encountering out-of-distribution and spurious connections. More broadly, this paper advocates for concepts, which offer interpretable, fine-grained quantitative measurement for visual semantics, as an important principle in data selection. We need selection approaches that can proactively shape downstream model behavior.

AI USE STATEMENT

In this work, we used generative AI tools to edit portions of the manuscript for clarity and readability, and as coding assistants to help implement and debug parts of the experimental code. All AI-assisted code was reviewed, tested, and verified for correctness by the authors. We did not use generative AI tools to propose or refine hypotheses, design experiments, analyze data, or interpret results; generating synthetic data, developing theoretical models, formulating mathematical claims, and writing proofs are not applicable to this work. All AI-edited text was reviewed by the authors to ensure it does not alter the scientific content. We take responsibility for the final content of this work, including text, claims, code, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our study does not involve human subjects, and no personal or identifiable information was collected. All experiments are conducted on publicly available datasets in accordance with their respective licenses. We emphasize that the released coresets and models are intended for scientific research purposes. Since our method selects subsets from existing datasets, the selected data naturally reflect the characteristics of their source datasets. We therefore recommend validating the selected subsets for the intended application before use beyond research settings.

REPRODUCIBILITY STATEMENT

We release our source code, the selected coresets, and the trained models at an anonymized repository. The experimental setup and training hyperparameters are presented in §A. The details of SAE training and evaluation are included in §B. Efficient implementation at scale is described in §A.3.

REFERENCES

Dimitris Achlioptas. Database-friendly random projections: Johnson-lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66(4):671–687, 2003.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *CVPR*, 2017.

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

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. In *ICML*, 2025.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, pp. 3558–3568, 2021.

Jinho Choi, Hyesu Lim, Steffen Schneider, and Jaegul Choo. Conceptscope: Characterizing dataset bias via disentangled visual concepts. In *NeurIPS*, 2025.

Arnav Mohanty Das, Gantavya Bhatt, Sahil Verma, Yiping Wang, Viswa Virinchi Muppirala, and Jeff Bilmes. Matched data, better models: Target aligned data filtering with sparse autoencoders. In *ICLR*, 2026.

Amaya Dharmasiri, William Yang, Polina Kirichenko, Lydia Liu, and Olga Russakovsky. The impact of coreset selection on spurious correlations and group robustness. In *NeurIPS*, volume 38, 2025.

Logan Engstrom, Axel Feldmann, and Aleksander Madry. DsDm: Model-aware dataset selection with datamodels. In *ICML*, 2024.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander T Toshev, and Vaishaal Shankar. Data filtering networks. In *ICLR*, 2024.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander J Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. Datacomp: In search of the next generation of multimodal datasets. In *NeurIPS*, pp. 27092–27112, 2023.

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 *ICLR*, 2025.

Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *ICML*, volume 97, pp. 2242–2251, 09–15 Jun 2019.

Amirata Ghorbani, James Wexler, James Zou, and Been Kim. Towards automatic concept-based explanations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *NeurIPS*, 2019.

Chengcheng Guo, Bo Zhao, and Yanbing Bai. DeepCore: A comprehensive library for coreset selection in deep learning. In *DEXA*, pp. 181–195, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

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. In *ICCV*, pp. 8340–8349, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, pp. 15262–15271, 2021b.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *EMNLP*, November 2021.

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *ICLR*, 2024.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *ICML*, volume 235, pp. 20617–20642, 2024.

Hervé Jégou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *IEEE TPAMI*, 33(1):117–128, 2011.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2021.

Krishnateja Killamsetty, Durga S, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. Grad-Match: Gradient matching based data subset selection for efficient deep model training. In *ICML*, pp. 5464–5474, 2021a.

Krishnateja Killamsetty, Durga Sivasubramanian, Ganesh Ramakrishnan, and Rishabh Iyer. Glister: Generalization based data subset selection for efficient and robust learning. In *AAAI*, volume 35, pp. 8110–8118, 05 2021b.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viégas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *ICML*, 2018.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *ICML*, volume 70, pp. 1885–1894, 06–11 Aug 2017.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *ICML*, 2020.

Jonathan Krause, Jia Deng, Michael Stark, and Li Fei-Fei. Collecting a large-scale dataset of fine-grained cars. In *Second Workshop on Fine-Grained Visual Categorization*, 28 Jun 2013.

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Large-scale long-tailed recognition in an open world. In *CVPR*, 2019.

Hantao Lou, Changye Li, Jiaming Ji, and Yaodong Yang. SAE-V: Interpreting multimodal models for enhanced alignment. In *ICML*, pp. 40344–40360, 2025.

Aengus Lynch, Gbètondji J.-S. Dovonon, Jean Kaddour, and Ricardo Silva. Spawrious: A benchmark for fine control of spurious correlation biases. *arXiv preprint arXiv:2303.05470*, 2023.

Da Ma, Gonghu Shang, Zhi Chen, Libo Qin, Yijie Luo, Hongshen Xu, Lei Pan, Shuai Fan, Kai Yu, and Lu Chen. Task-specific data selection for instruction tuning via monosemantic neuronal activations. In *NeurIPS*, 2025.

Adyasha Maharana, Prateek Yadav, and Mohit Bansal. $\mathbb{D}^2$ pruning: Message passing for balancing diversity & difficulty in data pruning. In *ICLR*, 2024.

Anas Mahmoud, Mostafa Elhoushi, Amro Abbas, Yu Yang, Newsha Ardalani, Hugh Leather, and Ari S. Morcos. Sieve: Multimodal dataset pruning using image captioning models. In *CVPR*, pp. 22423–22432, 2024.

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

Akshay Mehra, Trisha Mittal, Subhadra Gopalakrishnan, and Joshua Kimball. Coreset selection via llm-based concept bottlenecks. *arXiv preprint arXiv:2502.16733*, 2025.

Michel Minoux. Accelerated greedy algorithms for maximizing submodular set functions. In *Optimization Techniques*, pp. 234–243, 1978.

Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *ICML*, volume 119, pp. 6950–6960, 13–18 Jul 2020.

George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical Programming*, 14(1):265–294, 1978.

Andrei Liviu Nicolicioiu, Sarvjeet Singh Ghotra, Morgane M Moss, and Aaron Courville. Data selection through iterative self-filtering for vision-language settings. *TMLR*, 2026. ISSN 2835-8856.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024.

Mateusz Pach, Shyamgopal Karthik, Quentin Bouniot, Serge Belongie, and Zeynep Akata. Sparse autoencoders learn monosemantic features in vision-language models. In *NeurIPS*, volume 38, pp. 95706–95742, 2025.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. In *NeurIPS*, volume 34, pp. 20596–20607, 2021.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, pp. 2641–2649, 2015.

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 *ICML*, pp. 8748–8763, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5389–5400, 2019.

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*, 115(3):211–252, 2015.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *ICLR*, 2020.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *ICLR*, 2018.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. Beyond neural scaling laws: Beating power law scaling via data pruning. In *NeurIPS*, volume 35, pp. 19523–19536, 2022.

Samuel Stevens, Wei-Lun Chao, Tanya Berger-Wolf, and Yu Su. Interpretable and testable vision features via sparse autoencoders, 2025.

Haoru Tan, Sitong Wu, Wei Huang, Shizhen Zhao, and Xiaojuan Qi. Data pruning by information maximization. In *ICLR*, 2025.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An empirical study of example forgetting during deep neural network learning. In *ICLR*, 2019.

Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR*, pp. 1521–1528. IEEE, 2011.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025.

Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *CVPR*, pp. 595–604, 2015.

Huy V. Vo, Vasil Khalidov, Timothée Darcet, Théo Moutakanni, Nikita Smetanin, Marc Szafraniec, Hugo Touvron, Camille Couprie, Maxime Oquab, Armand Joulin, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Automatic data curation for self-supervised learning: A clustering-based approach. *TMLR*, 2024.

Zhijing Wan, Zhixiang Wang, Zheng Wang, Xin Xu, and Shin'Ichi Satoh. Foundation model insights and a multi-model approach for superior fine-grained one-shot subset selection. In *ICML*, volume 267, pp. 62084–62101, 13–19 Jul 2025.

Kai Wei, Rishabh Iyer, and Jeff Bilmes. Submodularity in data subset selection and active learning. In *ICML*, pp. 1954–1963, 2015.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting influential data for targeted instruction tuning. In *ICML*, 2024.

Xiaobo Xia, Jiale Liu, Jun Yu, Xu Shen, Bo Han, and Tongliang Liu. Moderate coreset: A universal method of data selection for real-world data-efficient deep learning. In *ICLR*, 2023.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. Data selection for language models via importance resampling. In *NeurIPS*, 2023.

Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying CLIP data. In *ICLR*, 2024.

Xianjun Yang, Shaoliang Nie, Lijuan Liu, Suchin Gururangan, Ujjwal Karn, Rui Hou, Madian Khabsa, and Yuning Mao. Diversity-driven data selection for language model tuning through sparse autoencoder. *arXiv preprint arXiv:2502.14050*, 2025.

Vladimir Zaigrajew, Hubert Baniecki, and Przemyslaw Biecek. Interpreting CLIP with hierarchical sparse autoencoders. In *ICML*, 2025.

Chenshuang Zhang, Fei Pan, Junmo Kim, In So Kweon, and Chengzhi Mao. ImageNet-D: Benchmarking neural network robustness on diffusion synthetic object. In *CVPR*, pp. 21752–21762, 2024.

Ziheng Zhang, Jianyang Gu, Arpita Chowdhury, Zheda Mai, David Carlyn, Tanya Berger-Wolf, Yu Su, and Wei-Lun Chao. Finer-cam: Spotting the difference reveals finer details for visual explanation. In *CVPR*, pp. 9611–9620, June 2025.

Haizhong Zheng, Rui Liu, Fan Lai, and Atul Prakash. Coverage-centric coreset selection for high pruning rates. In *ICLR*, 2023.

Yuzhe Zhou, Zhenglin Hua, Haiyun Guo, and Yuheng Jia. Rethinking representativeness and diversity in dynamic data selection, 2026.

Xiangxin Zhu, Dragomir Anguelov, and Deva Ramanan. Capturing long-tail distributions of object subcategories. In *CVPR*, 2014.

Table of Contents in Appendix

A Implementation Details **16**

A.1 Default Configuration 16

A.2 Training Hyperparameters 16

A.3 Scalable Selection: Neighbor Graph and Lazy Greedy 16

B Sparse Autoencoder Training and Evaluation **17**

B.1 Training Setup 17

B.2 Evaluation Metrics and Results 18

C Analysis Protocols **19**

C.1 Concept Support and Concept Groups 19

C.2 Pure-Rarity Selector 19

C.3 Visualization Methods 19

D Additional Ablations **20**

D.1 Concept and Local Representations 20

D.2 Local Support Weight β and Neighborhood Size k 20

D.3 Choice of Utility Function 21

D.4 SAE Features vs. Selection Objective 21

D.5 SAE Pretraining Data 22

D.6 Downstream Architecture 22

E Additional Results **23**

E.1 ImageNet-1K and Distribution Shift 23

E.2 Concept-Support Profiles of All Selectors 23

E.3 Additional SAE Concept Visualizations 23

E.4 Additional Finer-CAM Visualizations 24

F Relation to Concept-Based Data Selection **25**

A Implementation Details

A.1 Default Configuration

We train a TopK SAE (Makhzani & Frey, 2013) on layer-11 patch-token activations extracted from a frozen DINOv3-B encoder (Siméoni et al., 2025). We use 100 million patch tokens sampled from ImageNet-1K (Russakovsky et al., 2015) for training and 10 million patch tokens for validation. The SAE has 16,384 latent dimensions, with $k = 32$ active latents per token. After training, we run SAE inference on each candidate dataset and max-pool the patch-level activations to obtain one concept vector per image, on which data selection is performed. We use the logarithmic utility with $\beta = 0.05$ for fine-grained and spurious-correlation settings, $\beta = 2$ for general setting and $k = 3$ nearest neighbors by default. For classification, ResNet-50 (He et al., 2016) is initialized from ImageNet-pretrained weights, except on Tiny-ImageNet and ImageNet-1K, where it is trained from scratch. For multimodal pretraining, a ViT-B/16 is trained from scratch on the selected CC12M subsets. All selection methods share the same training configuration for each task.

A.2 Training Hyperparameters

We summarize the hyper-parameter configurations in Table 6 and Table 7 for classification, Table 8 for multimodal pretraining, and Table 9 for SAE training. ImageNet-1K uses a larger learning rate, a cosine schedule, and a longer training budget due to its scale.

A.3 Scalable Selection: Neighbor Graph and Lazy Greedy

Neighbor graph. On CC12M, we build the k-nearest-neighbor graph in SAE space with approximate search. Sparse concept vectors are randomly projected to 512 dimensions (Achlioptas, 2003), candidate neighbors are retrieved with a FAISS inverted-file index with product quantization (Jégou et al., 2011; Johnson et al., 2021), and the candidates are re-ranked by exact cosine similarity on the original sparse vectors before keeping the top k. The resulting neighborhood weights are materialized once as a sparse matrix, so that $\tilde{x}_i$ is computed a single time for every image. On a single CPU node with 56 threads, candidate retrieval takes 26 minutes and the exact re-ranking 8 minutes, so the graph for the 12M-image pool is ready in about 34 minutes.

Global lazy greedy. Selection runs as one lazy greedy pass (Minoux, 1978) over the whole pool. Each marginal gain touches only the nonzero entries of $\tilde{x}_i$, and the lazy evaluation refreshes about 50 candidates per selected image on average. The greedy pass takes 47 minutes on the same node, so the full ranking of CC12M costs about 1.3 hours in total and is reused for all ratios.

Table 6: The adopted hyper-parameter setting for image classification on ImageNet-1K.

Setting	Value
Model	ResNet-50
Initialization	None
GPUs	1× H100
Batch size	256
Optimizer	SGD
Learning rate	0.1
Momentum	0.9
Weight decay	1×10^{-4}
Scheduler	Cosine
Training steps	400,000

Table 7: The adopted hyper-parameter setting for image classification on other datasets.

Setting	Value
Model	ResNet-50
Initialization	ImageNet / None:Tiny-In1K
GPUs	1× H100
Batch size	256
Optimizer	SGD
Learning rate	0.001
Momentum	0.9
Weight decay	1×10^{-4}
Scheduler	Constant
Training steps	18,700

Table 8: The adopted hyper-parameter setting for multimodal pretraining on CC12M.

Setting	Value
Model	ViT-B/16
Initialization	None
GPUs	4× H100
Batch size	16,384
Optimizer	AdamW
Learning rate	1×10^{-3}
Weight decay	0.5
Scheduler	Cosine decay
Warmup	670 steps
Training steps	10,050

Table 9: The adopted hyper-parameter setting for SAE training.

Setting	Value
Encoder layer	Layer 11 patch tokens
SAE dictionary size	16,384
GPUs	1× H100
SAE activation	Top-K
Active latents/token	$k = 32$
Token batch size	16,384
Optimizer	Adam
Learning rate	1×10^{-3}
Warmup	500 steps
Training steps	6,104

B Sparse Autoencoder Training and Evaluation

B.1 Training Setup

SAE variants. We train three sparse autoencoder (SAE) variants: ReLU, TopK, and BatchTopK SAEs (Bricken et al., 2023; Makhzani & Frey, 2013; Bussmann et al., 2024). Given an activation $\mathbf{x} \in \mathbb{R}^d$ extracted from a DINOv3-B encoder, each SAE computes

$$\mathbf{z} = \phi(\mathbf{W}_{\text{enc}}\mathbf{x} + \mathbf{b}_{\text{enc}}), \qquad \hat{\mathbf{x}} = \mathbf{W}_{\text{dec}}\mathbf{z} + \mathbf{b}_{\text{dec}}.$$

where ϕ determines the sparsification mechanism.

Matryoshka training. All variants are trained with the Matryoshka reconstruction objective (Bussmann et al., 2025). Given nested latent prefixes $p_1 < \cdots < p_m = d_{\text{SAE}}$, we reconstruct the input from each prefix,

$$\hat{\mathbf{x}}^{(p_i)} = \mathbf{W}_{\text{dec}}[:, 1:p_i]\, \mathbf{z}_{1:p_i} + \mathbf{b}_{\text{dec}}.$$

and optimize

$$\mathcal{L}_{\text{SAE}} = \frac{1}{m}\sum_{i=1}^{m} \left\| \mathbf{x} - \hat{\mathbf{x}}^{(p_i)} \right\|_2^2 + \lambda_{\text{sp}}\mathcal{L}_{\text{sp}} + \lambda_{\text{aux}}\mathcal{L}_{\text{aux}}.$$

The nested objective encourages earlier latent dimensions to capture broadly useful structure while later dimensions refine the reconstruction. We use an ℓ_1 sparsity penalty for the ReLU SAE. For TopK and BatchTopK SAEs, sparsity is enforced directly by the TopK activation, and we additionally use the AuxK loss (Gao et al., 2025) to mitigate dead latents.

Training details. We train SAEs on 100 million patch-token activations from a DINOv3-B encoder. All variants use ten Matryoshka prefixes; TopK and BatchTopK use $k = 32$, and decoder columns are constrained to unit norm. Full training hyperparameters are summarized in Table 9.

Table 10: **Quantitative comparison of SAE variants on ImageNet-1K.** We compare ReLU SAEs with ℓ_1 sparsity coefficients $\lambda \in \{0.1, 0.01\}$, TopK SAEs with $k \in \{16, 32\}$ active latents per token, and BatchTopK SAEs with an average batch-wide activation budget of $k \in \{16, 32\}$ latents per token. Arrows indicate the preferred direction.

Run	L_0	Norm. MSE ↓	n_{dead} ↓	Mean KL ↓	Top-1 Drop ↓	CKNNA ↑	CS ↑
ReLU $\lambda = 0.1$	6.75	0.60	6,331	2.87	0.51	0.28	0.91
ReLU $\lambda = 0.01$	150.69	0.22	2,773	0.24	0.06	0.46	0.97
TopK-16	16.00	0.32	529	0.34	0.09	0.59	0.96
TopK-32	32.00	0.26	472	0.21	0.06	0.69	0.96
BatchTopK-16	16.04	0.32	230	0.35	0.09	0.64	0.96
BatchTopK-32	32.12	0.26	191	0.22	0.06	0.52	0.96

B.2 Evaluation Metrics and Results

Evaluation protocol. We evaluate the learned ReLU, TopK, and BatchTopK SAEs on held-out activation shards using complementary metrics that capture sparsity, reconstruction quality, representation preservation, and downstream fidelity.

L_0. L_0 denotes the mean number of active latent features per token. It characterizes the effective sparsity of the SAE representation and is interpreted jointly with reconstruction fidelity.

Number of dead latents. The number of dead latents, denoted n_{dead}, counts latent features that remain inactive or activate less frequently than a small threshold over the evaluation set (Bricken et al., 2023). Fewer dead latents indicate better utilization of the SAE dictionary.

Normalized MSE. Normalized mean squared error (MSE), also known as the fraction of variance unexplained (Gao et al., 2025), measures reconstruction fidelity by normalizing the squared reconstruction error by the variance of the mean-centered input. A value of zero indicates perfect reconstruction, while a value of one corresponds to predicting the mean activation.

Cosine similarity (CS). Cosine similarity measures the directional agreement between each original activation and its SAE reconstruction. It provides a complementary reconstruction metric to normalized MSE. Higher values indicate better reconstruction fidelity.

Centered kernel nearest-neighbor alignment (CKNNA). CKNNA (Huh et al., 2024) measures local representational alignment between the original backbone activations and the corresponding SAE latent representations. Higher CKNNA indicates better preservation of the local geometry of the original feature space.

Downstream fidelity. We replace the original patch-token activations at the trained layer with their SAE reconstructions, propagate both representations through the remaining frozen backbone layers, and apply the same fixed ImageNet-1K linear probe (Zaigrajew et al., 2025). We report the mean KL divergence between the original and reconstructed prediction distributions together with the decrease in Top-1 accuracy. Lower values indicate better preservation of the original model behavior.

Results. Table 10 summarizes the quantitative comparison across SAE variants. TopK-32 provides a favorable balance across sparsity, reconstruction quality, representation preservation, and downstream fidelity, and is therefore used in our main experiments.

C Analysis Protocols

C.1 Concept Support and Concept Groups

For each concept c, we measure its support in subset S as

$$m_c(S) = \sum_{I_i \in S} x_i[c].$$

We draw $1{,}000$ random subsets using the same budgets and use the resulting values of m_c as a reference distribution. A concept is *enhanced* if its support under our selection exceeds the 97.5th percentile of this distribution, *reduced* if it falls below the 2.5th percentile, and *unchanged* if it lies within the 40th–60th percentiles.

C.2 Pure-Rarity Selector

The pure-rarity selector isolates the effect of favoring rare concepts without any coverage or local-support term. For each concept c, let ρ_c be the fraction of pool images on which it is active. Each image is scored by the summed inverse prevalence of its active concepts,

$$s_i = \sum_{c:\, x_i[c]>0} \frac{1}{\rho_c}.$$

so that images carrying many rare concepts score highest.

C.3 Visualization Methods

SAE concept overlays. Given an image I, we encode each of its DINOv3 patch tokens with the SAE, giving patch-level concept activations $\mathbf{A}_I \in \mathbb{R}_{\geq 0}^{P \times C}$, where $A_{I,p,c}$ denotes the activation of concept c at patch p. To visualize a concept c, we retrieve the images on which it is most strongly active and show where in each image it fires.

Following Stevens et al. (2025), we arrange the activations of concept c on the patch grid and normalize them by the maximum activation of c over the displayed images:

$$\widehat{A}_{I,p,c} = \frac{A_{I,p,c}}{\max_{I',p'} A_{I',p',c}}.$$

Patches below a relative threshold are left untouched. The remaining patches are colored by $\widehat{A}_{I,p,c}$ with a continuous colormap, with opacity increasing with activation, and overlaid on the image. This highlights the regions that carry a concept without gradients or category supervision.

Finer-CAM attribution. We adopt Finer-CAM (Zhang et al., 2025) to visualize the regions that drive predictions of downstream ResNet-50 models trained on different selected subsets. For each image, we backpropagate the logit of its ground-truth category and apply Finer-CAM to the output of the final convolutional stage (`layer4`). The resulting heatmap is bilinearly upsampled to the input resolution, normalized to $[0, 1]$, and overlaid on the image.

D Additional Ablations

Table 11: **Ablation of representations used in concept coverage and local support.** SAE and DINO denote representations derived from SAE features and dense DINO features, respectively. We report accuracy for NABirds and worst-group accuracy (WGA) for Waterbirds. **Bold** and underlined entries indicate the best and second-best results, respectively.

Component		NABirds			Waterbirds		
Concept	Local	10%	30%	50%	10%	30%	50%
SAE	SAE	**42.9**	**67.4**	**75.2**	**68.6**	**77.9**	**79.3**
SAE	DINO	41.6	67.3	74.9	60.0	74.1	76.8
SAE	–	37.9	66.6	75.0	65.7	75.9	79.1
DINO	SAE	35.4	66.3	75.0	67.9	77.4	78.8
DINO	DINO	34.9	66.5	74.9	68.0	74.4	75.4
DINO	–	31.8	66.0	74.8	65.7	74.1	79.0
–	DINO	38.1	62.2	71.3	38.0	61.8	69.9
–	SAE	37.5	62.2	71.6	48.3	62.8	71.5

Table 12: **Ablation study on the neighborhood size k. Bold** and underlined entries indicate the best and second-best results, respectively.

k	NABirds Acc			Waterbirds WGA			Waterbirds Acc		
	10%	30%	50%	10%	30%	50%	10%	30%	50%
1	42.0	67.4	74.7	62.9	75.6	78.8	91.9	93.7	94.3
2	42.4	**67.5**	74.7	63.2	74.6	76.8	92.1	93.8	**94.9**
3	**42.9**	67.4	**75.2**	68.6	**77.9**	**79.3**	92.3	94.0	94.4
5	42.6	67.2	74.9	69.0	71.8	78.0	92.1	94.0	94.7
10	41.1	66.2	74.5	**71.5**	75.1	76.3	**92.7**	**94.2**	94.5
20	40.1	66.5	74.7	62.8	72.1	76.6	**92.7**	93.9	94.2

Table 13: **Effect of normalization** over neighbor similarity. We report accuracy for NABirds and worst-group accuracy (WGA) for Waterbirds. The better result in each column is bolded.

Assignment	NABirds Acc			Waterbirds WGA		
	10%	30%	50%	10%	30%	50%
Unnormalized	42.6	67.1	74.9	**69.1**	75.9	78.7
Row-normalized	**42.9**	**67.4**	**75.2**	68.6	**77.9**	**79.3**

D.1 Concept and Local Representations

Table 11 reports the per-rate results underlying Figure 3. Using SAE features for both components achieves the best result. Removing the concept component causes the largest drop, especially on Waterbirds at the 10% rate, where WGA falls from 68.6% to 38.0% and 48.3%. The advantage of SAE-based concept coverage is most pronounced at low selection rates: at 10% on NABirds, it outperforms all DINO-based concept variants by at least 6.2%.

Effect of similarity normalization. Table 13 reports the per-rate results underlying Table 3b. Row normalization improves performance in five of the six settings.

D.2 Local Support Weight β and Neighborhood Size k

Table 14: **Sensitivity to the local support weight β.** Comparison of performance across different local support weights and selection rates. We report classification accuracy (Acc.) for all classification datasets and worst-group accuracy (WGA) for Waterbirds and Spawrious. **Bold** and underlined entries indicate the best and second-best results, respectively.

β	NABird Acc			Stanford Cars Acc			Waterbirds Acc			Waterbirds WGA			Spawrious Acc			Spawrious WGA			Tiny-ImageNet Acc		
Selection Rate	10%	30%	50%	10%	30%	50%	10%	30%	50%	10%	30%	50%	10%	30%	50%	10%	30%	50%	10%	30%	50%
0	37.9	66.6	75.0	17.8	62.2	78.6	91.9	**94.6**	**94.5**	65.7	75.9	79.1	87.8	**87.3**	87.1	**76.1**	72.8	71.9	12.0	32.5	45.1
0.05	**42.9**	**67.4**	**75.2**	23.9	65.3	**78.8**	92.3	94.0	94.4	68.6	**77.9**	**79.3**	87.5	87.1	86.7	74.3	74.7	72.3	15.4	37.0	46.5
0.1	42.8	67.2	75.0	24.2	65.1	78.7	92.4	93.9	93.8	66.6	71.7	78.4	**88.1**	87.2	**87.4**	74.8	**76.9**	74.6	15.7	36.9	46.6
0.25	42.7	67.3	74.5	24.4	**65.8**	78.7	92.5	94.2	94.2	65.7	76.5	**79.3**	87.5	87.0	86.2	75.6	75.9	**74.9**	16.8	34.7	45.1
0.5	42.6	67.0	74.1	24.2	65.6	78.2	**92.8**	94.3	94.3	**69.8**	75.4	75.9	87.0	87.1	86.4	75.9	75.1	71.9	21.1	38.3	45.9
1.0	42.3	66.1	73.7	**25.3**	65.2	78.5	92.4	94.2	94.4	65.4	75.4	74.8	87.1	86.9	86.5	76.0	76.3	74.3	22.3	39.2	48.7
2.0	41.9	66.0	73.3	25.1	65.4	77.8	91.6	93.7	94.3	65.4	67.6	77.3	86.5	86.3	86.8	76.0	72.4	73.4	**23.5**	**40.4**	**50.0**

Sensitivity to the local support weight β. The local support weight β controls the balance between concept coverage and local support. Table 14 reports the full sweep underlying the summary in Figure 5a. The trend is consistent with the main paper's results: fine-grained and spurious-correlation settings favor a moderate amount of local support, with the strongest overall performance around $\beta = 0.05$, whereas general image classification benefits from a substantially larger weight, with the best performance at $\beta = 2$. This variation reflects differences in how concept coverage and local support interact across dataset regimes. We use the corresponding dataset-group setting throughout the main experiments.

Effect of neighborhood size k. We vary the number of nearest neighbors over $k \in \{1, 2, 3, 5, 10, 20\}$ in Table 12. Overall, $k = 3$ provides the most consistent performance across NABirds and Waterbirds,

while larger neighborhoods do not yield systematic improvements. We therefore use $k = 3$ as the default setting throughout our experiments.

D.3 Choice of Utility Function

Table 15: **Ablation study on the utility function.** We compare different utility functions used in our selection objective. Within each β group, the best and second-best results in each column are bolded and underlined, respectively.

β	Utility	NABirds Acc			Waterbirds Acc			Waterbirds WGA		
		10%	30%	50%	10%	30%	50%	10%	30%	50%
0	Linear	37.1	61.6	70.9	89.7	91.1	92.7	56.7	61.1	61.2
	Power ($\rho = 0.5$)	**39.9**	66.7	74.1	92.2	93.7	**94.6**	**68.3**	74.4	77.3
	Log	37.9	66.6	**75.0**	91.9	**94.6**	94.5	65.7	**75.9**	**79.1**
	Bounded exp.	37.9	**66.8**	**75.0**	**93.4**	94.0	94.5	67.1	74.1	**79.1**
0.05	Linear	37.8	61.1	69.9	90.4	90.9	91.8	50.6	59.4	61.2
	Power ($\rho = 0.5$)	40.3	66.6	74.1	92.1	93.0	93.7	65.4	65.3	74.8
	Log	**42.9**	**67.4**	**75.2**	92.3	**94.0**	**94.4**	68.6	**77.9**	**79.3**
	Bounded exp.	42.4	66.1	74.4	**92.6**	93.7	**94.4**	**69.6**	73.2	79.0

Our objective uses a logarithmic utility to impose diminishing returns on accumulated concept support. We ablate this choice by replacing the logarithm with a utility ϕ:

$$F_\phi(\mathcal{S}) = \sum_{c=1}^{C} \phi\left(\sum_{\boldsymbol{I}_i \in \mathcal{S}} \tilde{\boldsymbol{x}}_i[c]\right).$$

where $\tilde{x}_i[c]$ is the effective concept contribution defined in the §3. We compare four utilities:

$$\phi_{\text{linear}}(t) = t, \qquad \phi_{\text{power}}(t) = 2\big(\sqrt{1+t} - 1\big),$$
$$\phi_{\log}(t) = \log(1+t), \qquad \phi_{\exp}(t) = 1 - \exp(-t).$$

All four satisfy $\phi(0) = 0$ and $\phi'(0) = 1$, giving the same initial marginal value. The linear utility assigns constant value to additional support, whereas the remaining utilities introduce diminishing returns; the exponential utility additionally saturates at one. Table 15 compares these choices with $\beta = 0$ and $\beta = 0.05$. Nonlinear utilities generally outperform the linear alternative, supporting diminishing returns on accumulated concept support. We use the logarithmic utility for its strong overall performance across datasets and selection rates.

D.4 SAE Features vs. Selection Objective

Table 16: **Geometric selection in DINO and SAE feature spaces.** We compare K-Center using dense DINO features and sparse SAE features.The better result is **bolded**.

Selection Rate	NABirds Acc		Waterbirds WGA	
	K-Center (DINO)	K-Center (SAE)	K-Center (DINO)	K-Center (SAE)
10%	31.6	**32.7**	57.5	**61.2**
30%	64.9	**66.5**	64.2	**70.3**
50%	74.1	**75.0**	**79.4**	77.0

A natural question is whether the gains come from the SAE representation itself or from our selection objective. As shown in Table 16, replacing DINO features with SAE features partially improves K-Center. On NABirds, the improvement is modest and K-Center (SAE) still underperforms Random at the 10% budget. Nevertheless, K-Center (SAE) remains below our method on both datasets. These results show that the SAE representation is useful, but the main benefit comes from explicitly reallocating support across concepts.

Table 17: **Robustness to SAE pretraining data.** We compare SAEs trained on the target dataset with an SAE trained on ImageNet-1K. **Bold** and underlined entries indicate the best and second-best results, respectively.

Selection Rate	NABirds Acc		Waterbirds WGA		Waterbirds Acc	
	Target SAE	IN1K SAE	Target SAE	IN1K SAE	Target SAE	IN1K SAE
10%	**43.0**	42.9	63.4	**68.6**	91.7	**92.3**
30%	**67.4**	**67.4**	73.2	**77.9**	93.8	**94.0**
50%	74.4	**75.2**	78.7	**79.3**	**95.1**	94.4

D.5 SAE Pretraining Data

Table 17 compares SAEs trained on each target dataset with a single SAE pretrained on ImageNet-1K. Across NABirds and Waterbirds, the ImageNet-1K SAE performs comparably to target-trained SAEs. This indicates that the learned sparse concept space transfers well across datasets and does not require retraining an SAE for each new selection task. We therefore use a single ImageNet-1K-pretrained SAE throughout our main experiments.

D.6 Downstream Architecture

Table 18: **Cross-architecture generalization.** Comparison of subset selection performance using the same selected subsets with ResNet-50 and ViT-B/16. We report classification accuracy (Acc.) for NABirds, and worst-group accuracy (WGA) and overall accuracy (Acc.) for Waterbirds; each cell shows ResNet-50 / ViT-B/16. **Bold** and underlined entries indicate the best and second-best results for each architecture, respectively.

Method	NABirds Acc			Waterbirds WGA			Waterbirds Acc		
Selection Rate	10%	30%	50%	10%	30%	50%	10%	30%	50%
Random	35.6 / 40.2	64.9 / 65.9	73.3 / 73.8	52.3 / 54.3	67.5 / 66.0	68.8 / 68.6	84.1 / 76.7	89.6 / 85.8	91.6 / 85.9
K-Center	31.6 / 33.7	64.9 / 65.0	74.1 / 73.5	57.5 / 21.7	64.2 / 49.5	**79.4** / 62.3	89.4 / 85.5	89.9 / 84.9	91.7 / 88.7
Forgetting	22.4 / 26.9	63.3 / 63.6	74.3 / 73.7	31.6 / 3.9	60.1 / 25.6	73.4 / 54.4	73.8 / 75.7	85.1 / 81.7	91.0 / 89.8
EL2N	16.8 / 21.0	61.2 / 61.0	74.4 / 73.4	31.0 / 30.4	64.2 / 38.2	72.9 / 63.2	78.4 / 79.9	87.0 / 85.0	92.7 / 89.7
D2 Pruning	25.8 / 31.0	66.6 / 66.6	74.7 / 74.1	47.0 / 50.5	56.4 / 37.5	60.8 / 47.2	89.0 / 88.6	92.2 / 87.9	93.3 / 91.2
Moderate	34.6 / 37.7	64.9 / 66.1	73.7 / 73.8	60.5 / 60.6	68.0 / 57.5	71.1 / 61.4	91.2 / 88.0	91.8 / 90.7	92.5 / 88.7
CCS	39.7 / 43.5	65.5 / 66.9	73.6 / 73.6	52.8 / 38.8	67.1 / 55.9	69.9 / 64.5	86.3 / 73.6	90.4 / 86.4	92.0 / 87.0
InfoMax	22.5 / 27.2	64.7 / 64.9	72.7 / 72.8	22.6 / 30.5	61.7 / 42.8	70.9 / 68.4	74.4 / 65.4	88.3 / 85.3	91.8 / 89.2
RAM-APL	39.3 / 43.6	61.7 / 64.5	69.8 / 71.6	30.2 / 13.9	51.2 / 43.0	65.0 / 24.1	75.6 / 73.7	82.7 / 80.7	88.8 / 85.3
Ours	**42.9 / 44.9**	**67.4 / 67.6**	**75.2 / 74.4**	**68.6 / 63.7**	**77.9 / 68.2**	79.3 / **80.4**	**92.3 / 92.2**	**94.0 / 91.9**	**94.4 / 91.8**

We evaluate the same selected subsets with both ResNet-50 and ViT-B/16. As shown in Table 18, the advantage of our subsets is consistently preserved after switching the downstream architecture. With ViT-B/16, our method achieves the best NABirds accuracy, Waterbirds WGA, and Waterbirds overall accuracy across all three selection rates. The result shows that the benefit of our method remains stable across different downstream model families.

E Additional Results

E.1 ImageNet-1K and Distribution Shift

Table 19: **ImageNet-1K classification and robustness results.** We report top-1 accuracy (%) on four robustness benchmarks and the standard ImageNet-1K validation set at different selection rates. The best and second-best results are bolded and underlined, respectively.

Method	ImageNet-A			ImageNet-R			ImageNet-D			ImageNet-V2			ImageNet-1K Val		
	10%	30%	50%	10%	30%	50%	10%	30%	50%	10%	30%	50%	10%	30%	50%
Random	1.4	1.2	1.1	15.8	24.1	**25.9**	2.9	6.3	6.8	36.3	49.3	50.9	48.6	64.2	67.2
EL2N	1.2	1.0	1.1	9.3	19.4	23.2	1.4	3.6	6.1	13.4	23.4	45.2	11.2	45.4	64.1
CCS	1.3	1.6	1.4	16.0	24.5	25.1	2.7	6.2	6.9	**37.4**	49.1	**51.2**	**50.1**	64.8	**68.4**
Ours	**1.7**	**2.0**	**1.5**	**16.4**	**25.8**	**25.9**	**3.3**	**6.8**	**7.6**	37.2	**50.1**	50.7	49.2	**65.2**	67.7

Standard classification. We evaluate ImageNet-1K subset selection at the 10%, 30%, and 50% selection rates, comparing our method with Random, EL2N, and CCS under identical downstream training settings. As shown in Table 19, our method improves validation accuracy over random selection at all three rates, although the gains are modest compared with those on the fine-grained, spurious-correlation, and multimodal pretraining benchmarks. CCS is competitive on this benchmark.

The smaller gains on ImageNet-1K may reflect differences in which visual evidence limits learning across dataset regimes. In fine-grained and spurious-correlation settings, under-supported concepts and contextual variation are closely associated with where larger gains occur, as we discussed in §5. One possible explanation is that random ImageNet-1K subsets already preserve much of the evidence needed for standard classification, leaving less room for improvement through concept-support redistribution. Exploring how concept support should be characterized and reallocated in broad, general-purpose datasets is an important direction for future work.

Robustness under distribution shift. While our method does not establish a clear advantage on the standard validation set, we further compare the performance on robustness test sets. Specifically, we evaluate the same models on ImageNet-A (Hendrycks et al., 2021b), ImageNet-R (Hendrycks et al., 2021a), ImageNet-D (Zhang et al., 2024), and ImageNet-V2 (Recht et al., 2019). Our method achieves the best accuracy on ImageNet-A, ImageNet-R, and ImageNet-D at all three selection rates, while CCS performs better on ImageNet-V2 at the 10% and 50% rates. The benefit of our selection is therefore more pronounced under larger distribution shifts than on the standard validation set, benefiting from broader concept coverage.

E.2 Concept-Support Profiles of All Selectors

Table 20a extends the concept-support profile of Table 4 to all baselines on 10% Stanford Cars and NABirds subsets. The trend is consistent across selectors and datasets: every baseline, whether geometry-based, score-based, or distribution-based, remains close to the random profile, whereas our selection substantially spreads support more evenly and shifts it toward less prevalent concepts.

E.3 Additional SAE Concept Visualizations

Figure 8 provides additional examples of concepts whose support is most enhanced or reduced by our selection. Enhanced concepts include uncommon contextual evidence such as human hands, sign edges, human figures, and suspension bridges, while reduced concepts include recurring visual patterns such as waves, sky backgrounds, trees, and tires. These examples illustrate that our selection redistributes a fixed training budget across visual factors rather than simply increasing overall concept coverage. Support is shifted away from repeatedly represented patterns toward evidence that is less frequently retained under random selection.

Table 20: **Concept-support profiles of 10% subsets for all baselines on (a) NABirds and (b) Stanford Cars.** Every baseline stays close to the random profile on all three measures, whereas our selection covers more concepts, spreads its support more evenly, and shifts support toward less prevalent concepts.

Methods	Unique Coverage	Effective Number	Prevalence Ratio
Random	7,973	865	1.00
K-Center	8,215	893	0.97
Moderate	8,112	876	0.99
CCS	7,855	846	1.02
Forgetting	8,120	888	0.97
EL2N	8,181	910	0.94
D^2-Pruning	8,099	887	0.97
InfoMax	8,205	902	0.95
RAM-APL	7,449	790	1.09
SDM	7,918	891	0.95
Ours	**8,951**	**1,124**	**0.74**

(a) NABirds

Methods	Unique Coverage	Effective Number	Prevalence Ratio
Random	9,321	1,491	1.00
K-Center	9,320	1,481	1.00
Moderate	9,261	1,477	1.01
CCS	9,104	1,468	1.02
Forgetting	9,219	1,516	0.99
EL2N	9,296	1,532	0.97
D^2-Pruning	9,304	1,480	1.01
InfoMax	9,271	1,522	0.98
RAM-APL	8,792	1,390	1.06
SDM	9,026	1,544	0.95
Ours	**9,510**	**1,853**	**0.77**

(b) Stanford Cars

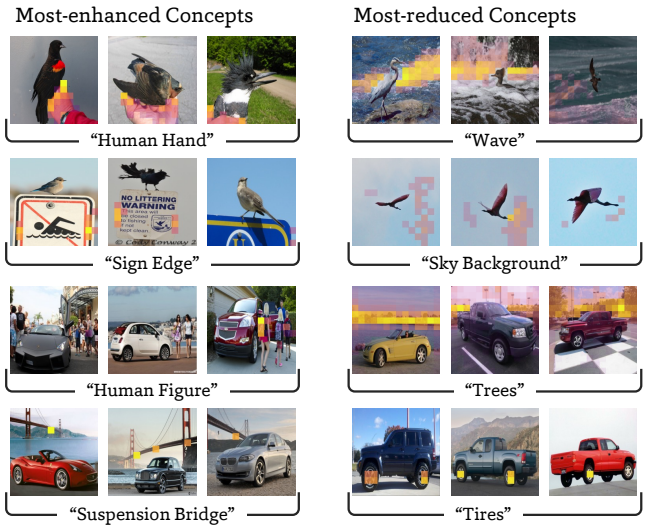


Figure 8: **Additional examples of enhanced and reduced SAE concepts.** We show representative images from concepts whose support changes most under our selection relative to random selection. Highlighted patches indicate high-activation regions.

E.4 Additional Finer-CAM Visualizations

Figure 9 provides additional Finer-CAM examples to complement the qualitative analysis in Figure 6. Across multiple enhanced concepts, models trained on our selected subsets consistently place more attribution on the target object, while models trained on random and prior works selected subsets more often emphasize the surrounding contextual evidence.

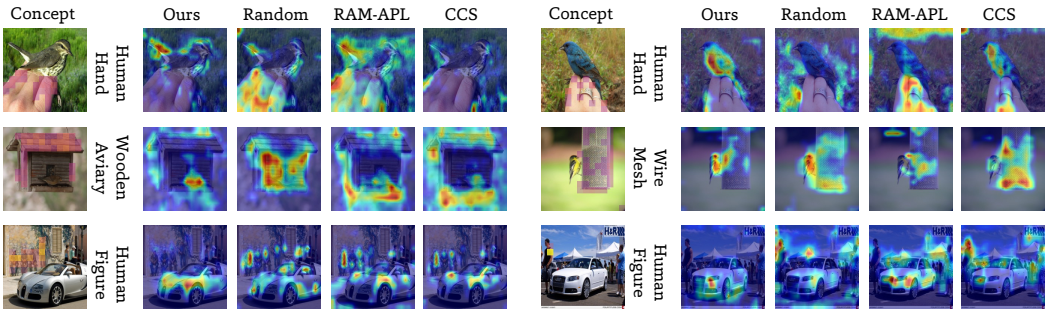


Figure 9: **Additional Finer-CAM visualizations on NABirds and Stanford Cars.** We show attribution maps for models trained on subsets selected by our method, Random, RAM-APL, and CCS. The *Concept* column highlights image patches with high activation for the corresponding SAE latent.

F Relation to Concept-Based Data Selection

This work studies data selection from the perspective of how a limited training budget is distributed across visual concepts. This direction is complementary to recent advances that use concepts as signals for selecting individual examples. Yang et al. (2025) use SAE features to promote feature diversity, Mehra et al. (2025) estimate sample difficulty through concept bottlenecks, and SAE-V (Lou et al., 2025) uses concept-level weights for multimodal filtering. These works demonstrate that concept representations can provide useful semantic information for data selection. Our work considers concept support at the level of the selected subset. We characterize how limited training support is distributed across concepts and how this allocation changes with subset selection. Because each image contains multiple visual factors, selecting diverse samples can still repeatedly reinforce already common evidence. Our results show that making this allocation explicit helps reveal where the training budget is spent and how these changes relate to downstream behavior.

A parallel line of work uses concept representations to align selection with a desired target. MONA (Ma et al., 2025) selects examples according to a target-task prototype, while SDM (Das et al., 2026) matches a reference concept distribution. These approaches are complementary to our setting. Related methods also differ in how the subset is constructed: D^2-Pruning (Maharana et al., 2024) propagates scalar difficulty information over a sample graph, whereas Zhou et al. (2026) control concept exposure dynamically by changing training samples across epochs. We instead study concept support as a property of a fixed selected subset. This perspective allows us to examine not only whether concept-aware selection improves performance, but also how changing the support of particular visual evidence is associated with fine-grained recognition, contextual robustness, and multimodal pretraining.