Investigating Demographic Attribute Representation in the Florence-2 Vision Encoder using Sparse Autoencoders

Ozan Bayiz
ozanbayiz@berkeley.edu

Kapil Malladi
kapilmalladi@berkeley.edu

Charlie Cooper
charlie.c@berkeley.edu

Raiyan Hammad Ausaf
raiyanausaf14@berkeley.edu

Abstract

Machine learning systems are prone to exhibiting misaligned behavior when inputs include sensitive attributes such as race, age or gender. Understanding how neural networks encode these attributes is essential for achieving fairness, transparency, interpretability, and overall alignment. This project investigates activations of the vision encoder (VE) of a vision language model (VLM) – in this case a vision transformer (ViT) – when provided input of images of people of different races, genders, and ages. We specifically use Microsoft's Florence-2-base model [1], and perform the following analyses: First, we determined the degree to which the VE's output activations contain features that are relevant to encoding demographic information by training linear probes to classify VE encoded images of people based on race, gender, and age. We then train sparse autoencoders (SAEs) to decompose these activations into human-interpretable latent features. Finally, we identify sparse dictionary features (SDFs) that are correlated with particular demographic features, providing insights into how these attributes might be represented.

Code is accessible at: https://github.com/ozanbayiz/idarve

1 Introduction

Vision Language Models (VLMs) like Microsoft's Florence-2 [1] represent a significant leap in AI, adeptly integrating visual and textual information for diverse tasks such as image captioning and visual question answering. However, these powerful models often inherit and amplify societal biases related to sensitive attributes like race, gender, and age from their vast training data, leading to potentially harmful stereotypes and eroding user trust.

Addressing this critical challenge requires moving beyond black-box evaluations to understand the internal mechanisms by which these biases are encoded, particularly within the vision encoder (VE) component. While techniques like Sparse Autoencoders (SAEs) have shown promise for decomposing representations and enhancing interpretability in large language models, their application towards understanding demographic bias within the visual pathways of VLMs remains relatively under-explored.

This project tackles this gap by investigating how race, gender, and age are encoded within the DaViT vision encoder of the Florence-2-base model [1]. We employ a multi-faceted approach, utilizing linear probes to assess the linear separability of demographic information and training SAEs to identify interpretable latent features correlated with these attributes. Our work aims to provide deeper insights into bias encoding in modern VLMs and explore the utility of SAE-based interpretability for fairness analysis in the vision-language domain.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

2 Related Work

This research integrates concepts and methodologies from several key areas: fairness in vision-language models (VLMs), techniques for probing neural representations, and mechanistic interpretability using Sparse Autoencoders (SAEs), particularly in the visual domain.

Probing Representations for Sensitive Attributes

A common first step in assessing bias is to determine if sensitive attribute information is decodable from model activations. Linear probing, as employed in our initial step, involves training simple classifiers on frozen model representations to predict attributes like race or gender. This technique has been widely used across modalities, including in vision models and VLMs, to quantify the presence of linearly separable demographic information at different network layers [e.g., 2, 3]. While successful probes indicate the presence of information, they do not confirm its causal role in downstream tasks [4], motivating deeper analysis.

Sparse Autoencoders for Mechanistic Interpretability

To move beyond simple decodability and understand *how* information is structured, we employ Sparse Autoencoders (SAEs). Primarily motivated by the superposition hypothesis – the idea that models represent more concepts than they have neurons, encoding them in overlapping, entangled patterns [5] – SAEs aim to decompose dense activation vectors into a sparse, potentially more interpretable basis of features (Sparse Dictionary Features or SDFs) [6]. Initial successes focused on Large Language Models (LLMs), demonstrating that SAEs can uncover human-understandable concepts within complex text representations [e.g., 7, 8].

Patch-Level SAEs for Vision Transformers

While LLM interpretability via SAEs is advancing, applying these techniques effectively to vision encoders, particularly within VLMs, is an active area of research. Critically relevant to our approach is the work by Lim et al. [9] on PatchSAE. Recognizing that Vision Transformers (ViTs) process images as sequences of patches, PatchSAE trains the SAE directly on these patch-level activations within CLIP's ViT. This enables the discovery of interpretable visual concepts with *spatial attribution*, localizing feature activity within the image [9]. This contrasts with earlier SAE applications that might focus on globally pooled representations (similar to our initial, revised SAE approach) and provides a more granular tool suitable for analyzing the distributed representations within ViT architectures like the DaViT encoder in Florence-2 [1]. Other studies have also explored SAEs on CLIP's vision encoder [Fry, 2024; Daujotas, 2024a, cited in 9].

Targeted Feature Discovery vs. Broad Interpretation

A key distinction of our research lies in its objective. Much prior SAE work aims broadly to discover and interpret the dictionary of features learned by the model – essentially asking "what concepts does the model represent?" [7, 8, 9]. In contrast, our investigation starts with specific attributes of interest – namely, sensitive demographic characteristics (race, gender, age) – and asks "how, and where spatially, are these pre-defined attributes encoded within the vision encoder's representations?" This involves identifying potentially complex patterns, possibly involving combinations of SDFs, that correlate with these specific attributes, rather than solely interpreting the most prominent features learned by the SAE. This targeted approach is crucial for fairness research, where understanding the representation of specific sensitive information is paramount.

Analyzing and Filtering SAE Features

Once an SAE is trained, identifying features relevant to a specific concept (like race) requires analysis of their activation patterns. Our filtering pipeline utilizes metrics such as activation frequency, mean activation value, and label entropy. These statistical measures are commonly employed or discussed in SAE literature for characterizing learned features and assessing their properties, including specificity [9, 6]. Visualizing top-activating inputs is also a standard technique for qualitative interpretation of SAE features [9].

3 Methodology

Our methodology investigates the encoding of demographic attributes, specifically race, within the vision encoder (VE) of the Florence-2-base model [1] through a sequence of analysis steps. To mitigate confounding factors arising from dataset bias during analysis, we primarily utilize the **FairFace dataset** [10]. FairFace is specifically chosen for its explicit balancing across 7 race categories (White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, Latino), gender, and age groups, making it well-suited for training models and analyzing correlations related to these attributes with reduced risk of learning spurious correlations due to demographic imbalance [.1960].

3.1 Quantifying Demographic Signal with Linear Probing

We begin by establishing a quantitative baseline confirming the presence of linearly decodable race information within the VE representations, thereby justifying deeper interpretability analysis. We train linear classifiers (probes) using VE activations extracted from FairFace images. Following standard practice, we initially average the patch-level activations from the VE's final layer (`(num_patches, hidden_size)`) into a single image-level vector (`(hidden_size)`) as input to the probe. A linear probe trained on these average-pooled activations achieved 62.15% accuracy in classifying the 7 balanced race categories from FairFace. This performance significantly surpasses the random chance baseline (14.3%), confirming that race-related information is sufficiently encoded in the VE activations to warrant investigation via Sparse Autoencoders (SAEs).

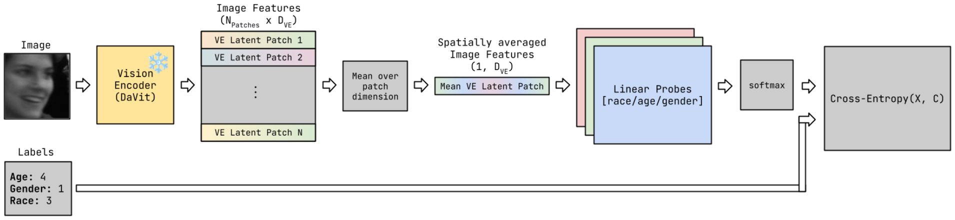


Figure 1: Linear probing pipeline. An image is processed by a Vision Encoder, from which patch features are extracted and averaged. These averaged features are then fed into separate linear probes to predict race, age, and gender. A softmax function generates output probabilities, and cross-entropy loss is calculated against ground truth labels.

3.2 Learning Interpretable Features via Patch-Level SAEs

Motivated by the probing results, we employ SAEs to decompose the VE's representations into a sparse dictionary of potentially interpretable features (SDFs) that capture underlying visual concepts. Our initial approach involved training an SAE on the same average-pooled VE activations used for linear probing (i.e., computing the mean over the patch dimension before SAE training). However, recognizing the potential information loss from this early aggregation step, and inspired by the PatchSAE methodology developed for CLIP [9], we revised our approach. We now train an overcomplete SAE directly on the individual patch activation vectors (`(num_samples * num_patches, hidden_size)`) from the final layer of Florence-2's VE [1], using the FairFace training set [10]. Sparsity is enforced via an L1 penalty on the SAE's hidden layer activations during training. This patch-level approach, consistent with [9], avoids premature information compression, crucially enables spatial analysis of learned features by examining activations on a per-patch basis, and allows for more flexible downstream analysis by deriving image-level representations from patch features (e.g., via mean or max aggregation) rather than being limited to pre-averaged inputs.

3.3 Identifying Candidate Race-Correlated SDFs

Subsequently, we systematically filter the extensive dictionary of learned SDFs (potentially tens of thousands) to identify a smaller, more manageable subset that exhibits the strongest statistical correlation with specific racial categories. This identification process utilizes a multi-stage filtering pipeline applied to SDF activation patterns computed across the FairFace validation set [10]. For

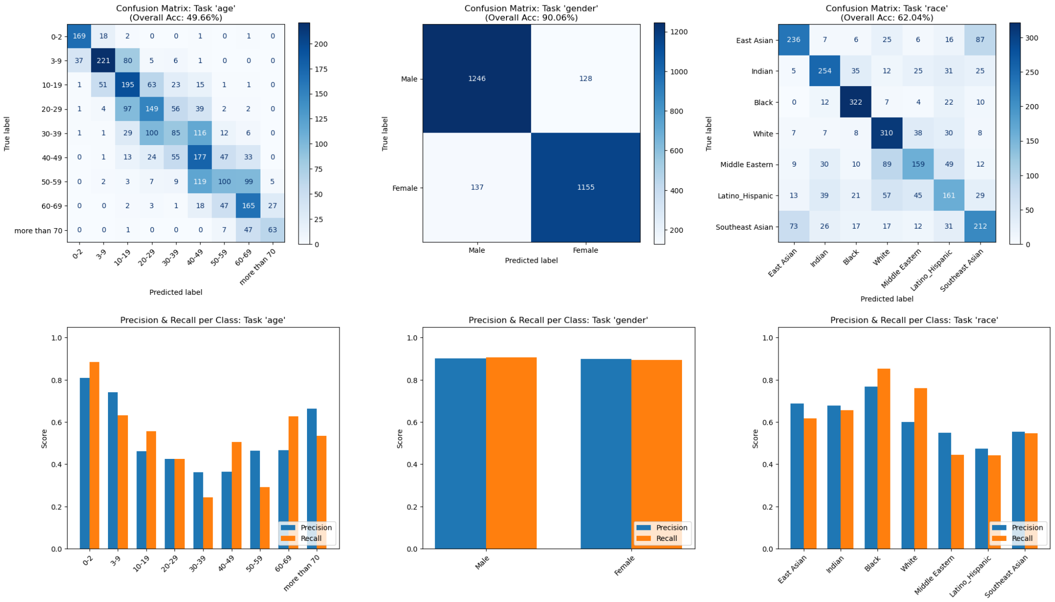


Figure 2: Linear probing results. The figure displays confusion matrices and per-class precision/recall bar plots for age, gender, and race classification tasks, evaluated using linear probes.

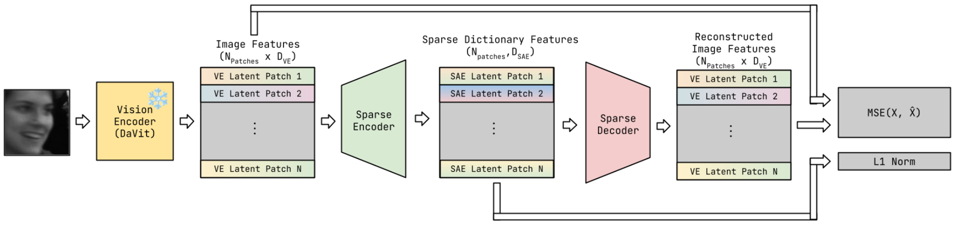


Figure 3: Patch-level Sparse Autoencoder (SAE) pipeline. This diagram illustrates an image input to a Vision Encoder, followed by the extraction of patch features. These features are then fed into a Sparse Encoder to produce SDF (Sparse Deep Features) patch features. Subsequently, the SDFs are processed by a Sparse Decoder to reconstruct the patch features. The pipeline calculates Mean Squared Error (MSE) loss between the original and reconstructed features, and an L_1 norm for promoting sparsity.

this analysis, SDF activations are initially aggregated to the image level by taking the mean across patches to associate a single activation value per SDF with each image's race label.

Intra-group Activation Frequency

The rationale behind this first step is to identify SDFs that are commonly activated when processing images belonging to a particular race group. For each SDF and each race category, we calculate the proportion of images within that category for which the SDF's aggregated activation surpasses a minimal threshold ($\tau > 0$). This yields a frequency score indicating how often the feature "fires" for that group. We select the top k_1 SDFs exhibiting the highest activation frequency *within each specific race group*, forming an initial pool of candidates based on prevalence.

Intra-group Mean Activation

Simply being frequently active might not be sufficient; we are also interested in features that activate *strongly* when present, suggesting higher salience or confidence according to the SAE. For each race group, we consider its k_1 candidate SDFs identified in the previous step. We then calculate the mean activation value for each of these SDFs, averaging *only* across those images within the group where the SDF was active (activation $> \tau$). By retaining the top k_2 SDFs per group ranked by this mean activation value, we prioritize features that are not only common but also exhibit strong activation signals within that group.

Inter-group Specificity with Label Entropy

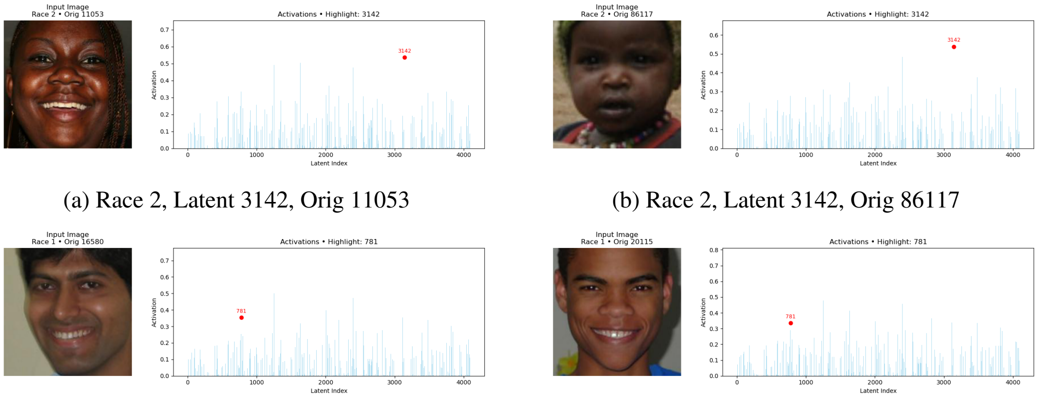
A feature might be frequent and strong within one group but also activate substantially for other groups. To isolate features more uniquely associated with a target group, we calculate the label entropy for each of the remaining k_2 candidate SDFs per group, following [9]. This involves summing the SDF's activations across all images within each race category, normalizing these sums to form a probability distribution (p_c) representing the SDF's activation across the different race categories (C), and then computing the Shannon entropy: $entropy = -\sum_{c \in C}(p_c \log p_c)$. A low entropy value signifies that the SDF's activation is concentrated within one or a few race groups, indicating higher specificity. Conversely, high entropy suggests the feature is broadly active across many groups. We retain the top k_3 SDFs *per group* that demonstrate the *lowest* label entropy, yielding our final set of candidate SDFs considered most statistically relevant and specific to each racial category based on their activation patterns. This multi-stage process systematically narrows down the vast SDF dictionary to focused lists of candidates for subsequent qualitative interpretation.

3.4 Visual Interpretation of Candidate SDFs

To qualitatively understand the visual semantics captured by the candidate SDFs identified as potentially race-correlated, we follow standard SAE interpretation techniques [9]. We visualize the top-k input image patches that maximally activate each candidate SDF. This analysis, aided by the spatial context inherent in the patch-level approach, helps associate abstract SDFs with concrete visual patterns, such as specific skin tones, hair features, facial structures, accessories, or background elements.

3.5 Findings

Several of our hypothesized race-associated latents demonstrate strong alignment with the intended demographic group. Notably, certain latents such as 3142 consistently activate on images labeled with the target race more than 80% of the time across the top 1000 activations, substantially above random chance. This suggests a robust and specific relationship between these latents and demographic attributes. In contrast, other latents (e.g., 781) exhibit much weaker alignment, with target race falling below 25%, indicating either noisier representations or potential entanglement with unrelated features.



(a) Race 2, Latent 3142, Orig 11053

(b) Race 2, Latent 3142, Orig 86117

(c) Race 1, Latent 781, Orig 16580

(d) Race 1, Latent 781, Orig 20115

Figure 4: Top-activating examples for two representative latents. Latent 3142 shows strong race alignment with Race 2, while latent 781 exhibits less consistent activation patterns across Race 1 individuals.

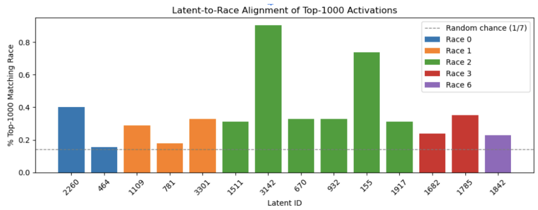
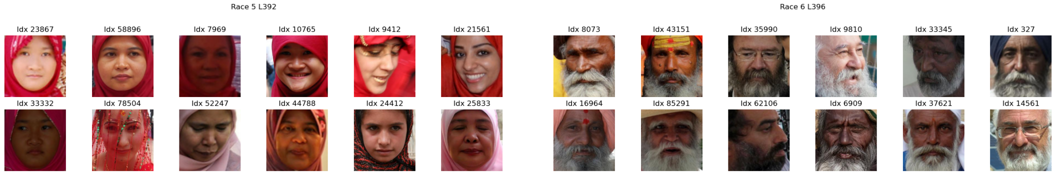


Figure 5: Percent of images correctly classified when the hypothesized latent is signaling strongly.

Further qualitative analysis of the Sparse Dictionary Features (SDFs) reveals that these latents often encode cultural or contextual markers rather than solely inherent facial characteristics. For example, one SDF found to be strongly associated with individuals labeled as ’Middle Eastern’ in the FairFace dataset predominantly activates on images of women who appear to be wearing headscarves. This observation underscores a tendency for AI systems to utilize visually distinctive proxy features when representing complex, socially-constructed categories like race. This mechanism—the reliance on salient cultural or contextual signifiers as computational shortcuts—demonstrates a key pathway through which algorithmic bias can manifest. The Vision Language Model (VLM), through its extensive pre-training, likely internalizes statistical correlations between various visual elements present in its training data, some of which serve as these proxies.



(a) Visual patterns including individuals who appear to be women wearing headscarves, strongly activate latent feature 392.

(b) Visual patterns including individuals who appear to be men with beards, strongly activating latent feature 396.

Figure 6: Top activating images for a latent feature identified by the Sparse Autoencoder as associated with varying races. These examples from the FairFace dataset highlight visual patterns, such as attire or facial features, that consistently activate this specific latent feature, revealing what the model has encoded.

Our subsequent analysis, correlating SDF activations with demographic labels, then brings to light how these learned shortcuts can lead to the reproduction of existing social categorizations. While leveraging such readily detectable visual cues may be computationally efficient, this approach risks reducing multifaceted identities to stereotypical visual markers, which are often neither universally present within nor exclusively confined to any single demographic group. These findings highlight the critical importance of meticulously scrutinizing the actual features AI systems learn and employ for encoding social categories, as these internal representations may inadvertently reflect and potentially amplify prevailing societal biases rather than capturing intrinsic or defining group characteristics.

3.6 Limitations & Further Directions

Our current method of using unsupervised Sparse Autoencoders (SAEs) to understand how this AI model represents race has shown promise, but also reveals limitations. While some identified internal features reliably capture race-specific visual information, others struggle to separate these demographic details from unrelated visual information. This highlights a clear need for further research.

One key direction is to refine our interpretability techniques. For instance, developing semi-supervised SAEs (SSSAEs), which can be guided with minimal supervision, could help isolate features related to predefined concepts like race more consistently and reliably. This would provide a clearer and more robust understanding of the model’s internal workings and how specific features might influence its behavior.

Beyond improving how we identify these features, a critical next step is to investigate the tangible impact of these race-correlated internal features on the AI's performance in real-world applications and its overall fairness. This involves quantitatively measuring whether these features lead to performance differences across various racial groups or contribute to biased outcomes, such as generating stereotypes in image captions or search results. Understanding this link between internal model features and observable downstream effects is crucial for grasping the real-world relevance of our findings. Additionally, it's important to explore potential confounding factors. These include the influence of the AI's original, vast training dataset (which might contain its own biases) and other image characteristics (like lighting or pose) that could inadvertently correlate with race. Broadening this analysis to different AI models, other demographic attributes like age and gender, and more diverse datasets will also be essential to determine how widely these encoding patterns apply.

Ultimately, the aim is to leverage this deeper understanding to build fairer and more transparent AI systems. Future efforts should focus on establishing causal relationships—for example, determining if a specific internal feature directly causes a biased outcome, perhaps through targeted experiments like feature editing or removal. The insights gained from reliably identifying race-related features and understanding their impact should then directly inform the design and implementation of effective bias mitigation strategies. This could involve developing new training methods, techniques to adjust the model's internal representations, or improved data handling practices to reduce the AI's reliance on problematic features. By pursuing these integrated research directions, we hope to move from simply identifying correlations to truly understanding their causal effects and, ultimately, to developing more robust and equitable Vision Language Models.

4 Conclusion

This paper presented a methodology for investigating the encoding of demographic attributes within the vision encoder of the Florence-2-base Vision Language Model [1]. Recognizing the critical need for understanding and mitigating bias in multimodal AI systems, our approach leverages a combination of linear probing and Sparse Autoencoder (SAE) techniques, specifically drawing inspiration from patch-level SAE analysis [9]. We first confirmed the presence of linearly decodable race information using probes trained on the balanced FairFace dataset [10]. Subsequently, we detailed a patch-level SAE training procedure and a multi-stage filtering pipeline designed to identify interpretable SDFs statistically correlated with specific racial categories based on activation frequency, mean activation, and label entropy. The methodology culminates in qualitative interpretation of these candidate features via visualization of top-activating patches.

The identification of such deeply encoded demographic features is a critical step, as their unexamined presence carries tangible implications for fairness in downstream applications. For instance, a VLM's reliance on these specific race-correlated SDFs could inadvertently lead to performance disparities across demographic groups or perpetuate stereotypical outputs in tasks like image captioning, visual search, or human-computer interaction.

By providing a structured approach to dissecting representations with spatial awareness, this work aims to contribute to the mechanistic understanding of bias in VLMs. Crucially, this granular understanding of how and where demographic information is encoded can directly inform the development of more fair and transparent AI systems. For example, it lays the groundwork for future interventions such as targeted feature pruning or the design of regularization techniques to discourage reliance on sensitive attributes, thereby reducing their impact on downstream task behavior. Ultimately, the insights and methodologies presented herein form a basis for continued efforts toward developing more robustly interpretable and equitable AI systems.

References

[1] Yuan, L., et al. (2023). Florence-2: A New Foundation Model for Computer Vision. *arXiv preprint arXiv:2311.06242*.

[2] Gustafson, E., et al. (2024). FairMedFM: Fairness Benchmarking for Medical Imaging Foundation Models. *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track.*

[3] Wu, H. (2025). Membership Inference Attacks on Large-Scale Models: A Survey. *arXiv preprint arXiv:2503.19338.*

[4] Nanda, N., et al. (2023). Emergent linear representations in world models of self-supervised sequence models. *Advances in Neural Information Processing Systems (NeurIPS).*

[5] Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Chen, A., ... Amodei, D. (2022). A Mathematical Framework for Transformer Circuits. *Transformer Circuits*. Retrieved from `https://transformer-circuits.pub/2022/toy_model/index.html`

[6] Bricken, T., Templeton, A., Batson, J., et al. (2023). Towards monosemanticity: Decomposing language models with dictionary learning. *Anthropic*. Retrieved from `https://transformer-circuits.pub/2023/monosemantic-features/index.html`

[7] Cunningham, H., Ewart, A., Riggs, L., Huben, R., Sharkey, L. (2023). Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*. Retrieved from `https://transformer-circuits.pub/2023/monosemantic-features/index.html`

[8] Templeton, A., Ganguli, D., Henighan, T., et al. (2024). Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. *Anthropic*. Retrieved from `https://transformer-circuits.pub/2024/scaling-monosemanticity/`

[9] Lim, H., Choi, J., Choo, J., Schneider, S. (2024). Sparse autoencoders reveal selective remapping of visual concepts during adaptation. *arXiv preprint arXiv:2412.05276*.

[10] Karkkainen, K., Joo, J. (2021). FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV).*