Decomposing SAE-Based Debiasing: Reconstruction Error Dominates Targeted Suppression in Vision-Language Models

Ozan Bayiz
ozanbayiz@example.edu

February 17, 2026

Abstract

Sparse autoencoders (SAEs) are increasingly used to identify and suppress demographic features in vision encoders, with reported fairness improvements on classification and retrieval benchmarks. We show that on free-form generation tasks, these effects are confounded by a *reconstruction-error confound*: the SAE's encode-decode process itself—without modifying any features—accounts for a substantial, architecture-dependent portion of the observed demographic content reduction. Through a six-method decomposition (noise, SAE passthrough, random suppression, targeted SDF suppression, LEACE, and S&P Top-K projection) across three VLMs with distinct vision encoder architectures (PaliGemma2, Qwen2-VL, Qwen3-VL), we separate SAE-based intervention effects into generic perturbation ($\sim$4 pp), structured reconstruction error (0–16 pp, architecture-dependent), and genuine targeted suppression ($\sim$12 pp). The reconstruction component scales monotonically with SAE sparsity and varies dramatically across models: it dominates on PaliGemma2 (83% baseline gender mention rate), vanishes on Qwen3-VL, and produces *opposite-direction* effects on Qwen2-VL where suppression *increases* demographic content by converting neutral language into gendered hallucinations. We further show that LEACE, a linear concept erasure method that bypasses the SAE reconstruction path, reveals a VQA–captioning dissociation: VQA accuracy drops to chance (50%) under gender erasure while captioning is nearly unaffected, indicating that these tasks access demographic information through different mechanisms. All results are validated with LLM-as-judge evaluation (n=1,000, bootstrap CIs), which also exposes a 77–98% false-positive rate in keyword-based race metrics. Our decomposition framework provides a necessary diagnostic for any SAE-based intervention study.

1 Introduction

Sparse autoencoders (SAEs) have emerged as a promising tool for mechanistic interpretability of neural networks [Cunningham et al., 2024, Bricken et al., 2023], offering a way to decompose dense, polysemantic representations into sparse, human-interpretable features. A growing body of work applies SAEs to vision encoders in vision-language models (VLMs), identifying features that encode demographic attributes—such as gender, race, and age—and suppressing them to reduce biased outputs [Sasse et al., 2024, Pach et al., 2025, Joseph et al., 2025]. These approaches report substantial fairness improvements on classification and retrieval benchmarks, suggesting that SAE-based feature suppression is an effective debiasing tool.

We show that this conclusion requires substantial qualification. A dominant mechanism behind reported debiasing effects is not targeted feature suppression, but the *reconstruction error* imposed by the SAE's encode-decode process itself. Passing vision encoder activations through an SAE—*without modifying any features*—accounts for a large, architecture-dependent portion of demographic content reduction in generated text. This "passthrough" effect, which we term the **reconstruction-error confound**, arises because SAEs with finite dictionary size and top-k sparsity constraints preferentially lose demographic nuance during reconstruction, as these attributes occupy a small, linearly-structured subspace that is poorly preserved by sparse coding. The SAE acts as an unintentional demographic filter, systematically discarding fine-grained attribute information while preserving coarse semantic content.

Through a systematic series of experiments across three VLMs with distinct vision encoder architectures—PaliGemma2 (SigLIP) [Beyer et al., 2024], Qwen2-VL (custom ViT) [Wang et al., 2024], and Qwen3-VL (DeepStack ViT) [Bai et al., 2025]—we quantitatively decompose the full SAE intervention effect into three components:

1. **Generic perturbation** (∼4 pp): The effect of matched-magnitude Gaussian noise, which barely affects demographic content.

2. **Structured reconstruction error** (∼17 ppon PaliGemma2, ∼0 ppon Qwen3-VL): The additional effect of the SAE encode–decode pipeline beyond generic noise, which varies dramatically across architectures.

3. **Targeted feature suppression** (∼12 pp): The genuine incremental effect of zeroing demographic-aligned sparse dictionary features (SDFs), attributable to feature identification.

This decomposition reveals that prior work conflates components (2) and (3), inflating claims about the efficacy of targeted feature identification. The bottleneck component is not a property of the SAE alone, but depends critically on the interaction between the VLM's language model and the vision encoder: models that generate highly gendered language (like PaliGemma2, with 83% baseline gender mention rate) are far more susceptible to the bottleneck than models that default to neutral language (like Qwen2-VL, with 6% baseline). We further show that:

- The bottleneck effect **scales monotonically** with SAE reconstruction quality: reducing the top-k sparsity parameter from 64 to 16 increases gender content reduction from 21 ppto 52 pp, with no change to the features suppressed (Section 4.2).

- **LEACE** [Belrose et al., 2023], a linear concept erasure method that bypasses the SAE entirely, reveals a **VQA–captioning dissociation**: it reduces VQA gender accuracy to chance (50%) while leaving caption content nearly unchanged, indicating that these tasks access demographic information through different mechanisms (Section 4.4).

- The same SAE-based intervention can have **qualitatively opposite effects** across VLMs: suppressing gender-aligned features reduces gender content by 25 ppon Qwen3-VL but *increases* it by 24 ppon Qwen2-VL, because Qwen2-VL's neutral language generation is disrupted by SAE reconstruction (Section 4.3).

- We validate all results with an independent **LLM-as-judge** evaluation, finding that keyword-based demographic content metrics have 77–98% false-positive rates for race (due to polysemous colour terms) while achieving <3% error for gender (Section 4.5).

Our results have immediate implications for the SAE-for-debiasing research programme: reported improvements must be decomposed against passthrough and noise baselines before attributing them to feature-level interpretability. We release code and all experimental artefacts to enable this decomposition on future work.

2 Related Work

Sparse autoencoders for interpretability. SAEs decompose neural network activations into sparse, monosemantic features by training an overcomplete dictionary with a sparsity penalty [Cunningham et al., 2024, Bricken et al., 2023]. BatchTopK SAEs [Bussmann et al., 2024] replace the L_1 penalty with a deterministic top-k selection, offering better reconstruction–sparsity trade-offs. Recent work extends SAEs to vision transformers [Stevens et al., 2025], CLIP encoders [Joseph et al., 2025], and the vision towers of VLMs [Pach et al., 2025]. Matryoshka SAEs [Zaigrajew et al., 2025] learn hierarchical features at multiple granularities, achieving near-perfect reconstruction (0.99 cosine similarity on CLIP). SAEBench [Karvonen et al., 2025] provides a comprehensive evaluation suite for SAEs, revealing that proxy metrics (e.g., reconstruction quality) do not reliably predict practical effectiveness. Our work builds on this foundation but critically examines the confound introduced by SAE reconstruction error during downstream VLM generation.

SAE-based VLM debiasing. Sasse et al. [2024] train SAEs on CLIP-family vision encoders across five VLMs and suppress demographic features, reporting 5–15 point fairness improvements on classification, VQA, and captioning benchmarks. Bărbălau et al. [2025] propose Select & Project (S&P) Top-K, which uses the SAE encoder weights to define a concept direction for orthogonal projection, reporting up to 3.2× fairness gains over standard SAE suppression on CelebA and FairFace classification tasks. Neither work reports passthrough baselines that would isolate reconstruction error from targeted suppression. We show that when evaluated on free-form captioning with passthrough controls, S&P Top-K has negligible downstream impact ($<$0.5 pp), and standard SAE suppression effects are substantially inflated by reconstruction error on certain architectures.

Linear concept erasure. LEACE (Least-squares Concept Erasure; Belrose et al. 2023) removes concept information via an optimal linear projection that guarantees no linear probe can recover the erased attribute. SPLINCE [Holstege et al., 2025] extends this framework by preserving task-relevant covariance while removing concepts. Unlike SAE-based methods, these projection methods introduce no reconstruction bottleneck—the projection preserves $>$99.9% of the activation variance. We use LEACE as a bottleneck-free baseline to isolate genuinely targeted effects from the SAE reconstruction confound.

VLM fairness and debiasing. A growing literature addresses demographic bias in VLMs through training-time interventions [Zhang et al., 2025], test-time representation editing [Gerych et al., 2024], feature pruning [Jung et al., 2024], and contrastive projection [Molahasani et al., 2025]. These methods typically evaluate on benchmark classification tasks. Our contribution is orthogonal: we provide a diagnostic framework for *any* intervention that modifies vision encoder activations, decomposing its effect into bottleneck and targeted components.

Evaluation of demographic content. Keyword-based demographic content rate (DCR) metrics count occurrences of demographic terms in generated text [Zhao et al., 2017]. We show this approach has severe limitations for racial attributes due to polysemous colour terms, and complement it with LLM-as-judge evaluation using both binary classification (Qwen3-8B; Yang et al. 2025) and pairwise comparison (JudgeLRM-7B; Chen et al. 2025).

Table 1: Vision-language models used in experiments.

VLM	VE Architecture	VE dim	Tokens	SAE dim	Recon. cos sim
PaliGemma2 3B	SigLIP ViT	1152	1024	4608	0.945
Qwen2-VL 2B	Custom ViT	1536	64	6144	0.959
Qwen3-VL 2B	DeepStack ViT	2048	196	8192	0.975

3 Method

We study how demographic information flows from vision encoder (VE) representations through vision-language models to generated text. Our framework consists of four stages: (1) extract VE latents, (2) train SAEs and identify demographic features, (3) intervene on VE activations via forward hooks during VLM inference, and (4) evaluate the downstream effect on generated text.

3.1 Vision Encoder Latent Extraction

For each VLM, we extract vision encoder activations from the FairFace dataset [Kärkkäinen and Joo, 2021], which contains ∼87K face images annotated for age (9 classes), gender (2 classes), and race (7 classes). We process each image through the VLM's vision encoder and store the resulting token-level activations $\mathbf{H} \in \mathbb{R}^{T \times d}$, where T is the number of visual tokens and d is the hidden dimension. Table 1 summarises the three VLMs and their vision encoder architectures.

3.2 SAE Training and Demographic Feature Discovery

We train BatchTopK SAEs [Bussmann et al., 2024] on the extracted VE latents. Each SAE maps a d-dimensional input to a $4d$-dimensional sparse code via:

$$\mathbf{z} = \mathrm{TopK}\big(\mathrm{ReLU}(\mathbf{W}_{\mathrm{enc}}\,\mathbf{x} + \mathbf{b}_{\mathrm{enc}}),\ k\big), \qquad \hat{\mathbf{x}} = \mathbf{W}_{\mathrm{dec}}\,\mathbf{z} + \mathbf{b}_{\mathrm{dec}}, \tag{1}$$

where $\mathrm{TopK}(\cdot, k)$ retains only the k largest activations and zeros the rest (k=64 for all models).

To identify **sparse dictionary features** (SDFs) aligned with demographic attributes, we use a three-stage filtering pipeline:

1. **Frequency filtering**: Select features that activate in >10% of samples from at least one demographic class.
2. **Mean-activation filtering**: Rank features by the ratio of their highest class-conditional mean activation to their lowest, retaining the top 50 per class.
3. **Entropy filtering**: Among candidates, select features with high inter-class variance (low entropy), prioritising features that discriminate between demographic groups.

This yields 50–80 unique SDFs per attribute per model (e.g., 56 gender SDFs for PaliGemma2, 59 for Qwen3-VL).

3.3 Intervention Methods

All interventions are applied via forward hooks on the VLM's vision encoder output during inference, modifying activations before they reach the language model. We compare six intervention methods, ordered by their degree of perturbation:

1. Noise perturbation. We add Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ directly to VE activations, where σ is calibrated to match the magnitude of SAE reconstruction error. This isolates the effect of generic perturbation at the same scale.

2. SAE passthrough. Activations are encoded through the SAE and decoded *without modifying any features*: $\hat{\mathbf{x}} = \mathrm{Dec}(\mathrm{Enc}(\mathbf{x}))$. This isolates the structured information bottleneck imposed by SAE reconstruction.

3. Random feature suppression. Same as passthrough, but m randomly selected non-SDF features are zeroed after encoding. This controls for the effect of zeroing *any* features, regardless of their demographic alignment.

4. Targeted SDF suppression. The m identified SDFs for a target attribute are zeroed after SAE encoding: $z_i = 0$ for all $i \in \mathcal{S}_{\mathrm{attr}}$. This is the standard approach in prior work [Sasse et al., 2024].

5. LEACE erasure. We fit a Least-squares Concept Erasure projector [Belrose et al., 2023] on mean-pooled VE latents and their demographic labels, then apply the projection to each visual token during inference. LEACE removes the complete linear concept subspace without introducing reconstruction error.

6. S&P Top-K projection. Following Bărbălau et al. [2025], we compute a rank-1 concept direction from the SAE encoder weights of the top-k SDFs and orthogonally project VE activations to remove this direction.

3.4 Evaluation

We evaluate interventions on two downstream tasks:

Image captioning. We prompt the VLM with "Describe this image" and measure demographic content rate (DCR)—the fraction of captions that mention a given attribute—using keyword matching. We report DCR delta ($\Delta\mathrm{DCR} = \mathrm{DCR}_{\mathrm{modified}} - \mathrm{DCR}_{\mathrm{original}}$) on the same images with and without intervention. To address keyword limitations, we independently evaluate with an LLM-as-judge (Section 4.5).

Visual question answering (VQA). We ask direct demographic questions (e.g., "What is the gender of the person in this image?") and measure accuracy change under intervention. This provides a complementary metric that is robust to captioning style differences.

We additionally report BERTScore [Zhang et al., 2020] F1 between original and modified captions as a measure of overall text quality preservation, and compute 95% bootstrap confidence intervals (10,000 resamples) for all effect sizes.

4 Experiments

All experiments use FairFace images with demographic labels. Unless otherwise noted, results report n=1,000 samples with 95% bootstrap confidence intervals. We focus on gender as the primary

Table 2: **Bottleneck decomposition on PaliGemma2.** Gender DCR delta (percentage points) for each intervention method on image captioning (n=1,000). The SAE passthrough (encode–decode without modification) accounts for the majority of the effect. All 95% CIs from 10,000 bootstrap resamples.

Intervention	**Gender ΔDCR**	**95% CI**	**BERTScore**	**Incremental**
No intervention (baseline)	—	—	—	—
Noise (calibrated)	$-4.0\,pp$	[−6.0, −2.0]	0.978	$-4.0\,pp$
SAE passthrough	$-16.1\,pp$	[−18.4, −12.6]	0.933	$-12.1\,pp$
Random suppress (m=56)	$-16.1\,pp$	[−18.4, −12.6]	0.933	$0.0\,pp$
SDF suppress (m=56)	**−28.2 pp**	[−31.1, −24.8]	0.931	**−12.1 pp**
LEACE (gender)	$-0.8\,pp$	[−2.2, +0.8]	0.974	—
S&P Top-K (k=64)	$-0.4\,pp$	—	0.998	—

attribute because gender keyword DCR is well-calibrated (1–18% false-positive rate across VLMs; Section 4.5).

4.1 The Bottleneck Decomposition

Table 2 presents the core result on PaliGemma2. The full SDF suppression effect ($-28.2\,pp$ gender DCR) decomposes cleanly:

- **Generic perturbation (14%)**: Matched-magnitude Gaussian noise reduces gender DCR by only $4.0\,pp$, with BERTScore 0.978 indicating minimal text disruption. Generic noise at the scale of SAE reconstruction error is insufficient to explain the observed effects.

- **Structured bottleneck (43%)**: SAE passthrough—encoding and decoding activations without any feature modification—reduces gender DCR by $16.1\,pp$ (CI: $[-18.4, -12.6]$). This is entirely reconstruction error, yet accounts for nearly half the "suppression" effect. Critically, random feature suppression produces the *identical* effect ($p = 1.0$, paired permutation test), confirming that which features are zeroed is irrelevant at this level.

- **Targeted suppression (43%)**: Suppressing the 56 identified gender SDFs adds a further $12.1\,pp$ reduction beyond passthrough (CI: $[-15.0, -9.2]$, $p < 0.001$ vs. passthrough). This is the genuine interpretability-driven component—it requires identifying the correct features. It is statistically significant, but accounts for less than half of the total effect reported by standard SAE suppression evaluations.

LEACE and S&P Top-K, which do not pass through the SAE bottleneck, produce negligible caption DCR changes ($< 1\,pp$), confirming that the large effects in the SAE conditions are primarily bottleneck-driven.

4.2 SAE Quality Ablation

If the bottleneck hypothesis is correct, worsening SAE reconstruction should increase the bottleneck effect. We test this by varying the effective k parameter at inference time (reducing sparsity $\Rightarrow$ worse reconstruction).

Table 3 confirms the prediction: halving k from 64 to 32 nearly doubles the gender DCR drop ($16.1 \rightarrow 31.8\,pp$), and quartering it to 16 more than triples it ($52.4\,pp$). All pairwise differences

Table 3: **Bottleneck effect scales with reconstruction quality (PaliGemma2).** Gender DCR delta under passthrough (no feature modification) with varying SAE sparsity.

Effective k	**Gender ΔDCR**	**95% CI**	**BERTScore**	**vs. k=64**
64 (default)	$-16.1\,pp$	[−18.4, −12.6]	0.933	—
32	$-31.8\,pp$	[−36.6, −27.2]	0.918	$p < 0.0001$
16	$-52.4\,pp$	[−57.4, −47.4]	0.904	$p < 0.0001$

Table 4: **Reconstruction quality under varying dictionary size (PaliGemma2, mean-pooled).** All SAEs trained for 50 epochs with k=64. Quadrupling dictionary capacity yields negligible cosine similarity improvement; nearly all additional features remain dead.

Dict size	**Cosine sim**	**Val loss**	**Alive features**	**Dead (%)**
$4d$ (4,608)	0.998	0.0098	577	87.5%
$8d$ (9,216)	0.998	0.0099	559	93.9%
$16d$ (18,432)	0.998	0.0095	583	96.8%

are significant ($p < 0.0001$). Importantly, BERTScore degrades only moderately (0.933 → 0.904), meaning the VLM still generates coherent English—it simply contains far less gendered language. This confirms that the SAE's sparse reconstruction preferentially loses demographic information, constituting a structured information bottleneck rather than generic degradation.

Dictionary size ablation. The sparsity ablation above varies k at fixed dictionary size ($4d$). A natural question is whether increasing dictionary capacity could reduce the reconstruction error and hence the bottleneck effect. We train SAEs at 4×, 8×, and 16× dictionary size on mean-pooled PaliGemma2 VE features (Table 4), holding k=64 fixed.

The result is unambiguous: quadrupling dictionary capacity from $4d$ to $16d$ improves cosine similarity by only 5×10^{-5} (0.99823 → 0.99828). The number of alive features saturates at ∼560–580 regardless of dictionary size, with >87% of features remaining dead. This confirms that the bottleneck is fundamentally *sparsity-limited* (k=64), not *capacity-limited*: the top-k constraint restricts reconstruction to a 64-dimensional subspace per token regardless of how many candidate features are available. The additional features in the $8d$ and $16d$ dictionaries simply fail to activate, contributing nothing to reconstruction quality. This result explains why the k-ablation (Table 3) produces dramatic effects while dictionary size does not, and suggests that higher-fidelity SAE architectures (e.g., Matryoshka SAEs [Zaigrajew et al., 2025]) that modify the sparsity mechanism rather than increasing capacity may be needed to mitigate the reconstruction-error confound.

4.3 Cross-Model Validation

Table 5 reveals that the bottleneck decomposition is fundamentally **architecture-dependent**:

PaliGemma2 (SigLIP VE). High baseline gender DCR (83%). The bottleneck dominates: passthrough alone reduces gender content by $16.1\,pp$. SDF suppression adds $12.1\,pp$ beyond the bottleneck.

Qwen2-VL (custom ViT). Very low baseline gender DCR (6%). The model defaults to neutral language ("a person" instead of "a man/woman"). Passthrough has *zero* effect (CI spans zero).

Table 5: **Cross-model comparison of gender captioning interventions.** Gender DCR delta (pp) with LLM-judge validation where available. The bottleneck decomposition is architecture-dependent: dominant on PaliGemma2, absent on Qwen3-VL, and reversed on Qwen2-VL.

		VLM		
Method	**Metric**	PaliGemma2	Qwen2-VL	Qwen3-VL
Passthrough	KW ΔDCR	−16.1	−0.6	+0.4
	LLM ΔDCR	−15.5	—	+3.2
SDF suppress	KW ΔDCR	−28.2	+23.8	−21.2
	LLM ΔDCR	−27.9	—	−12.9
LEACE	KW ΔDCR	−0.8	−4.2	−12.9
	LLM ΔDCR	−0.7	—	−12.9
Baseline gender DCR		83%	6%	88%

Strikingly, SDF suppression **increases** gender content by 23.8 *pp*—the opposite direction. Of 500 samples under suppression, 127 (25.4%) *gained* gender-specific language while only 8 (1.6%) lost it (Appendix B.4). The dominant pattern is that neutral descriptions ("a person," "the child") are replaced by gendered hallucinations ("a man," "a boy"), sometimes assigning the *incorrect* gender. This is consistent with SAE reconstruction disrupting the visual features that support the model's neutral-language strategy, causing fallback to more specific (and often inaccurate) descriptions. This opposite-direction effect is confirmed by LLM-judge evaluation (Section 4.5).

Qwen3-VL (DeepStack ViT). High baseline gender DCR (88%). The bottleneck is *absent*: passthrough changes gender DCR by only +0.4 *pp* (LLM: +3.2 *pp*, CI: $[+0.9, +5.5]$). SDF suppression provides a genuine −21.2 *pp* reduction (LLM: −12.9 *pp*, CI: $[-16.5, -9.3]$), entirely attributable to targeted feature removal. LEACE is also effective (−12.9 *pp*). The Qwen3-VL SAE achieves the highest reconstruction quality among our models (cosine similarity 0.975 vs. 0.945 for PaliGemma2), consistent with the hypothesis that better reconstruction reduces the bottleneck. This is the cleanest demonstration that SAE-based intervention *can* work when reconstruction quality is sufficient.

Implications. The cross-model comparison demonstrates that the bottleneck illusion is not a universal property of SAEs but emerges from the interaction between reconstruction error and the VLM's baseline captioning behaviour. The reconstruction quality ordering (Qwen3-VL: 0.975 > Qwen2-VL: 0.959 > PaliGemma2: 0.945) is consistent with the bottleneck severity ordering (PaliGemma2: dominant, Qwen2-VL: absent, Qwen3-VL: absent), though baseline gender DCR (83%, 6%, 88%) also varies across models. Disentangling the contributions of reconstruction quality and language model priors requires controlled experiments that we leave to future work.

4.4 LEACE: Targeted Concept Erasure Without the Bottleneck

LEACE provides a bottleneck-free comparison method (Table 6). On PaliGemma2, LEACE gender erasure drops VQA accuracy to exactly 50%—chance for binary classification—while producing negligible caption DCR change (−0.8 *pp*, CI spans zero). The high BERTScore (0.974) and cosine similarity (0.9991) confirm that LEACE introduces minimal perturbation to the overall representation.

Table 6: **LEACE reveals dual processing pathways.** VQA accuracy delta and caption DCR delta under LEACE gender erasure across three VLMs.

VLM	VQA Acc. Δ	Caption Gender ΔDCR	BERTScore	Cos sim
PaliGemma2	$-9.0\,pp$ (59% → 50%)	$-0.8\,pp$	0.974	0.9991
Qwen2-VL	$-17.5\,pp$	$-4.2\,pp$	—	—
Qwen3-VL	$-11.0\,pp$	$-12.9\,pp$	0.957	—

Table 7: **Keyword-based race DCR has 77–98% false-positive rates.** Baseline DCR (original captions, no intervention) measured by keyword matching vs. LLM-as-judge binary classification (Qwen3-8B).

VLM	Attribute	KW DCR	LLM DCR	FP Rate
PaliGemma2	Race	11.4%	0.2%	98.2%
PaliGemma2	Gender	83.0%	82.2%	1.0%
Qwen2-VL	Race	21.6%	5.0%	76.9%
Qwen2-VL	Gender	6.0%	5.8%	3.3%
Qwen3-VL	Race	45.0%	9.5%	78.9%
Qwen3-VL	Gender	88.0%	72.0%	18.2%

This VQA–captioning dissociation admits several interpretations: (1) VQA reads gender from a compact linear subspace fully removable by LEACE, while captioning synthesises gender from nonlinear or distributed features resistant to linear erasure; (2) the mean-pooled LEACE projection incompletely erases token-level gender information that captioning relies on (see Appendix A); or (3) the language model's internal priors compensate for erased visual gender cues during free-form generation but not during directed VQA. The PaliGemma2 model's high baseline gender DCR (83%) is consistent with strong language-side priors that could sustain gendered captions even without visual input. We cannot fully disambiguate these hypotheses with the current experiments, though the near-unity cosine similarity under LEACE argues against explanation (2) as the dominant factor.

On Qwen3-VL, LEACE is more effective on captions ($-12.9\,pp$), suggesting that gender encoding in the DeepStack ViT has a stronger linear component accessible to captioning, or that the Qwen3-VL language model has weaker compensatory priors. The cross-model variation in LEACE effectiveness provides complementary evidence about the interaction between VE representation geometry and language model generation strategy.

4.5 LLM-as-Judge Validation

A critical methodological finding concerns the validity of keyword-based DCR for different attributes (Table 7). We evaluate all captions with two independent LLM judges: Qwen3-8B [Yang et al., 2025] for binary classification ("Does this caption mention [attribute]?") and JudgeLRM-7B [Chen et al., 2025] for pairwise comparison ("Which caption contains more [attribute] content?").

Race keyword DCR is unreliable. Polysemous terms like "white" (clothing), "black" (hair), "light" (lighting), and "dark" (shadows) generate false positives at 77–98% rates across all three VLMs. The LLM judges confirm that VLMs almost never mention race in free-form captions (PaliGemma2: 0.2%, Qwen2-VL: 5.0%). This means all keyword-based race captioning results in the literature—including ours—are measuring noise in false-positive rates, not genuine changes in racial content.

Table 8: **VQA accuracy delta under intervention** (gender attribute, 200 samples per condition). VQA accuracy measures direct demographic identification, complementing the indirect caption DCR metric.

Method	PaliGemma2	Qwen2-VL	Qwen3-VL
Passthrough	$+4.5\,pp$	$-1.0\,pp$	$0.0\,pp$
SDF suppress	$+4.0\,pp$	$-21.5\,pp$	$-10.0\,pp$
LEACE	$-9.0\,pp$	$-17.5\,pp$	$-11.0\,pp$

Gender keyword DCR is well-calibrated. For gender, keyword and LLM-based metrics agree within 1–3 ppacross all VLMs, with false-positive rates of 1–18%. This validates our use of keyword DCR for gender throughout the paper.

LLM judge confirms all main effects. For the n=1,000 experiments on PaliGemma2 and Qwen3-VL (Table 5), LLM-based gender DCR deltas differ from keyword-based deltas by $< 1.5\,pp$ on PaliGemma2 and $< 8.3\,pp$ on Qwen3-VL (where the 18% keyword FP rate inflates keyword-based estimates). All confidence intervals remain significant and all directional conclusions are unchanged.

4.6 VQA Provides Complementary Evidence

VQA results (Table 8, n=200) reveal a consistent pattern: LEACE reduces VQA accuracy substantially on all three VLMs (-9 to $-17.5\,pp$), confirming that it removes the linear gender subspace used for direct classification. SDF suppression shows variable effects: negligible on PaliGemma2 ($+4.0\,pp$), strong on Qwen2-VL ($-21.5\,pp$) and Qwen3-VL ($-10.0\,pp$). This contrasts with the captioning results, where SDF suppression has the largest effect on PaliGemma2. The divergence provides evidence that the SAE bottleneck affects free-form captioning (which synthesises gender from many features) more than VQA (which reads a compact linear subspace). We note that the VQA sample size (n=200) is smaller than the captioning experiments (n=1,000), and these results should be interpreted accordingly; future work should validate with larger samples and bootstrap CIs.

5 Discussion

Implications for SAE-based debiasing. Our decomposition reveals that published SAE-based debiasing results [Sasse et al., 2024, Pach et al., 2025] may conflate reconstruction error with targeted feature effects when the SAE bottleneck is not controlled for. On PaliGemma2, passthrough accounts for 57% of the total SDF suppression effect; on Qwen2-VL, the same intervention produces a $24\,pp$ *increase* in demographic content. These results do not invalidate prior findings on classification benchmarks, where reconstruction error may affect the measured attribute differently, but they demonstrate that free-form generation tasks require passthrough baselines. We recommend that future work report passthrough and random suppression baselines alongside any SAE-based intervention result.

The structured bottleneck is not generic noise. An important nuance: the SAE bottleneck is not random degradation. Matched-magnitude Gaussian noise produces only $4\,pp$ gender DCR change, while the SAE bottleneck produces $16\,pp$. The SAE's sparse coding constraint systematically under-represents the demographic subspace because demographic attributes occupy a low-rank, linearly-encoded portion of the representation space—exactly the kind of information that top-k sparsity tends to discard. This is consistent with our finding that LEACE (which removes a rank-1 gender subspace) drops VQA accuracy to chance: gender is concentrated in a small subspace that sparse reconstruction inadvertently disrupts.

In this sense, the SAE acts as an *unintentional demographic filter*: it preferentially preserves coarse semantic content (object identity, scene layout) while losing fine-grained demographic attributes during sparse reconstruction. This filtering is a property of the interaction between the representation's geometry and the sparsity constraint, not a design choice. Whether this unintentional filtering should be considered a form of "debiasing" is partly a question of framing: it reduces demographic content, but without the interpretability guarantees that motivate SAE-based approaches. Our contribution is to quantify this component separately from the targeted component that does rely on feature identification.

Scope: output-layer vs. internal SAEs. Our experiments apply SAEs to the vision encoder's *output* (the modal interface between VE and LM). Most mechanistic interpretability work applies SAEs to internal residual streams of transformers (e.g., intermediate ViT layers or LM residual streams). The bottleneck confound may be maximised at the VE output, where information density is highest and the LM is most sensitive to perturbation. Internal-layer SAEs modify representations that are further processed by subsequent transformer layers, which may partially compensate for reconstruction error—analogous to how residual connections help error correction in deep networks. We leave systematic internal-layer ablation to future work, noting that it requires not only retraining SAEs at each depth but also a hook mechanism that intercepts, reconstructs, and reinjects activations *within* the VE forward pass, rather than at the modal interface. Future work should also examine whether the bottleneck confound extends to SAEs applied within the language model's residual stream, where the error correction capacity of subsequent LM layers may differ from that of VE layers.

Architecture-dependent bottleneck: a feature, not a bug? The cross-model comparison suggests a practical insight: VLMs whose language models default to neutral language (like Qwen2-VL) are naturally robust to SAE reconstruction artifacts, while VLMs with strong demographic language priors (like PaliGemma2) are highly susceptible. If the goal is VLM debiasing, the bottleneck effect—while confounding for interpretability research—may itself be a useful mechanism. A carefully calibrated SAE bottleneck could serve as a simple, non-targeted debiasing tool, though at the cost of some general caption quality (BERTScore ∼0.93).

LEACE as a diagnostic tool. LEACE's contrasting behaviour across tasks (VQA to chance, captioning nearly unchanged on PaliGemma2) provides a diagnostic for how demographic information is accessed in different generation modes. When LEACE is effective on captioning (as on Qwen3-VL), it indicates a stronger linear component in the VE's demographic encoding that captioning depends on. When it is ineffective on captioning but effective on VQA (as on PaliGemma2), the information used for captioning is either encoded nonlinearly, distributed across token positions in a way the mean-pooled projection does not fully capture, or compensated for by language model priors. This VQA–captioning dissociation was not previously documented and motivates future work to disentangle VE-side encoding geometry from LM-side generation strategy.

Limitations. Our study has several limitations. (1) We evaluate only three VLMs, all in the 2–3B parameter range; larger models may exhibit different bottleneck characteristics. (2) FairFace consists of cropped, aligned face images where the face dominates the visual field; this represents the easiest case for vision encoder reconstruction. In complex scenes (e.g., COCO, Visual Genome) where faces are small or occluded, the bottleneck may manifest as hallucination rather than debiasing, and the interaction between reconstruction error and demographic content could differ substantially. (3) Our keyword DCR metric, while validated by LLM judges for gender, is unreliable for race; we address this limitation but cannot fully resolve it within the current evaluation framework. (4) Sample sizes, while larger than some prior work (n=1,000 for captioning, n=200 for VQA), may be insufficient to detect small interaction effects, and VQA results lack bootstrap CIs. (5) We do not evaluate on downstream fairness benchmarks (e.g., bias amplification scores), focusing instead on the mechanistic decomposition. (6) The cross-model comparison confounds SAE reconstruction quality with VLM architecture; the bottleneck may be absent on Qwen3-VL partly because its SAE has higher cosine similarity (0.975 vs. 0.945 for PaliGemma2), not solely due to architectural properties. Ablating reconstruction quality on a single model (e.g., via dictionary size variation or higher-fidelity SAE architectures such as Matryoshka SAEs [Zaigrajew et al., 2025]) would isolate this confound. (7) The LEACE projection is fit on mean-pooled features but applied per-token, which may incompletely erase spatially-varying concept information (see Appendix A).

6 Conclusion

We have shown that SAE-based interventions in vision-language models are subject to a substantial *reconstruction-error confound*: the SAE's encode-decode process itself, independent of any feature modification, accounts for an architecture-dependent portion of reported debiasing effects. Through a systematic six-method decomposition across three VLMs, we separate the full intervention effect into generic noise (∼4 pp), structured reconstruction error (0–16 ppdepending on architecture), and targeted suppression (∼12 pp). The reconstruction component scales monotonically with SAE quality, confirming it is a property of sparse coding rather than feature identification. On Qwen2-VL, the same intervention produces the opposite effect entirely, demonstrating that architecture-dependent language generation strategies can amplify or invert reconstruction artifacts (Appendix B.4).

Our results do not invalidate SAE-based interpretability—targeted SDF suppression produces statistically significant effects beyond the reconstruction baseline on all tested models. However, they demonstrate that the magnitude of these effects has been substantially overstated in work that lacks passthrough controls. We recommend that future studies report passthrough and noise baselines, use LLM-based evaluation for racial content (where keyword metrics have >77% false-positive rates), and validate across VLMs with different language generation strategies. LEACE provides a complementary, reconstruction-free baseline that reveals the linear structure of demographic encoding and the dissociation between VQA and captioning information pathways.

References

Shuai Bai et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Antonio Bărbălau, Cristian Daniel Păduraru, Teodor Poncu, Alexandru Tifrea, and Elena Burceanu. Rethinking sparse autoencoders: Select-and-Project for fairness and control from encoder features alone. *arXiv preprint arXiv:2509.10809*, 2025.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Rabin, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *NeurIPS*, 2023.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Adriana Romero, Xiaohua Zhai, and Neil Houlsby. PaLI-Gemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

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

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

Fanqi Chen et al. JudgeLRM: Large reasoning models as a judge. *arXiv preprint arXiv:2504.00050*, 2025.

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

Walter Gerych et al. BendVLM: Test-time debiasing of vision-language embeddings. In *NeurIPS*, 2024.

Floris Holstege, Shauli Ravfogel, and Bram Wouters. Preserving task-relevant information under linear concept removal. *arXiv preprint arXiv:2506.10703*, 2025.

Sonia Joseph et al. Steering CLIP's vision transformer with sparse autoencoders. In *MIV Workshop at CVPR*, 2025.

Hoin Jung, Taeuk Jang, and Xiaoqian Wang. A unified debiasing approach for vision-language models across modalities and tasks. In *NeurIPS*, 2024.

Kimmo Kärkkäinen and Jungseock Joo. FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *WACV*, 2021.

Adam Karvonen et al. SAEBench: A comprehensive benchmark for sparse autoencoders in language model interpretability. In *ICML*, 2025.

Majid Molahasani et al. PRISM: Reducing spurious implicit biases in vision-language models with LLM-guided embedding projection. In *ICCV*, 2025.

Mateusz Pach et al. Sparse autoencoders learn monosemantic features in vision-language models. *arXiv preprint arXiv:2504.02821*, 2025.

Kuleen Sasse, Shan Chen, Jackson Pond, Danielle Bitterman, and John Osborne. debiaSAE: Benchmarking and mitigating vision-language model bias. *arXiv preprint arXiv:2410.13146*, 2024.

Samuel Stevens et al. Interpretable and testable vision features via sparse autoencoders. *arXiv preprint arXiv:2502.06755*, 2025.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Mikołaj Zaigrajew et al. Interpreting CLIP with hierarchical sparse autoencoders. In *ICML*, 2025.

Haoyu Zhang, Yangyang Guo, and Mohan Kankanhalli. Joint vision-language social bias removal for CLIP. In *CVPR*, 2025.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *ICLR*, 2020.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *EMNLP*, 2017.

A Experimental Details

A.1 SAE Training Hyperparameters

All SAEs use the BatchTopK architecture [Bussmann et al., 2024] with dictionary size = $4 \times d$ and k=64. Training uses the Adam optimiser with learning rate 3×10^{-4}, batch size 128, for 10 epochs on FairFace training set latents (73,732 images). Table 9 reports reconstruction quality on the validation set (13,012 images).

Table 9: SAE reconstruction quality by model.

VLM	Input dim	Dict size	Cosine sim	Dead features (%)
PaliGemma2	1152	4608	0.945	0.07%
Qwen2-VL	1536	6144	0.959	1.2%
Qwen3-VL	2048	8192	0.975	—

A.2 SDF Discovery Pipeline Details

For each demographic attribute with C classes, we identify 50 candidate SDFs per class using a three-stage pipeline:

1. **Frequency filter**: Retain features that activate (value > 0) in >10% of samples from at least one class.
2. **Mean-activation ranking**: For each class c, compute the mean activation $\bar{a}_{c,j}$ of feature j across all samples with label c. Rank features by $\max_c \bar{a}_{c,j} / (\min_c \bar{a}_{c,j} + \epsilon)$ and retain the top 50 per class.
3. **Entropy filter**: Compute the Shannon entropy of each feature's class-conditional activation distribution. Retain features with entropy below the median (high discriminability).

After deduplication across classes, this yields 50–80 unique SDFs per attribute. We measure *alignment rate*: the fraction of SDFs that have significantly different mean activations across demographic classes (Kruskal-Wallis $p < 0.05$). Alignment rates range from 0.185 (age) to 0.550 (gender) across models, indicating that gender is most cleanly encoded in the SAE feature space.

A.3 LEACE Implementation

We use the `concept-erasure` library [Belrose et al., 2023] to fit a LeaceEraser on mean-pooled VE latents (averaged across the token dimension) with demographic labels. The eraser is applied per-token during VLM inference via a forward hook on the vision encoder output. For binary attributes (gender), LEACE removes a rank-1 subspace; for multi-class attributes (race: 7 classes, age: 9 classes), it removes a rank-(C−1) subspace.

Mean-pooled fitting, per-token application. We fit the LEACE projection on mean-pooled VE features because demographic labels are image-level (not token-level). The resulting projection is then applied identically to each visual token during inference. This assumes the concept direction is shared across spatial positions. On PaliGemma2, the high BERTScore (0.974) and cosine similarity (0.9991) under LEACE confirm that this approximation introduces minimal distortion. However, if demographic information is encoded differently across spatial tokens (e.g., concentrated in face-region tokens), the mean-pooled projection may incompletely erase the concept at the token level. This could contribute to LEACE's limited captioning effect on PaliGemma2, in addition to the nonlinear encoding hypothesis discussed in Section 4.4.

A.4 LLM Judge Implementation

Binary classification (Qwen3-8B). Each caption is evaluated independently with the prompt: "Does the following caption mention or imply anything about the [attribute] of the person(s) described? Answer only YES or NO." We compute LLM-DCR as the fraction of "YES" responses.

Pairwise comparison (JudgeLRM-7B). Original and modified captions are presented side-by-side (order randomised) with the prompt: "Which caption contains more information about the [attribute] of the person(s) described? Answer A, B, or TIE." Position bias is controlled by randomisation with a fixed seed.

Both judges achieve 0% parse failure rate across all evaluations (0 / 54,000+ binary prompts; 0 / 26,100+ pairwise prompts).

B Additional Results

B.1 Dose-Response Under Feature Amplification

Table 10: Gender DCR delta under feature amplification (PaliGemma2, n=500).

Alpha	**Gender ΔDCR**	**Race ΔDCR**	**BERTScore**
0 (suppress)	$-27.6\,pp$	$-2.8\,pp$	0.932
1 (passthrough)	$-21.0\,pp$	$-4.6\,pp$	0.933
3 (amplify)	$-17.4\,pp$	$-2.8\,pp$	0.920
5 (amplify)	$-25.4\,pp$	$-3.2\,pp$	0.910
10 (amplify)	$-69.4\,pp$	$+2.8\,pp$	0.888

All amplification levels produce negative gender DCR deltas (Table 10), which is initially counterintuitive: amplifying gender-aligned features should increase gender content. However, all conditions pass through the SAE bottleneck, which imposes a baseline ∼21 ppreduction (alpha=1 passthrough). The dose-response is non-monotonic: moderate amplification (alpha=3) partially compensates for the bottleneck ($-17.4\,pp$ vs. $-21.0\,pp$ for passthrough), while extreme amplification (alpha=10) overwhelms the language model's capacity to integrate the exaggerated features, producing $-69.4\,pp$ reduction. This pattern reinforces the interpretation that the SAE bottleneck is the dominant mechanism: even amplification cannot fully overcome the information loss from sparse reconstruction on PaliGemma2.

B.2 Intersectional Analysis

ANOVA interaction effects across 177 intersectional SDFs (age × gender × race): 0/177 features pass the strict criterion (Bonferroni $p < 0.05$ AND partial $\eta^2 > 0.01$). Mutual information decomposition reveals mean synergy of -0.007 bits (redundancy dominates). The SAE learns disentangled, additive representations of demographics—intersectional interactions exist statistically but explain negligible variance.

B.3 Race-Gender Feature Entanglement

Ablation of which race SDFs drive gender DCR changes on PaliGemma2: the 20 "universal demographic detector" features (shared across all 7 race classes) account for 90% of the incremental gender DCR effect ($-7.2\,pp$ vs. $-0.8\,pp$ for 20 class-specific features). This confirms that race-gender entanglement is encoded in features detecting demographic salience generally, not race-specific features.

B.4 Qualitative Analysis of the Qwen2-VL Reverse Effect

Table 11 shows representative examples of the opposite-direction effect on Qwen2-VL (Section 4.3). Of 500 samples under gender SDF suppression, 127 (25.4%) *gained* gender-specific language while only 8 (1.6%) lost it. The dominant pattern is clear: Qwen2-VL's original captions use neutral language ("a person," "the child," "the individual") that is replaced by gendered descriptions ("a man," "a boy," "he") after suppression. In several cases (samples 9, 25, 71), the model assigns the *incorrect* gender, and in others (sample 21), the model hallucinates entirely new scene content. This is consistent with SAE reconstruction disrupting the visual features that support the model's neutral-language strategy, causing fallback to more specific (and often inaccurate) descriptions.

Table 11: **Qualitative examples of the Qwen2-VL reverse effect.** Gender SDF suppression converts neutral captions into gendered ones. Ground-truth gender labels shown for reference; suppressed captions sometimes assign the wrong gender.

Label	**Original caption**	**After gender SDF suppression**
Male, 40s	The image shows **a person** wearing glasses and a jacket. The background is blurred…	The image shows **a man** wearing a pair of glasses and a suit. **He** is holding a small object…
Female, 20s	The image shows a blurred face of **a person**. The individual appears to be wearing earrings…	The individual appears to be **a man** with short hair and is wearing earrings…
Female, 60s	The image shows a close-up of **a person's** face. The individual appears to be elderly…	A close-up of a person's face, likely **a man**, with a serious expression. The person has short, neatly combed hair…
Female, 30s	**The individual** has a fair complexion and is wearing a blue top…	The individual appears to be **a man** with short hair, wearing a light-colored shirt…
Male, 30s	The image shows **a close-up of a person's face**. The individual has short, dark hair and is smiling…	A close-up of **a man's face**… **He** appears to be smiling and is wearing a light-colored shirt…