Scalable Auditing of Large Image Datasets Using Sparse Linear Representations

Alex Fan, Arnav Gurudatt, Karsten Kropp, Gabriel López-Asiaín

University of Chicago August 27, 2026

Abstract. Deep neural networks trained on image datasets frequently exploit spurious correlations that are difficult to detect because the relevant biases are often unknown *a priori* and modern datasets are too large for manual inspection. We introduce an interactive web tool for concept-level dataset auditing that combines CLIP's semantic embedding space with SpLiCE [1], which decomposes image embeddings into sparse, human-readable concept representations. Users compose queries as weighted combinations of natural-language concepts, retrieve matching images via nearest-neighbor search, and inspect an enrichment panel that automatically surfaces concepts overrepresented in the retrieved subset relative to the dataset baseline. This workflow supports hypothesis-free exploration of potential dataset biases: rather than specifying a suspected association in advance, users discover candidate associations directly from what the retrieval returns. We demonstrate the tool on MSCOCO, where it surfaces asymmetric gendered associations and spurious concept co-occurrences that would be difficult to anticipate manually.

1 Introduction

Classifiers trained on image datasets often learn spurious correlations—statistical associations between a label and an incidental visual feature that fail to generalize [2]. These associations may be artifacts of dataset construction, annotation practices, or broader social biases embedded in the training data, rather than features that are genuinely relevant to the target class. In high-stakes settings, these failures can reinforce existing inequities by encoding biased associations into downstream prediction, retrieval, or decision-support systems. For example, a multi-modal image generation model that uses textual prompts may learn to render gender-neutral descriptions in gender-stereotyped ways, depicting women disproportionately in care, household, shopping, fashion, or intimate-wear contexts while depicting men disproportionately in business, technical, transportation, or physical-labor contexts [3]. Despite the social and technical consequences of these learned spurious associations, detecting them remains difficult; existing methods require either labeled subgroups, manual inspection of large collections, or prior knowledge and hypotheses of which spurious feature to look for.

CLIP [4] offers a way out of this impasse. By embedding images and text into a shared semantic space, it enables navigation of large image collections through natural language queries alone, with no labeled subgroups or task-specific training. However, CLIP embeddings are high-dimensional and opaque, which makes it difficult to understand what associations the model has internalized. SpLiCE [1] addresses this by decomposing CLIP embeddings into sparse, non-negative linear combinations of human-readable concepts: rather than an opaque 512-dimensional vector, an image is expressed as a weighted sum of concepts such as "dog", "grass", or "running", requiring no training or concept labels.

We build on both to introduce an interactive auditing tool, demonstrated on the Microsoft Common Objects in Context (MSCOCO) [5] dataset, that allows users to compose queries as weighted combinations of natural-language concepts, retrieve matching images via nearest-neighbor search with no ML inference at query time, and visually inspect the results. Alongside the retrieved images, an enrichment ratio panel ranks concepts by their overrepresentation in the retrieved subset relative to the dataset baseline, surfacing associations the user may not have anticipated. Surfaced concepts can be injected back into the query builder directly, creating a human-in-the-loop interface in which the user iteratively refines or redirects the retrieval based on what they observe.

2 Background

CLIP. Contrastive Language–Image Pretraining (CLIP) [4] learns a shared semantic embedding space between images and natural language by training on approximately 400 million image–text pairs collected from the internet. CLIP uses a dual-encoder architecture consisting of an image encoder and a text encoder trained jointly with a contrastive objective. During training, corresponding image–text pairs are encouraged to occupy nearby regions of the embedding space, while unrelated pairs are pushed apart. This contrastive formulation enables the model to

learn high-level semantic relationships without requiring manually curated class labels.

The resulting embedding space captures rich semantic structure across both modalities. Semantically related concepts such as "dog", "puppy", and "golden retriever" cluster together geometrically, while unrelated concepts occupy distant regions of the space. Because both images and text are embedded into the same latent geometry, CLIP supports zero-shot retrieval and classification through nearest-neighbour similarity alone. In this work, we use the ViT-B/32 variant via OpenCLIP [6], which provides publicly reproducible CLIP weights and inference pipelines.

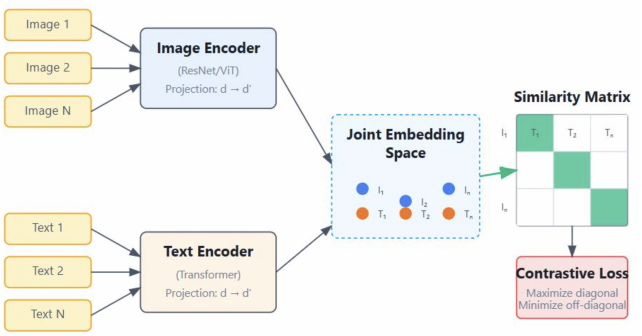

Figure 1: Overview of the CLIP architecture. Images and text are encoded into a shared embedding space using dual encoders trained with a contrastive objective, allowing semantically related image-text pairs to align geometrically.

SpLiCE. Although CLIP embeddings exhibit strong semantic structure, the learned representation space remains difficult to interpret directly because each embedding dimension lacks a clear human-readable meaning. Bhalla et al. introduced Sparse Linear Concept Embeddings (SpLiCE) [1] as a method for decomposing CLIP embeddings into sparse combinations of interpretable natural-language concepts. Rather than treating an image embedding as an opaque point in $\mathbb{R}^d$, SpLiCE expresses the embedding as a weighted combination of concepts drawn from a large vocabulary of phrases.

Given a CLIP image embedding $\mathbf{z} \in \mathbb{R}^d$, SpLiCE solves the following optimization problem for a vector $\mathbf{w}$:

$$\min_{\mathbf{w} \geq 0} \|\mathbf{z} - W\mathbf{w}\|_2^2 + \lambda \|\mathbf{w}\|_1,$$

where $W \in \mathbb{R}^{d \times V}$ is a concept dictionary constructed from CLIP text embeddings. For this work, we construct W from the V=12,000 most frequent concepts present in the MSCOCO caption corpus. The non-negativity constraint ensures that concepts contribute additively, while the ℓ_1 penalty encourages sparsity so that only a small number of concepts receive substantial weight. Optimization is performed using the alternating direction method of multipliers (ADMM) algorithm, which decomposes the constrained optimization problem into simpler subproblems that can be solved iteratively while enforcing non-negativity and sparsity. The resulting vector $\mathbf{w}$ provides an interpretable semantic decomposition of the image, enabling downstream analysis of concept co-occurrence and contextual associations.

Enrichment ratio. To identify concepts that are disproportionately associated with a retrieved image subset, we compare the frequency of each concept within the retrieved set against its baseline frequency across the full dataset. For a retrieved subset $S \subset \mathcal{D}$ and concept c, we define the retrieved-set frequency as

$$f_c(S) = \frac{1}{|S|} \sum_{i \in S} \mathbf{1}\{w_{ic} > 0\},$$

where w_{ic} is the SpLiCE weight assigned to concept c for image i. This measures the fraction of retrieved images in which the concept appears with nonzero activation.

We then define the enrichment ratio

$$\rho_c = \frac{f_c(S)}{f_c(\mathcal{D})},$$

where $f_c(\mathcal{D})$ is the baseline frequency of concept c across the entire dataset. Values $\rho_c > 1$ indicate that a concept appears more frequently in the retrieved subset than expected from the dataset baseline. For example, if a concept appears in 18% of the retrieved images but only 5% of the dataset overall, then its enrichment ratio is $0.18/0.05 = 3.6$.

Note that a high enrichment ratio does not directly imply a spurious or biased association; surfaced concepts are candidates that require human interpretation in light of the query and the retrieved images. Some may reflect social biases, such as "male" appearing enriched for "doctor"; others may be entirely expected, such as "snow" appearing enriched for "skiing".

MSCOCO. Microsoft Common Objects in Context (MSCOCO) [5] is a large-scale image dataset introduced to advance object recognition in realistic, cluttered scenes. Images were sourced from Flickr and selected to depict everyday objects in their natural context rather than in isolation. The 2017 split, which we use in this work, contains approximately 118,000 training images across 80 object categories. Because images were drawn from Flickr without explicit demographic controls and annotated by crowdworkers, the dataset reflects biases present in both the source platform and its annotator pool, making it a natural testbed for concept-level bias auditing.

FAISS. FAISS (Facebook AI Similarity Search) [7] is a library for efficient nearest-neighbor search over dense vector collections. Given a query vector, FAISS retrieves the most similar vectors from a prebuilt index using exact

or approximate search strategies. In this work we use an exact flat index over the CLIP image embeddings of MSCOCO, which allows us to retrieve the top-k most similar images to any query vector in milliseconds without rerunning any neural network inference at query time.

3 System Design

3.1 Implementation

The system is implemented as a lightweight web application with a FastAPI backend and a Svelte 5 frontend built with Vite. All representation work is performed offline: CLIP embeddings are extracted for MSCOCO images using OpenCLIP [6], SpLiCE decompositions are precomputed using the reference SpLiCE implementation [1], and a FAISS flat index is built over the resulting embeddings. These artifacts, together with dataset-level concept statistics and image metadata, are loaded into memory at startup.

3.2 Interaction Model

Users compose a query by adjusting concept weight sliders. The backend constructs $\mathbf{q} = \sum_c w_c \mathbf{e}_c$, where w_c is the user-assigned weight for concept c and $\mathbf{e}_c$ is its CLIP text embedding. The normalized query vector is passed to the FAISS index, which returns the nearest images in real time. Concept weights are rescaled to sum to one as sliders are adjusted, keeping the query a convex combination of concept embeddings. Concepts surfaced by the enrichment panel can be injected back into the query builder with a single click, allowing users to iteratively redirect the retrieval based on what they observe.

3.3 Unexpected Concepts Panel

For each retrieved subset, the backend computes enrichment ratios and retrieved-set frequencies over the full set of returned images. The panel exposes both ranking modes interactively, with horizontal bars visualizing relative prominence. Query concepts are excluded from the ranking to avoid trivial returns, and globally dominant concepts are masked to reduce noise. Clicking any surfaced concept injects it directly into the query builder, closing the human-in-the-loop cycle between retrieval, inspection, and refinement. Figure 2 shows a screenshot of the tool.

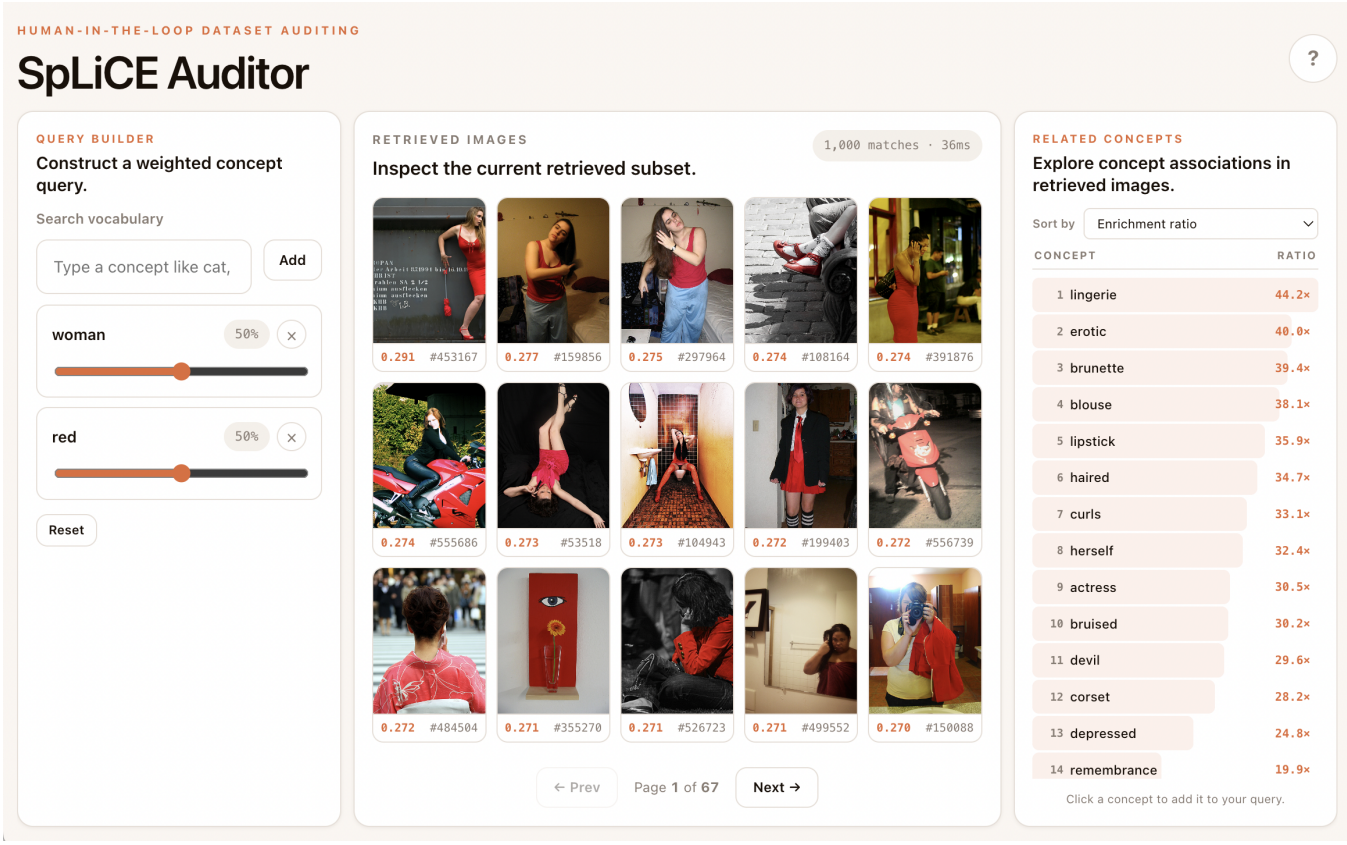

Figure 2: Bias-auditing retrieval example. Querying for "woman" and "red" surfaces neutral contextual associations such as "blouse" and "lipstick", but also lewd associations like "lingerie" and "erotic." The enrichment scores allow users to inspect how concepts are distributed within the retrieved subset relative to the broader dataset.

4 Case Study: Gender Bias Auditing

We first examine whether our auditing tool can recover asymmetric gender associations in MSCOCO. Bhalla et al. [1] report an association between "woman" and swimwear-related concepts in CIFAR-100; we observe a similar pattern in MSCOCO, but with a broader set of sexualized, appearance-based, and emotionally coded associations. Table 1 compares representative retrieved-set frequency and enrichment results for the queries "woman" and "man."

Query	Concept	Freq.	Enrichment
woman	pregnant	6%	50.2×
woman	lingerie	4%	59.9×
woman	erotic	4%	58.3×
woman	actress	4%	51.2×
woman	cleavage	3%	42.0×
woman	crying	3%	30.6×
woman	bruised	3%	32.7×
woman	curls	3%	47.3×
man	beard	4%	50.3×
man	mustache	3%	26.5×
man	collar	3%	27.8×
man	formal	2%	28.3×
man	flexing	2%	27.3×
man	gentleman	2%	24.4×
man	manager	2%	20.9×
man	gangster	2%	19.9×
man	faculty	1%	10.2×

Table 1: Top concepts associated with "woman" and "man" queries. Frequency is the share of retrieved images in which the concept has nonzero SpLiCE weight. Enrichment compares that frequency to the concept's dataset-wide baseline.

The "woman" query surfaces a sexualized and appearance-centered set of associations: "lingerie" and "erotic" are enriched by $\sim 60\times$, while "pregnant," "cleavage," "crying," and "bruised" point to a broader pattern of emotional and bodily associations. By contrast, the "man" query is dominated by facial hair, formality, and power-coded occupations such as "manager," "faculty," and "gentleman," with comparatively few sexually explicit concepts. These results suggest that the representation associates women more strongly with sexuality and emotion, and men more strongly with professionalism and institutional status.

These asymmetric associations are not confined to explicitly gendered queries. Table 2 shows results for "doctor," "nursing," and "manager." The "doctor" query retrieves overwhelmingly male images and surfaces concepts such as "manager," "faculty," and "beard" alongside medically relevant terms, directly overlapping with the "man" query. The "manager" query shows the same pattern. By contrast, "nursing" retrievals are overwhelmingly women, with associated concepts that are gendered but not necessarily spurious, since the term refers both to a healthcare occupation and to infant feeding. Taken together, these findings indicate that asymmetric gender structure propagates beyond explicit demographic labels and into apparently neutral occupational concepts.

Query	Concept	Freq.	Enrich.
doctor	scientist	10%	71.6×
doctor	dr	9%	67.4×
doctor	medic	7%	61.2×
doctor	manager	5%	73.2×
doctor	faculty	5%	37.3×
doctor	beard	4%	50.2×
nursing	medical	6%	75.0×
nursing	clinic	5%	48.3×
nursing	surgical	5%	44.4×
nursing	birth	4%	40.3×
nursing	pregnant	3%	26.7×
nursing	nurses	3%	33.1×
manager	associate	6%	87.9×
manager	employee	5%	41.7×
manager	faculty	5%	34.3×
manager	agent	5%	55.5×
manager	beard	3%	34.0×
manager	formal	3%	33.3×

Table 2: Top concepts associated with occupational queries. Both "doctor" and "manager" surface appearance and status concepts that overlap directly with the "man" query, suggesting the representation links professional authority with male-coded visual features.

5 Discussion

5.1 Limitations and Failure Cases

The tool inherits the imperfections already present in CLIP and SpLiCE. CLIP embeddings encode associations learned from web-scale data that may not reflect semantic correctness, and SpLiCE decompositions are an approximation of those embeddings rather than a ground-truth concept attribution. This means that enriched concepts can reflect artifacts of the learned representation rather than genuine properties of the dataset. The "man" query illustrates this: as shown in Figure 3, several top retrievals contain images with text such as "Main Street" or bus labels reading "MAN," suggesting that CLIP's embedding space partially responds to visual text patterns rather than purely semantic content.

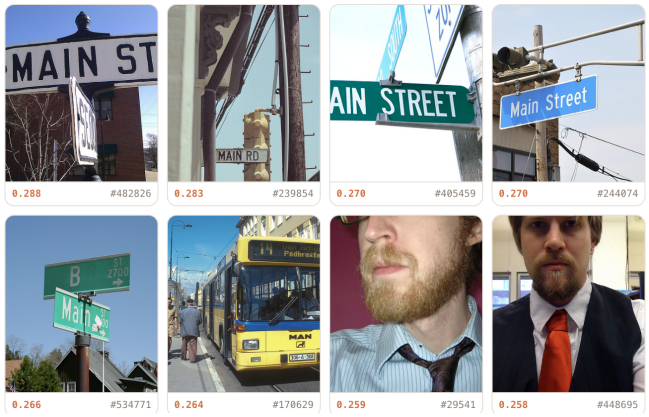

Figure 3: Example of non-semantic shortcut associations for the query "man." Several highly ranked retrievals contain "Main Street" signs or other textual patterns containing the letters "m", "a", and "n", suggesting the embedding representation partially responds to textual substrings within images.

Cosine similarity is also an imperfect retrieval mechanism. The system will always return the top-k images above a threshold, even when the dataset contains no true examples of the queried concept. For example, querying "kangaroo" retrieves koalas, marsupial-adjacent animals, and zoo imagery from Australia rather than kangaroos themselves, as shown in Figure 4. These failure cases reinforce why the system should be treated as a human-in-the-loop auditing interface: retrieved images must be visually inspected to determine whether a surfaced association reflects a genuine dataset property or an artifact of the underlying representations.

Figure 4: Retrievals for the query "kangaroo" surface koalas and zoo imagery from Australia rather than kangaroos, illustrating how cosine similarity can match semantically adjacent concepts when the queried concept is absent from the dataset.

5.2 Future Directions

Two natural extensions would strengthen the auditing framework. First, the concept vocabulary used for SpLiCE decomposition could be more carefully curated. The current vocabulary is derived from raw caption frequency in MSCOCO, which means it includes generic or ambiguous terms while potentially missing concepts that are more diagnostically useful for bias detection. A structured vocabulary grounded in social science literature or demographic taxonomies could surface more meaningful associations and reduce noise in the enrichment panel.

Second, the findings produced by the tool are currently qualitative. The enrichment panel surfaces candidate associations and gives the user a first indication of their strength, but it does not provide a rigorous way to confirm whether a flagged association reflects a genuine bias or how large its effect is. A natural next step would be to pair the tool with a more systematic downstream evaluation protocol, so that hints produced by the tool can be turned into confirmed, quantifiable findings.

References

[1] U. Bhalla, A. Oesterling, S. Srinivas, F. P. Calmon, and H. Lakkaraju, "Interpreting CLIP with sparse linear concept embeddings (SpLiCE)," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[2] S. Wu, M. Yuksekgonul, L. Zhang, and J. Zou, "Discover and cure: Concept-aware mitigation of spurious correlation," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2023.

[3] L. Girrbach, S. Alaniz, G. Smith, and Z. Akata, "A large scale analysis of gender biases in text-to-image generative models," 2025.

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

[5] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *European Conference on Computer Vision (ECCV)*, pp. 740–755, 2014.

[6] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev, "Reproducible scaling laws for contrastive language–image learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829, 2023.

[7] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2019.