BIAS BEFORE PIXELS: A SURROGATE FRAMEWORK FOR AUDITING DEMOGRAPHIC BIAS IN TEXT-TO-IMAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-image (T2I) models reproduce demographic biases from their training data, but auditing this bias traditionally requires generating and classifying large image batches for every evaluated prompt. We train a surrogate that predicts the demographic distribution of a model's generated images directly from the prompt, so new prompts can be audited without generating any images. The surrogate can reuse the target model's own text encoder, or use an external encoder that requires no access to the model's weights, making it also suitable for API-only systems. We construct a dataset of long and short prompts paired with measured gender and race distributions for three T2I models (FLUX.2-klein, SDXL Turbo, Z-Image Turbo), on which our surrogate improves gender bias-detection accuracy by 11–25% over the strongest text-based or diffusion-internal baseline, with consistent gains on race. Since each audit takes a single forward pass, we can compare bias across models and find concepts and words whose demographic skew differs sharply between them. This efficiency also lets us test all 9,396 object–activity combinations, which shows that activities shape gender bias more than objects. Finally, the surrogate improves an existing debiasing pipeline by predicting failed interventions before image generation, raising gender fairness from 0.62 to 0.74 without reducing image quality. Surrogates thus make it practical to audit bias at scale, before generating a single image and even without access to model weights.

1 INTRODUCTION

Text-to-image (T2I) generative models reproduce and often amplify demographic biases present in their training data, for instance, associating certain professions or activities disproportionately with one gender even when a prompt makes no explicit reference to gender at all (Kim et al., 2025; Parihar et al., 2024; Shi et al., 2025; Girrbach et al., 2025). Such biases are documented across model families and identity axes, from defaulting to light-skinned men to falling back on stereotypes and skin-tone stigma (Ghosh & Caliskan, 2023; Sun et al., 2024; Fraser et al., 2023; Wilson et al., 2025; Zhou et al., 2024), trace back in part to web-scraped training data (Girrbach et al., 2026; Seshadri et al., 2024), and are surveyed by Wan et al. (2024). At deployment scale, these biases amplify viewer-facing harms (Hao et al., 2024) and measurably shift viewers' gender and racial stereotypes (AlDahoul et al., 2025). Yet auditing this bias traditionally requires generating large image batches per prompt and classifying the people depicted—a costly, multi-stage pipeline that must be repeated for every prompt and model under audit.

We address this bottleneck using a *surrogate*, a model trained to predict, directly from a text prompt, the attribute distribution a fixed generative model G would produce, without invoking G at inference time. The formulation is not tied to fixed categories, so it applies to any attribute for which its empirical ground truth can be built. We evaluate it on gender and race, two of the most consistently documented axes of T2I bias.

Because the surrogate is trained purely on (prompt, empirical distribution) pairs, it supports two deployment modes. First, it can reuse G's own text encoder when available, pairing with existing debiasing pipelines at minimal cost. Second, it can use an external encoder with no access to G's

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

weights or activations at inference, so it can also audit closed, API-only models. In our experiments, this external-encoder variant is empirically *more* reliable than existing white-box alternatives.

Although trained only on prompts paired with measured distributions, our surrogate empirically picks up fine-grained lexical detail. For example, appending an incidental clause (*"with a book left on the counter"*) to the neutral prompt *"A person working as a baker"* shifts FLUX.2-klein's predicted distribution to a pronounced female skew, as the model's generations confirm (Figure 1). We construct a new dataset of long and short prompts with measured gender and race distributions for three architecturally distinct T2I models: *FLUX.2-klein*, *SDXL Turbo*, and *Z-Image Turbo*. Because the surrogate reduces bias estimation to a single forward pass, we can compare the *same* prompt across models and identify concepts and words whose demographic associations differ sharply (Sections C.4 and 5.2). It also lets us exhaustively test object–activity combinations to see which one drives bias (Section 5.3) and audit the unseen TIFA benchmark without generating a single image (Section 5.4).

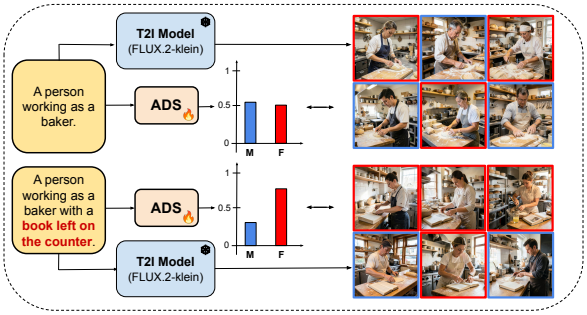


Figure 1: An incidental clause shifts the surrogate's prediction (left) to a female skew, which the model's generations confirm (right; border marks male/female).

Across our constructed dataset, our surrogate outperforms text-embedding (Zou et al., 2023) and diffusion-internal (Li et al., 2024) baselines, improving gender bias-detection accuracy by 11–25% and lowering the total variation (TV) distance on race by 0.09–0.34 compared with the strongest baseline. Beyond auditing, we show that the surrogate improves existing debiasing pipelines (Section 5.5). Indeed, its prediction on a steered embedding anticipates the generated image, so we use it to catch likely debiasing failures and fix them by reselecting the steering perturbation, without retraining. This raises gender fairness from 0.62 to 0.74 without reducing image quality.

In summary, we contribute: (1) a surrogate framework that predicts a T2I model's attribute distribution from the prompt alone, in a native-encoder and a fully black-box variant; (2) a dataset pairing tens of thousands of long and short prompts with measured gender and race distributions for three T2I models, with an evaluation against text-embedding and diffusion-internal baselines; (3) a cross-model comparison of concepts and words whose demographic association differs between models, an exhaustive test of object–activity combinations, and an audit of the unseen TIFA benchmark, all made practical by single-pass auditing; and (4) a demonstration that the surrogate improves an existing debiasing pipeline by anticipating its failures before generation.

2 Related Work

A growing body of work interprets diffusion models' internal computation to understand where bias arises. DAAM (Tang et al., 2023) aggregates cross-attention into pixel-level attribution maps, and Diffusion Lens (Toker et al., 2024) probes the text encoder's intermediate layers to show that concepts compose gradually across depth. OVAM (Marcos-Manchón et al., 2024) extends attention maps to words outside the prompt via training-free token optimization.

A related line uses this internal access to discover bias specifically. Li et al. (2024) recover a linear attribute direction from the diffusion latents, and Chakraborty et al. (2026) read bias off cross-attention overlap between demographic and semantic concepts. Moreover, Wu et al. (2024) show that gender cues also skew object choice and layout. Closest in structure to our surrogate is the Attribute Distribution Predictor (ADP) of Parihar et al. (2024), a small MLP that also outputs a full attribute distribution, but from a diffusion U-Net's latents during denoising. It therefore cannot audit a prompt before generation starts, whereas our surrogate works from text alone.

Other work discovers bias from the output distribution itself. Seshadri et al. (2024) trace most of Stable Diffusion's apparent bias amplification to a mismatch between gendered training captions and gender-neutral evaluation prompts, and OpenBias (D'Incà et al., 2024) combines an LLM, the generator, and a VQA model to propose and quantify open-set biases without a predefined list. All

of these still require running the generative model, as do large-scale generate-and-classify studies of bias across models, occupations, and identity groups (Luccioni et al., 2023; Mannering, 2023; Bianchi et al., 2023; Cho et al., 2023; Naik & Nushi, 2023; Wang et al., 2023; Zhang et al., 2024; Ungless et al., 2023; Bakr et al., 2023; Jha et al., 2024; Vice et al., 2025; Wang et al., 2024). The most direct antecedent is Girrbach et al. (2025). We adopt their detection-classification pipeline (YOLOv10 + InternVL-2) for ground truth (Section 3.2), but instead of generating and classifying images for every prompt and model, we train a surrogate once and predict the same signal for new prompts or unseen benchmarks directly from text. Finally, debiasing methods steer generation toward balanced outputs without retraining the generator (Jiang et al., 2025; Fu et al., 2026; Shi et al., 2025), and we show that our surrogate makes such interventions more reliable (Section 5.5).

3 METHOD

In this section, we build a dataset of prompts paired with measured attribute distributions (Section 3.2), train a text-only surrogate to predict these distributions (Section 3.3), and fit it with a distribution-matching objective (Section 3.4).

3.1 PROBLEM FORMULATION

Let G be a fixed, target T2I model being audited, and $\mathcal{K}$ be a finite set of attribute categories with $K = |\mathcal{K}|$. In this work, we evaluate two instantiations, $\mathcal{K} = \{\text{male}, \text{female}\}$ for gender and $\mathcal{K} = \{\text{White, Black, Asian, Indian, Middle Eastern, Latino/Hispanic}\}$ for race.[1] Δ^{K-1} denotes the probability simplex over $\mathcal{K}$. For a prompt x, generating many images with G and classifying the attributes of the people in them induces an empirical distribution $P_G(x) \in \Delta^{K-1}$. Our goal is a surrogate $f_\phi : x \mapsto \hat{P}(x) \in \Delta^{K-1}$, with trainable parameters ϕ separate from G, such that

$$\hat{P}(x) = f_\phi(x) \approx P_G(x). \tag{1}$$

The surrogate is trained once on prompts with measured $P_G(x)$ (Section 3.2) and then queried with a single forward pass, without invoking G. We write $\hat{P}(x)$ for the surrogate's prediction throughout, suppressing ϕ, and reuse the same notation for baselines.

3.2 DATASET CONSTRUCTION

Given that no suitable dataset existed, we build one from long, descriptive captions and short, templated prompts with controlled wording, which we call *Descriptive* and *Templated* prompts.

Descriptive prompts. We sample $\sim$13k prompts from two public human-image captioning datasets, ***Human Caption-HQ*** (Dai et al., 2025) and ***FaceCaption-15M*** (Dai et al., 2024), downweighting the longest Human Caption-HQ captions to balance lengths with the shorter FaceCaption-15M ones. Raw captions do not mention race or ethnicity (although detailed ones may carry correlated facial cues), but they often contain gendered words and appearance markers, so we rewrite them in two stages. First, a regex pass neutralizes gendered words and appearance cues, keeping explicit pronouns in 10% of prompts for diversity and baseline construction (Section 4.1). Second, we add random context to subsets of prompts, either a scene, occupation, activity, and object, or a demographic or cultural descriptor (region, religion, or ethnicity). Most prompts are thus neutral in gender and ethnicity, so bias stems from context rather than wording. A similarly small share with explicit gender or ethnicity cues keeps such prompts in distribution for both attributes. Exact percentages and vocabularies are in Section B.1.

Templated prompts. We also incorporate the corpus of short, everyday-situation prompts from Girrbach et al. (2025) ($\sim$15k prompts covering $\sim$3k concepts across activities, objects, occupations, and contexts). Each template uses one of five neutral subject nouns (*person*, *individual*, *friend*, *colleague*, *someone*), giving five prompts per concept, e.g., "*a high-quality picture of a person working as veterinary assistant*". As these prompts are neutral in gender and ethnicity, any bias can be attributed to the *concept* rather than the subject word.

Combining both prompt styles yields six splits: `Train-Desc`/`Train-Tmpl` ($\approx$10,100 / $\approx$12,000 prompts), `Val-Desc`/`Val-Tmpl` ($\approx$1,000 each), and `Test-Desc`/`Test-Tmpl` ($\approx$2,000 each). The Templated splits are divided by concept, so no validation or test concept appears during training.

[1] Our method applies to any attribute for which ground truth can be obtained.

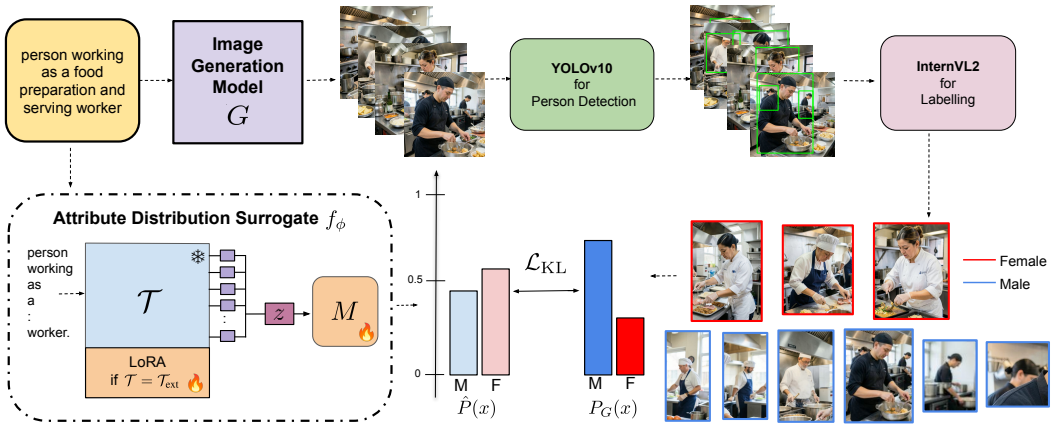


Figure 2: A prompt is encoded by a LoRA-adapted (or the target model's own, frozen) text encoder; token embeddings are mean-pooled and projected by a small MLP head to a K-way softmax distribution $\hat{P}(x)$, trained with a KL objective against the empirical ground-truth distribution $P_G(x)$.

Ground-truth distribution extraction. For each target model G, we generate 1 image per `Train-Desc` prompt, 10 per `Train-Tmpl` prompt, and 40 per validation/test prompt for more stable estimates. A single image often contains several people and thus already yields a distribution. Beyond that, we favor more diverse training prompts over more images per prompt, expecting overlapping prompts to share signal (Section B.2). We detect people with **YOLOv10** and classify each crop into one of the K categories (or a residual *Not Clear/Uncertain* label) with **InternVL-2**, following Girrbach et al. (2025). Per prompt, we discard uncertain labels and normalize the remaining counts, $p^{(k)} = \#k / \sum_{k' \in \mathcal{K}} \#k'$ (where $\#k$ is the number of crops labeled k), yielding $P_G(x) = [p^{(k)}]_{k \in \mathcal{K}} \in \Delta^{K-1}$. Prompts with only uncertain detections are dropped entirely. We train on both prompt styles together but evaluate them separately. For SDXL Turbo, whose CLIP text encoders truncate prompts at 77 tokens, we use only the Templated split.

3.3 Surrogate Architecture

Our *Attribute Distribution Surrogate* (ADS) has two variants that share one design (Figure 2). A text encoder $\mathcal{T}$ maps prompt $x = (x_1, \ldots, x_L)$ to hidden states $Z = \mathcal{T}(x) \in \mathbb{R}^{L \times d}$, which are mean-pooled into $z \in \mathbb{R}^d$. A small MLP head M with two hidden layers of width 512 and 256 (Section B.4) maps z to K logits, and a softmax gives $\hat{P}(x)$. The variants differ only in where $\mathcal{T}$ comes from and whether it is LoRA-adapted.

ADS-External supplies its own pretrained text encoder, entirely independent of G, frozen and adapted via LoRA on its query/value attention projections (following Hu et al., 2022), so only these low-rank matrices and the MLP head are trained. Because $\mathcal{T}$ is independent of G, this variant also works for black-box models, and G is queried only once to collect training labels.

ADS-Native reuses G's text encoder unchanged since LoRA would alter the embedding G actually generates from. Because this embedding is already computed during generation, ADS-Native adds only a small MLP pass, letting a deployment with access to G audit every prompt *while* generating.

3.4 Training Objective

We train f_ϕ to match $\hat{P}(x)$ to the empirical ground-truth distribution $P_G(x)$ via the *forward* Kullback–Leibler divergence

$$\mathcal{L}_{\text{KL}} = \sum_{k \in \mathcal{K}} P_G^{(k)}(x) \log \left(\frac{P_G^{(k)}(x)}{\hat{P}^{(k)}(x)} \right), \tag{2}$$

averaged over the training batch. Its mode-covering behavior penalizes the surrogate for assigning low probability to any category present in the ground truth, whereas the mode-seeking reverse KL would let it collapse onto the dominant category (Singh et al., 2025).

4 EXPERIMENTAL SETUP

To ensure our results are not an artifact of a single model family, we evaluate on three T2I backbones chosen to span different generative paradigms: **FLUX.2-klein** (a 4B flow-matching Diffusion Transformer), **SDXL Turbo** (a latent-diffusion UNet), and **Z-Image Turbo** (a single-stream Diffusion Transformer). We train a separate surrogate for each target model on that model's ground truth. ADS-External uses T5-XXL as its encoder, while ADS-Native uses each model's own encoder, which is Qwen3-4B for FLUX.2-klein and Z-Image Turbo, and the concatenation of CLIP ViT-L and OpenCLIP ViT-bigG for SDXL Turbo.

4.1 BASELINES

We compare against four baselines, two that read the prompt's text embedding and two that require running the diffusion model itself. All of them predict $\hat{P}(x)$ and are scored against the same ground truth and metrics, with full formulations in Section B.3. For race, we compare against the two probes only, since RepE and InterpretFlux recover a single direction between two categories and do not extend straightforwardly to six. Ablations of encoder scale, LoRA, and training data are in Section C.2.

Text Embedding Probe. The ADS-Native architecture (the target model's frozen encoder and the same MLP head), trained on generic prompts labeled by their explicit pronouns instead of on our dataset. Comparing it with ADS-Native therefore isolates the effect of our training data.

RepE Reading Vector (Zou et al., 2023). The first principal component of embedding differences between paired prompts (e.g., "a photo of a male plumber" vs. "a photo of a female plumber"). New prompts are projected onto it and calibrated with logistic regression.

InterpretFlux (Li et al., 2024). A concept direction discovered in the diffusion transformer's intermediate activations, which requires running G at inference. We evaluate it on FLUX.2-klein, where we identified a suitable latent space (Section B.3).

Diffusion Activation Probe. (Parihar et al., 2024) A one-hidden-layer MLP (width 512, ReLU) on pooled activations from G's early denoiser blocks. Like InterpretFlux, it is white-box, and it tests whether diffusion-internal signal goes beyond the text embedding.

4.2 EVALUATION METRICS

We use three metrics. Bias detection accuracy (gender only) and TV distance are averaged over test prompts, while Spearman correlation (gender) and entropy correlation (race) are computed across them. Together, they measure whether the verdict is right, how close the predicted distribution is, and whether the surrogate tracks how strongly biased (gender) or how mixed (race) each prompt is.

Bias Detection Accuracy. For gender, we write $P(\text{Male})$ for a distribution's male probability, so that $P(\text{Female}) = 1 - P(\text{Male})$. With a threshold $\tau = \frac{1}{3}$, we assign each distribution P a bias class $C_\tau(P)$, which is Male-Biased if $P(\text{Male}) \geq 1 - \tau$, Female-Biased if $P(\text{Male}) \leq \tau$, and Neutral otherwise, and we call $(\frac{1}{3}, \frac{2}{3})$ the neutral band. Accuracy is the fraction of test prompts whose predicted and ground-truth distributions fall in the same class, $C_\tau(\hat{P}(x)) = C_\tau(P_G(x))$, since audits typically need a discrete verdict rather than an exact probability. We discuss the choice of τ in Section C.1.

Total Variation Distance. $\text{TV}(P_G(x), \hat{P}(x)) = \frac{1}{2}\sum_{k\in\mathcal{K}} |P_G^{(k)}(x) - \hat{P}^{(k)}(x)|$ (0 for identical distributions, 1 for disjoint, lower is better) is the largest error the prediction makes on the total probability of any group of categories. For gender, it is just the absolute error in $P(\text{Male})$.

Spearman and Entropy Correlation. For gender, we report the Spearman correlation of predicted and ground-truth $P(\text{Male})$, checking whether the surrogate orders prompts by bias correctly, independent of calibration. Race has no single ranking axis, so we report entropy correlation (Singh et al., 2025), i.e., the Pearson correlation between Shannon entropies $H(P_G(x))$ and $H(\hat{P}(x))$. A high value means the surrogate is uncertain where the ground truth itself is mixed.

5 EXPERIMENTAL RESULTS

We first compare our surrogate with the baselines on gender and race (Section 5.1). We then use it to compare bias across models at the concept level (Section 5.2), to test all object–activity combi-

Table 1: Comparison of our surrogate (ADS variants) against four distinct baselines for Gender; SDXL Turbo has no Descriptive split (Section 3.2).

Model	Method	Acc.↑	TV Dist.↓	Spearman↑
Descriptive				
FLUX.2-klein	Diff. Activation Probe	0.7221	0.1832	0.7597
	RepE Reading Vec.	0.3665	0.3641	0.3608
	InterpretFlux	0.3610	0.4436	0.0194
	Text Embedding Probe	0.5273	0.2718	0.7191
	ADS-Native (Ours)	0.7927	0.1310	0.8659
	ADS-External (Ours)	**0.8348**	**0.0977**	**0.9073**
Z-Image Turbo	Diff. Activation Probe	0.7200	0.2030	0.6988
	RepE Reading Vec.	0.7014	0.2259	0.7396
	Text Embedding Probe	0.5239	0.3529	0.3275
	ADS-Native (Ours)	0.8135	0.1349	0.8329
	ADS-External (Ours)	**0.8441**	**0.1163**	**0.8657**
Templated				
FLUX.2-klein	Diff. Activation Probe	0.5025	0.2439	0.4715
	RepE Reading Vec.	0.3660	0.2946	0.1220
	InterpretFlux	0.3680	0.3447	−0.0304
	Text Embedding Probe	0.3835	0.2431	0.4317
	ADS-Native (Ours)	0.7355	0.1187	0.8557
	ADS-External (Ours)	**0.7505**	**0.1138**	**0.8573**
SDXL Turbo	Diff. Activation Probe	0.5351	0.3047	0.4879
	RepE Reading Vec.	0.4624	0.3380	0.1325
	Text Embedding Probe	0.2731	0.3050	0.1951
	ADS-Native (Ours)	**0.7880**	0.1275	0.7588
	ADS-External (Ours)	0.7698	**0.1257**	**0.7815**
Z-Image Turbo	Diff. Activation Probe	0.5344	0.2590	0.3708
	RepE Reading Vec.	0.3474	0.3129	0.0925
	Text Embedding Probe	0.4646	0.3354	0.1633
	ADS-Native (Ours)	0.6843	0.1660	0.6797
	ADS-External (Ours)	**0.7320**	**0.1487**	**0.7286**

Table 2: Our surrogate (ADS variants) for Race against a diffusion-internal-activation probe and a frozen text-embedding probe.

Model	Method	TV Dist.↓	Entr. Corr.↑
Descriptive			
FLUX.2-klein	Diff. Activation Probe	0.5200	−0.0339
	Text Embedding Probe	0.5534	−0.0269
	ADS-Native (Ours)	**0.3415**	**0.3558**
	ADS-External (Ours)	0.3664	0.3333
Z-Image Turbo	Diff. Activation Probe	0.6239	0.0936
	Text Embedding Probe	0.7180	0.0643
	ADS-Native (Ours)	**0.3591**	0.2668
	ADS-External (Ours)	0.3732	**0.2823**
Templated			
FLUX.2-klein	Diff. Activation Probe	0.5140	0.1298
	Text Embedding Probe	0.7414	0.0008
	ADS-Native (Ours)	**0.1732**	**0.6517**
	ADS-External (Ours)	0.1749	0.5928
SDXL Turbo	Diff. Activation Probe	0.2650	−0.0213
	Text Embedding Probe	0.6338	−0.0331
	ADS-Native (Ours)	**0.1720**	**0.4729**
	ADS-External (Ours)	0.1793	0.4350
Z-Image Turbo	Diff. Activation Probe	0.5395	0.0769
	Text Embedding Probe	0.7708	0.0445
	ADS-Native (Ours)	0.2289	**0.3122**
	ADS-External (Ours)	**0.2228**	0.2891

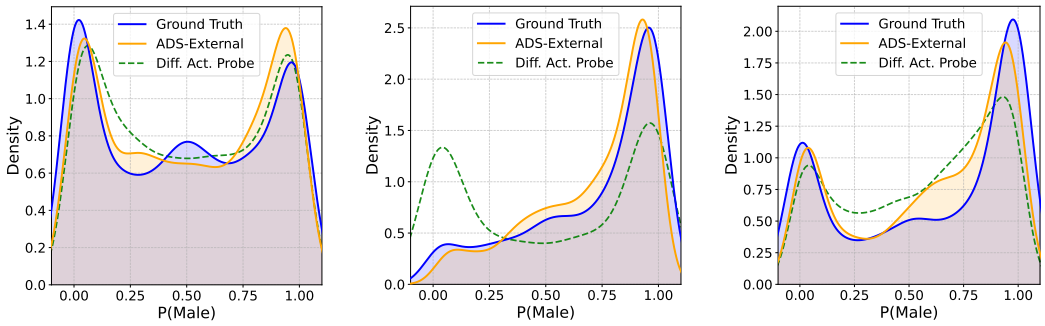
nations (Section 5.3), and to audit the unseen TIFA benchmark before any generation (Section 5.4), before applying it to debiasing (Section 5.5). Sections 5.1 and 5.2 cover all three models, while the remaining sections focus on FLUX.2-klein to showcase capabilities the surrogate newly enables.

5.1 Main Results

Gender. Table 1 compares our surrogate against four baselines across both splits and all three models, and ours wins on accuracy, TV distance, and Spearman in every block. We observe two patterns. First, diffusion-internal access alone does not guarantee a better signal. InterpretFlux performs worst overall, while the Diffusion Activation Probe, with the same access, is the strongest baseline overall yet still trails our surrogate, so how the signal is extracted matters more than whether such access exists. Second, the gap to the plain Text Embedding Probe (e.g., FLUX.2-klein `Test-Desc` 0.53 vs. 0.83 accuracy) stems from two compounding factors, isolated in Table 8. Curated data alone recovers most of it (0.53 → 0.79), and LoRA adds a further, consistent gain. In short, the dataset determines how much signal is available, and LoRA determines how well it is used.

Race. Table 2 reports race, where we compare against the two probes (Section 4.1). Both surrogate variants clearly outperform them. ADS-Native achieves the best TV distance and entropy correlation in most blocks, and ADS-External is close behind and even ahead on some Z-Image Turbo metrics. The baselines' entropy correlations are near zero or negative (e.g., −0.03 for FLUX.2-klein Descriptive), so their predictions carry essentially no information about how mixed a prompt is, whereas ADS-Native reaches 0.27 to 0.65.

Distributional fidelity. Figure 3 compares the kernel density estimates of our surrogate's predicted versus ground-truth $P(\text{Male})$ for all models (both test splits jointly, Templated for SDXL Turbo). For visual comparison, we include the Diffusion Activation Probe (green), our strongest baseline from Table 1. Across all three models, our surrogate closely tracks the shape of the ground truth, with most mass near the extremes and a smaller peak near 0.5. In contrast, while the baseline reasonably approximates the ground truth on FLUX.2-klein and Z-Image Turbo, it degrades noticeably on SDXL Turbo. There, it introduces a spurious low-probability mode and clearly undershoots the main peak near $P(\text{Male}) \approx 1$, which our surrogate does not show. A similar alignment holds for race (Section C.7).



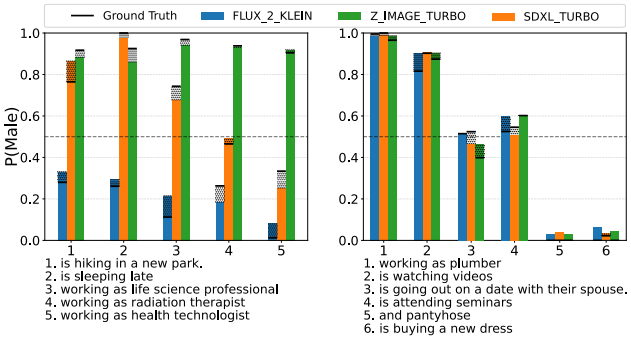
(a) FLUX.2-klein (b) SDXL Turbo (Templated) (c) Z-Image Turbo

Figure 3: Predicted vs. ground-truth $P(\text{Male})$ density per model, with the strongest baseline, Diffusion Activation Probe (green).

5.2 Cross-Model Comparison at Concept Level

We measure each unseen `Test-Tmpl` concept's $\bar{P}(\text{Male})$ by averaging the predicted or ground-truth $P(\text{Male})$ over its five subject-word variants (Section 3.2). Every comparison below can be checked against actual generations, shown as black ticks in Figure 4.

Concepts diverge sharply by model. For each concept, we rank by the largest gap in $\bar{P}(\text{Male})$ between any two models (Figure 4, (a)). The most divergent concepts show gaps of 0.6–0.8 for identical text. For "*working as a health technologist*", for instance, $\bar{P}(\text{Male})$ is about 0.08 in FLUX.2-klein but 0.92 in Z-Image Turbo. The ground truth closely matches these predictions, so the gaps are *real* model differences rather than prediction artifacts.



(a) Most divergent (b) Agreement & neutrality

Figure 4: Selected `Test-Tmpl` concepts. Bars show the surrogate's predicted $\bar{P}(\text{Male})$ per model, and black ticks mark the ground truth from generated images. (a) The five most divergent concepts. (b) Six concepts where all three models agree, whether on a clear bias direction or on Neutral.

Models agree at both extremes, and at neither. The three models also agree on many concepts, both on clear biases and on neutrality (Figure 4, (b)). Across all three models, the ground truth exceeds 0.81 for "*is watching videos*" and 0.96 for "*working as plumber*", and stays under 0.03 for "*and pantyhose*" and "*is buying a new dress*". In between, concepts like "*is attending seminars*" and "*is going out on a date with their spouse*" land squarely in the neutral band (0.40–0.60) for every model. The neutral cases likely arise because they depict companions or groups rather than one clearly gendered individual. With predictions matching the ground truth here as well, the surrogate does not report bias indiscriminately but reproduces the level of skew the models exhibit.

5.3 Combinatorial Bias Analysis

Beyond qualitative examples, we exhaustively cross every object and activity concept from `Test-Tmpl`, evaluating all 58 $\times$ 162 = 9,396 object–activity combinations in one batched pass, a sweep that would require generating and classifying images for 9,396 new prompts under the traditional pipeline. Cross-tabulating every combination by the object's and activity's own bias class (Figure 5) shows that the interaction is asymmetric. A Male-classified object paired with a Neutral activity collapses into the Neutral band 89.9% of the time, while a Male-classified activity paired with a Neutral object neutralizes only 33% of the time. We confirm this by isolating matched Male-

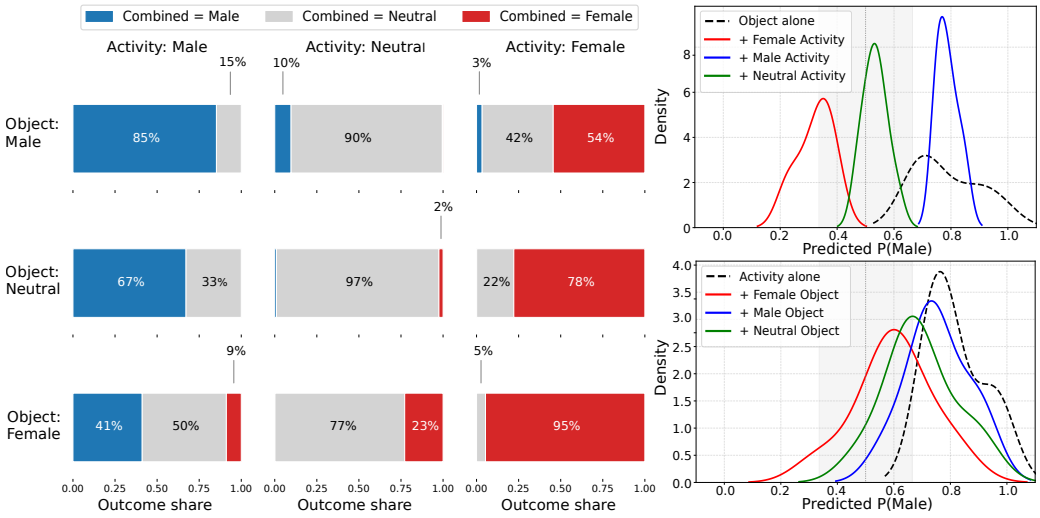


Figure 5: Left: the combined prediction's outcome class as a function of the object's and activity's own bias classes. Top Right: Average effect of adding activities across a subset of male biased objects. Bottom Right: Average effect of adding objects across a subset of male biased activities.

only subsets. Adding female activities shifts a male object cluster by 0.35 on average, while adding female objects shifts a male activity cluster by only ∼0.15 (Figure 5, top-/bottom-right), with the same pattern holding for neutral and male additions. Activities therefore shift gender bias more than objects, which shows *why* a model is biased, not just *that* it is. A ground-truth check on a smaller sample of generated prompts is consistent with this pattern (Section C.5).

5.4 Pre-Generation Auditing on an External Benchmark

Our surrogate also lets us audit a benchmark *before* committing any compute to image generation. We demonstrate this on TIFA (Hu et al., 2023), a widely used T2I evaluation benchmark whose 4,081 captions (sourced from COCO, DrawBench, PartiPrompts, and PaintSkills) were never seen during surrogate training, running the discovery pass over the entire benchmark in seconds.



Figure 6: Most-neutral prompts from the TIFA benchmark, color-coded by the source of their neutrality. Shaded band marks the Neutral zone.

Unpacking caption neutrality and word-level stereotypes. Among TIFA captions where $P(\text{Male})$ nears 0.5 (Figure 6), the surrogate disentangles three distinct neutrality sources: no human referent ("*a cow eating a green leafy plant*"), mixed-gender crowds ("*a crowd of people watching fireworks by a park*"), and canceling cues ("*This man is showing off the kite he made for his wife*"). An auditor can thus see why a caption is neutral before generating any image, instead of relying on a single aggregate rate. Ranking TIFA content words by frequency-averaged predictions reveals that the surrogate transfers learned cultural co-occurrences to unseen benchmarks without explicit gender language (Figure 13). The top male-associated words are *kilt*, *shouting*, *studded*, *tophat*, and *garbage* (average $P(\text{Male})$ above 0.98), while the top female-associated words are *dress*, *apples*, *flower*, *chocolate*, and *pink* (average $P(\text{Female})$ of 0.79–0.94).

5.5 Application: Reducing Flips in Direction-Guided Debiasing

Beyond auditing, f_ϕ can predict whether a debiasing intervention will fail before any image is generated. Direction-guided debiasing methods steer diffusion generation toward a balanced demographic distribution without retraining the generator, randomly steering each generation toward one

Table 3: Debiasing quality on the same 1,200 generations for gender and race. Fairness is balanced over realized labels, CLIP is CLIP ViT-B/16 prompt-image similarity, MUSIQ assesses visual quality (Ke et al., 2021), and Avg is the unweighted mean of the three.

Pipeline	Gender				Race			
	Fairness ↑	MUSIQ ↑	CLIP ↑	Avg ↑	Fairness ↑	MUSIQ ↑	CLIP ↑	Avg ↑
No debiasing (FLUX.2-klein)	0.0867	**0.7510**	**0.6810**	0.5062	0.4116	**0.7510**	**0.6810**	0.6145
FairImagen (Fu et al., 2026)	0.6233	0.7413	0.6441	0.6696	0.6555	0.7463	0.6423	0.6814
+ surrogate-guided reselection	**0.7450**	0.7429	0.6438	**0.7106**	**0.6951**	0.7457	0.6408	**0.6939**

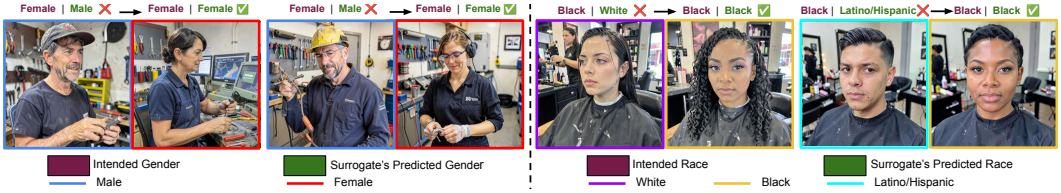


Figure 7: Example generations before (FairImagen) and after (surrogate-guided reselection) for *"a mechanic"* (left) and *"a hair stylist"* (right). Our surrogate detects the flip before generation, and perturbation reselection restores the intended attribute. Colored border represents the ground truth.

pole of an attribute direction so the *aggregate* output balances out (Jiang et al., 2025; Li et al., 2024; Shi et al., 2025; Fu et al., 2026). We build on FairImagen (Fu et al., 2026): given a prompt x and intended direction $k^\star \in \mathcal{K}$, FairImagen projects x's embedding into a subspace that removes attribute-correlated variance, then adds back a scaled copy of $k^\star$'s direction to steer toward it.

Steering does not always succeed. When the realized label $y_G(x)$ (the class detected in the generated image) does not match the intended direction $k^\star$ (a *flip*), Fairness (how balanced the realized labels are, Table 3) suffers directly, so reducing the flip rate suffices to improve it. Our key observation is that f_ϕ's prediction $\hat{y}(x) = \arg\max_{k\in\mathcal{K}} \hat{P}^{(k)}(x)$ on the *steered* embedding, computed instantly with no generation, anticipates $y_G(x)$ well even when it disagrees with $k^\star$. A high flip catch rate, $\mathbb{P}[\hat{y}(x) = y_G(x) \mid y_G(x) \neq k^\star]$, makes pre-generation correction possible, and with reliable labels, the surrogate catches 51% of flips (Section C.6). To exploit this, we keep, for each direction, a pool of candidate steering perturbations already computed by FairImagen (empirical noise vectors added to the text embedding, at no extra cost). Whenever $\hat{y}(x) \neq k^\star$, we apply each perturbation to produce candidate steered embeddings, batch-evaluate them in a single forward pass through f_ϕ, and select one at random among those for which $\hat{y}(x) = k^\star$. Since f_ϕ is lightweight, this vetting adds only seconds of overhead, and the diffusion model is still invoked only once.

Surrogate-based perturbation reselection improves metrics. Correcting surrogate-flagged flips raises Fairness substantially in both settings (Gender: 0.6233 → 0.7450; Race: 0.6555 → 0.6951; Table 3) while leaving MUSIQ and CLIP score essentially unchanged. Figure 7 shows representative flips corrected by our procedure. The reselection is deliberately simple. It searches a fixed pool for a candidate that the surrogate predicts will succeed, rather than learning a correction, yet it already removes more than half of the realized gender flips (163 to 71 of 1,200 generations, Table 9).

6 Conclusion

We introduced a surrogate framework that predicts the demographic distribution of a T2I model's generated images directly from the prompt. It can use the target model's own text encoder or an external one that needs no access to the model's weights, and in both forms it outperforms text-based and diffusion-internal baselines on gender and race. Because each audit takes a single forward pass, the surrogate enables analyses that would be impractical with image generation. They reveal that models diverge sharply on identical prompts and that activities shape bias more than objects, and they agree with the ground truth where we could check. The surrogate also makes an existing debiasing pipeline more reliable by anticipating failed interventions before generation. Together, these results argue that bias auditing should be comparative, granular, and pre-generation, and that a cheap surrogate makes all three practical at once. We discuss limitations, including ground-truth classifier bias and the lack of validation on a closed API model, in Section A.

Ethics Statement

This work studies and quantifies demographic bias in generative models with the explicit goal of making it easier to audit. The surrogate is intended for auditing, not for generating biased content, and the dataset we construct is intended for measurement and research purposes. Our ground-truth labels rely on an automatic classifier (InternVL-2) applied to generated images, which reduces gender to a binary, visually perceived category and race to six fixed, visually perceived categories (White, Black, Asian, Indian, Middle Eastern, Latino/Hispanic). Both are simplifications of real, non-binary and non-discrete human attributes, made for tractability of measurement as in previous works such as Fu et al. (2026), and should not be read as an endorsement of a binary view of gender or of race as a small set of visually discernible boxes. We encourage future work to extend this framework to more inclusive attribute definitions for both.

Reproducibility Statement

Our surrogate architecture, training objective, and hyperparameters are fully specified in Sections 3.3 and 3.4 and Section B.4, including encoder choice, LoRA configuration, quantization scheme, optimizer, and learning rate. All evaluation metrics are defined precisely in Section 4.2, and the baselines are detailed in Section B.3. Dataset construction, including prompt sources, splits, and the ground-truth extraction pipeline, is described in Section 3.2 and Section B.1. The debiasing application's steering, perturbation pool, and reselection procedure are described in Section 5.5. We will release the dataset and code upon publication.

AI Use Statement

In this work, we used generative AI to help implement the pipeline and analysis code as well as to construct the demographic vocabulary and templates used to enrich a subset of prompts (Section B.1). We did not use AI to develop the surrogate architecture and objectives, which are our own contributions. We also used it for creating figures, drafting and editing text, literature search, and suggesting structure and keywords. All AI-assisted code, data, figures, and text were reviewed and verified by the authors, who take full responsibility for the final content of this work.

References

Nouar AlDahoul, Talal Rahwan, and Yasir Zaki. Ai-generated faces influence gender stereotypes and racial homogenization. *Scientific reports*, 15(1):14449, 2025.

Eslam Mohamed Bakr, Pengzhan Sun, Xiaoqian Shen, Faizan Farooq Khan, Li Erran Li, and Mohamed Elhoseiny. Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models. In *2023 ieee/cvf international conference on computer vision (iccv)*, pp. 19984–19996. IEEE, 2023.

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, pp. 1493–1504, 2023.

Rajatsubhra Chakraborty, Xujun Che, Depeng Xu, Cori Faklaris, Xi Niu, and Shuhan Yuan. Biasmap: Leveraging cross-attentions to discover and mitigate hidden social biases in text-to-image generation. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 285–296, 2026.

Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3020–3031. IEEE, 2023.

Dawei Dai, YuTang Li, YingGe Liu, Mingming Jia, Zhang YuanHui, and Guoyin Wang. 15m multimodal facial image-text dataset. *arXiv preprint arXiv:2407.08515*, 2024.

Dawei Dai, Long Xu, Yutang Li, Yuanhui Zhang, and Shuyin Xia. Humanvlm: Foundation for human-scene vision-language model. *Information Fusion*, 123:103271, 2025.

Moreno D'Incà, Elia Peruzzo, Massimiliano Mancini, Dejia Xu, Vidit Goel, Xingqian Xu, Zhangyang Wang, Humphrey Shi, and Nicu Sebe. Openbias: Open-set bias detection in text-to-image generative models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12225–12235, 2024.

Kathleen C. Fraser, Isar Nejadgholi, and Svetlana Kiritchenko. A friendly face: Do text-to-image systems rely on stereotypes when the input is under-specified? In *The AAAI-23 Workshop on Creative AI Across Modalities*, 2023.

Zihao Fu, Ryan Brown, Shun Shao, Kai Rawal, Eoin Delaney, and Chris Russell. Fairimagen: Post-processing for bias mitigation in text-to-image models. *Advances in Neural Information Processing Systems*, 38:29488–29512, 2026.

Sourojit Ghosh and Aylin Caliskan. 'person'== light-skinned, western man, and sexualization of women of color: Stereotypes in stable diffusion. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 6971–6985, 2023.

Leander Girrbach, Stephan Alaniz, Genevieve Smith, and Zeynep Akata. A large scale analysis of gender biases in text-to-image generative models, 2025. URL https://arxiv.org/abs/2503.23398.

Leander Girrbach, Stephan Alaniz, Genevieve Smith, Zeynep Akata, et al. Person-centric annotations of laion-400m: Auditing bias and its transfer to models. In *International Conference on Learning Representations*, volume 2026, pp. 124867–124919, 2026.

Susan Hao, Renee Shelby, Yuchi Liu, Hansa Srinivasan, Mukul Bhutani, Burcu Karagol Ayan, Ryan Poplin, Shivani Poddar, and Sarah Laszlo. Harm amplification in text-to-image models. *arXiv preprint arXiv:2402.01787*, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *2023 ieee/cvf international conference on computer vision (iccv)*, pp. 20349–20360. IEEE, 2023.

Akshita Jha, Vinodkumar Prabhakaran, Remi Denton, Sarah Laszlo, Shachi Dave, Rida Qadri, Chandan Reddy, and Sunipa Dev. Visage: A global-scale analysis of visual stereotypes in text-to-image generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12333–12347, 2024.

Yilei Jiang, Wei-Hong Li, Yiyuan Zhang, Minghong Cai, and Xiangyu Yue. Fairgen: Enhancing fairness in text-to-image diffusion models via self-discovering latent directions. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18411–18420. IEEE, 2025.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5128–5137. IEEE, 2021.

Eunji Kim, Siwon Kim, Minjun Park, Rahim Entezari, and Sungroh Yoon. Rethinking training for de-biasing text-to-image generation: Unlocking the potential of stable diffusion. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13361–13370. IEEE, 2025.

Hang Li, Chengzhi Shen, Philip Torr, Volker Tresp, and Jindong Gu. Self-discovering interpretable diffusion latent directions for responsible text-to-image generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12006–12016. IEEE, 2024.

Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Evaluating societal representations in diffusion models. *Advances in neural information processing systems*, 36:56338–56351, 2023.

Harvey Mannering. Analysing gender bias in text-to-image models using object detection. *arXiv preprint arXiv:2307.08025*, 2023.

Pablo Marcos-Manchón, Roberto Alcover-Couso, Juan C SanMiguel, and Jose M Martínez. Open-vocabulary attention maps with token optimization for semantic segmentation in diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9242–9252. IEEE, 2024.

Ranjita Naik and Besmira Nushi. Social biases through the text-to-image generation lens. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 786–808, 2023.

Rishubh Parihar, Abhijnya Bhat, Abhipsa Basu, Saswat Mallick, Jogendra Nath Kundu, and R Venkatesh Babu. Balancing act: Distribution-guided debiasing in diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6668–6678. IEEE, 2024.

Preethi Seshadri, Sameer Singh, and Yanai Elazar. The bias amplification paradox in text-to-image generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6367–6384, 2024.

Yingdong Shi, Changming Li, Yifan Wang, Yongxiang Zhao, Anqi Pang, Sibei Yang, Jingyi Yu, and Kan Ren. Dissecting and mitigating diffusion bias via mechanistic interpretability. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8192–8202. IEEE, 2025.

Agamdeep Singh, Ashish Tiwari, Hosein Hasanbeig, and Priyanshu Gupta. Distributions in, distributions out: The case for soft-label training. 2025. URL `https://api.semanticscholar.org/CorpusID:283080357`.

Luhang Sun, Mian Wei, Yibing Sun, Yoo Ji Suh, Liwei Shen, and Sijia Yang. Smiling women pitching down: auditing representational and presentational gender biases in image-generative ai. *Journal of Computer-Mediated Communication*, 29(1):zmad045, 2024.

Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Türe. What the daam: Interpreting stable diffusion using cross attention. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5644–5659, 2023.

Michael Toker, Hadas Orgad, Mor Ventura, Dana Arad, and Yonatan Belinkov. Diffusion lens: Interpreting text encoders in text-to-image pipelines. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Eddie L Ungless, Björn Ross, and Anne Lauscher. Stereotypes and smut: The (mis) representation of non-cisgender identities by text-to-image models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 7919–7942, 2023.

Jordan Vice, Naveed Akhtar, Richard Hartley, and Ajmal Mian. Exploring bias in over 100 text-to-image generative models. *arXiv preprint arXiv:2503.08012*, 2025.

Yixin Wan, Arjun Subramonian, Anaelia Ovalle, Zongyu Lin, Ashima Suvarna, Christina Chance, Hritik Bansal, Rebecca Pattichis, and Kai-Wei Chang. Survey of bias in text-to-image generation: Definition, evaluation, and mitigation. *arXiv preprint arXiv:2404.01030*, 2024.

Jialu Wang, Xinyue Liu, Zonglin Di, Yang Liu, and Xin Wang. T2iat: Measuring valence and stereotypical biases in text-to-image generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 2560–2574, 2023.

Wenxuan Wang, Haonan Bai, Jen-tse Huang, Yuxuan Wan, Youliang Yuan, Haoyi Qiu, Nanyun Peng, and Michael Lyu. New job, new gender? measuring the social bias in image generation models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 3781–3789, 2024.

Kyra Wilson, Sourojit Ghosh, and Aylin Caliskan. Bias amplification in stable diffusion's representation of stigma through skin tones and their homogeneity. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pp. 2705–2717, 2025.

Yankun Wu, Yuta Nakashima, and Noa Garcia. Stable diffusion exposed: Gender bias from prompt to image. In *Proceedings of the AAAI/ACM conference on AI, ethics, and society*, volume 7, pp. 1648–1659, 2024.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 3485–3492. IEEE, 2010.

Yanzhe Zhang, Lu Jiang, Greg Turk, and Diyi Yang. Auditing gender presentation differences in text-to-image models. In *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pp. 1–10, 2024.

Mi Zhou, Vibhanshu Abhishek, Timothy Derdenger, Jaymo Kim, and Kannan Srinivasan. Bias in generative ai. *arXiv preprint arXiv:2403.02726*, 2024.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

SUPPLEMENTARY MATERIAL

A LIMITATIONS

First, we attempted to debias generation more directly by backpropagating through ADS-Native to optimize a prompt's embedding until it reached a balanced predicted distribution for a given attribute, hoping this would translate into a balanced distribution of generated images; instead, the optimized embedding simply fools the surrogate itself, and the realized image distribution does not change, indicating that better strategies are needed to turn the surrogate's predicted distribution into an actual debiasing signal. Second, while the surrogate performs well on both gender and race, its race results are comparatively more modest, suggesting the approach may need further tuning as the number of attribute categories grows. Third, we describe ADS-External as black-box since it supplies its own text encoder, but we do not empirically validate this on a genuinely closed, API-gated T2I model, due to the cost of doing so at scale. Fourth, our ground truth relies on YOLOv10 and InternVL-2; although both have been validated, any bias they carry propagates into our labels. Finally, our cross-model analysis (Sections C.4 and 5.2) is necessarily limited to the three T2I backbones we evaluate; whether the specific patterns we observe (e.g., FLUX.2-klein's flatter bias distribution, Section C.3) generalize to other model families is an open question.

B ADDITIONAL METHODOLOGY DETAILS

B.1 PROMPT CONSTRUCTION VOCABULARY

Descriptive prompts. Human Caption-HQ's raw captions skew long and descriptive (e.g., "*Visible in the background is a ski slope. This individual is wearing a necktie...*"), so we downweight the longest captions when sampling (character-length-inverse weighting for 90% of the draw, uniform for the rest) to retain a meaningful share of shorter ones; FaceCaption-15M is predominantly short by comparison.

Captions from both sources routinely use gendered nouns, pronouns, and appearance cues (*man*, *she*, *father*, *beard*, *dress*), so we pass every sampled caption through a two-stage rewriting script. First, a regex-based pass neutralizes gendered language on 90% of prompts (*he/she* $\rightarrow$ *they/the person*; *father/mother* $\rightarrow$ *parent*; *boy/girl* $\rightarrow$ *child*), with a further pass on an independently sampled 40% genericizing explicit gendered appearance cues (*beard* $\rightarrow$ *facial features*, *dress* $\rightarrow$ *clothing*, *makeup* removed). We keep gendered pronouns in the remaining 10% for diversity and to help the model generalize to explicitly gendered prompts.

The same script enriches a subset of the neutralized captions with templated context: 50% receive a randomly sampled scene setting (drawn from a $\sim$425-category vocabulary derived from the SUN scene taxonomy (Xiao et al., 2010)) plus a randomly assigned occupation, activity, and object, and an independent 35% receive a demographic or cultural descriptor (region, religion, or ethnicity, each from a fixed vocabulary), all appended in random order. After splitting, the number of prompts with explicit gender/ethnic cues is nearly balanced across the training set.

Demographic and cultural descriptors. Table 4 lists the six template patterns used to inject a region, ethnicity, or religion descriptor into 35% of `Descriptive` prompts, together with the values each one draws from.

Gender-neutralization replacement rules. Table 5 lists the regex-based replacement rules applied when neutralizing gendered language in sampled captions (Section 3.2); each rule is applied independently at the stated sampling rate, so a given caption may receive several substitutions.

Scene-setting vocabulary. 50% of `Descriptive` prompts receive a randomly sampled scene setting, drawn from 425 SUN-derived (Xiao et al., 2010) scene categories and inserted via one of six opener templates (e.g., "*Visible in the background is a/an* {}.", "*Taking place at a/an* {}.", "*Located in a/an* {}."); Table 6 shows a representative alphabetical sample, since listing all 425 is impractical.

Templated-prompt vocabulary. *Templated* prompts (Section 3.2) vary along two independent axes: the neutral subject noun, one of {*person*, *individual*, *friend*, *colleague*, *someone*}, and the underlying concept, drawn from the four-way taxonomy of Girrbach et al. (2025) summarized in Table 7.

Dropping the *Uncertain* Labels. This discard-and-renormalize step could bias $P_G(x)$ if uncertainty tracked specific demographic presentations rather than detection quality; empirically, how-

Table 4: Demographic/cultural descriptor injection templates and their vocabularies used while constructing *Descriptive* prompts.

Template	**Vocabulary**
Ethnicity	
"*The person appears to be of* {} *descent.*"	Asian, Black, Hispanic, Indigenous, Middle Eastern, Mixed, South Asian, White
"*Showcasing* {} *cultural features.*"	(same 8 values)
Region	
"*They are native to* {}."	Africa, Asia, Europe, Middle East, North America, Oceania, South America, Southeast Asia
"*The image features a person originating from* {}."	(same 8 values)
Religion	
"*They might be associated with* {}."	Buddhism, Christianity, Hinduism, Islam, Judaism, Secular/Non-religious, Sikhism
"*Reflecting* {} *traditions.*"	(same 7 values)

Table 5: Gender-neutralization replacement rules applied to sampled *Descriptive* captions.

Original	**Replacement**	**Applied to**
he / she	they / the person	90% of prompts
father / mother	parent	90% of prompts
boy / girl	child	90% of prompts
beard	facial features	40% of prompts
dress	clothing	40% of prompts
makeup	(removed)	40% of prompts

ever, uncertain labels are concentrated in small, secondary detections (median bounding-box area 8–28$\times$ smaller than confidently-classified crops) and rarely affect a prompt's focal subject (uncertain for only 0.5–14% of the largest detection per image, vs. 22–27% pooled across all detections), so $P_G(x)$ is best understood as the distribution over a model's *confidently-classifiable* demographic presentations, rather than a joint estimate that also captures how often it renders visually ambiguous individuals.

B.2 Image Budget: Diversity vs. Per-Prompt Variance

We generate only 1 image per prompt in `Train-Desc`, but 10 per prompt in `Train-Tmpl` and 40 per prompt for validation and test (Section 3.2). This is a deliberate diversity/variance tradeoff, not a shortcut. With a fixed generation budget, we can spend it two ways: fewer prompts with more images each, which lowers the variance of each prompt's ground truth but covers less ground overall, or more prompts with fewer images each, which covers far more ground but makes each prompt's ground truth noisier. Since every `Train-Desc` prompt is a unique caption, we choose the second option and spend the budget on covering more prompts and contexts, as training benefits more from seeing diverse prompts than from a very clean label on each one. Similar prompts also share signal, and a single image often contains several people, so even one image per prompt already yields a distribution. Validation and test splits on the other hand need a stable, low-variance ground truth to compare against, so they use many more images per prompt instead.

B.3 Baseline Method Details

Section 4.1 introduces our four baselines briefly; we give the full formulation for each here.

Text Embedding Probe. Let $h(x) \in \mathbb{R}^d$ denote the mean-pooled output of the target model's own text encoder $\mathcal{T}$ for x, with $\mathcal{T}$ held entirely *frozen*, no LoRA, no gradient updates whatsoever. The same MLP head as ADS-Native maps this directly to a prediction, $\hat{P}(x)$. This is architecturally identical to our ADS-Native surrogate; the comparison isolates the effect of our curated training

Table 6: Representative sample of the 425-category scene-setting vocabulary, used while constructing the *Descriptive* prompt corpus.

Sampled scene categories
abbey, athletic_field_outdoor, bank_indoor, beach_house, booth_indoor, cafeteria, chalet, cockpit, creek, dining_hall, elevator_door, fire_station, game_room, hangar_outdoor, hunting_lodge_outdoor, kasbah, lighthouse, military_hospital, nursery, parking_garage_indoor, plaza, raceway, rock_arch, shower, street, tennis_court_outdoor, tree_house, waiting_room, wine_cellar

Table 7: Concept taxonomy underlying `Templated` prompts, from Girrbach et al. (2025).

Mode	**# Concepts**	**Example**
Activity	1,405	"*is going for a run*"
Context	737	"*at a job interview*"
Occupation	575	"*working as a veterinary assistant*"
Object	500	"*and a tie*"

data specifically. We report two variants that differ only in training data: trained on a generic, off-the-shelf prompt set (*Text Embedding Probe*), or on our own curated dataset (Section 3.2) instead (*ADS-Native*, Table 1), isolating training-data quality from encoder choice. We construct the generic dataset from prompts in the *Descriptive* split, explicitly containing pronouns (he/she/his/her), and label them accordingly. We only consider prompts that describe one individual, so we are certain that the generated image contains one individual whose gender is depicted by the pronouns used in the prompt. We further validate through the generated images that the models have good prompt adherence capability.

RepE Reading Vector (Zou et al., 2023). A single *reading vector* $v \in \mathbb{R}^d$ is computed once, offline: we embed many contrastive prompt pairs (x^+, x^-) (e.g., "a photo of a male plumber" vs. "a photo of a female plumber") in the same pooled representation space $h(x)$ as above, collect their differences $h(x^+) - h(x^-)$, and take the first principal component of these differences as v. At inference, a novel prompt's representation is projected onto this fixed direction and mapped to a probability by logistic regression, $\hat{p}_{\text{male}}(x) = \sigma(a\langle h(x), v\rangle + b)$, $\hat{P}(x) = [\hat{p}_{\text{male}}(x), 1 - \hat{p}_{\text{male}}(x)]^\top$, with scalars a, b fit on a held-out calibration set.

InterpretFlux (Li et al., 2024). This baseline operates outside the text-embedding space entirely: a concept direction $u \in \mathbb{R}^{d_z}$ is discovered within the *diffusion transformer's own* intermediate activation space z (rather than $\mathcal{T}$'s), by training a small MLP that injects directional perturbations along candidate directions into z and optimizes the flow-matching denoising objective G was originally trained with. At inference, the alignment of u with G's prompt-conditioned intermediate activations for x is converted to a probability via cosine similarity. Because this requires running (part of) G itself at inference time, InterpretFlux is, like the Diffusion Activation Probe, a white-box, diffusion-internal baseline. The original authors developed this pipeline for Stable Diffusion, which has a semantic *h-space*; we find a similar space in FLUX just after the dual-stream blocks and first validate the original approach there before constructing our baseline.

Diffusion Activation Probe (Parihar et al., 2024). A one-hidden-layer MLP (width 512, ReLU), $\hat{P}(x) = \text{Softmax}(\text{MLP}(h_{\text{diff}}(x)))$, where $h_{\text{diff}}(x) = \text{pool}(Z_G(x))$ a pooled activation collected from G's own early transformer blocks (or U-Net blocks) at inference time, at exactly midway of the generation trajectory ($t = T/2$), rather than from its text encoder. This is then trained on the same generic dataset as Text Embedding Probe. This tests whether G's internal representations carry demographic signal beyond what its text encoder alone provides; like InterpretFlux, it requires invoking G itself at inference, forfeiting the black-box property our External surrogate retains.

B.4 Surrogate Configuration

ADS-External encodes prompts with `google/t5-v1_1-xxl` (∼5B params, 4-bit NF4 double-quantized, bfloat16), adapted via LoRA on query/value projections (r=16, α=32, dropout 0.1); trained with AdamW (learning rate 2×10^{-4}, weight decay 0.01, batch 32, 10 epochs, grad-clip 1.0, gradient checkpointing, max sequence length 512), selecting the lowest-validation-KL checkpoint. ADS-Native reuses the same MLP head and data but uses the target model's own, unadapted encoder

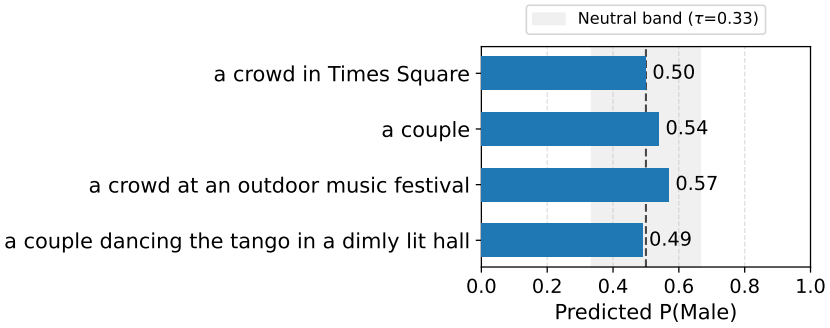


Figure 8: FLUX.2-klein's predicted $P(\text{Male})$ for *couple*/*crowd* prompts, all classified Neutral.

(no LoRA); trained with AdamW (sweep-selected hyperparameters, batch 256, up to 500 epochs, patience 30), again by lowest validation loss. The MLP head is Linear(d, 512)–LayerNorm–ReLU–Dropout–Linear(512, 256)–LayerNorm–ReLU–Dropout–Linear(256, K), with dropout 0.1, K=2 for gender, and K=6 for race.

C Additional Results

C.1 Threshold (τ) Ablation

Section 4.2 uses a fixed threshold $\tau = \frac{1}{3}$ to decide if a distribution counts as Male-Biased, Female-Biased, or Neutral. We chose this value a priori, because it splits $[0, 1]$ into equal thirds, and did not tune it on our results. Changing τ does change accuracy: as τ approaches 0, the neutral band covers almost the whole range and nearly every prompt becomes Neutral, while as τ approaches $\frac{1}{2}$, the band disappears. Neither extreme is inherently better, so we keep $\tau = \frac{1}{3}$, which leaves a wide, meaningful neutral band.

C.2 Encoder-Scale and Training-Data Ablation

Table 8 isolates encoder scale for gender, by comparing our T5-XXL surrogate against otherwise-identical T5-Base/T5-Large LoRA variants, against a no LoRA variant but using T5-XXL and against ADS-Native.

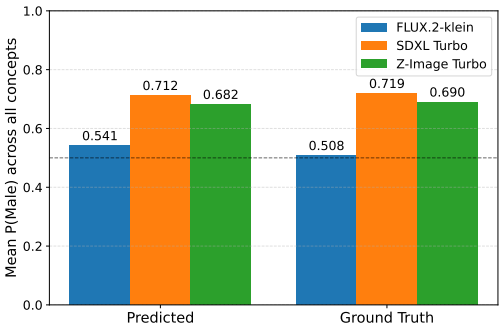
Encoder scale generally helps, but not consistently: T5-XXL is the best or tied-best encoder on accuracy and Spearman correlation in 9 of 10 block–metric pairs, though T5-Large sometimes does better on KL divergence (e.g., 0.1179 vs. 0.1288 on FLUX.2-klein `Test-Desc`). So scaling the encoder helps more with getting the classification and ranking right than with fine-grained calibration. Separately, comparing the Text Embedding Probe in Table 1 and ADS-Native shows what actually drives the bigger improvement: training the same probe on our curated dataset instead of generic data raises accuracy substantially on its own (e.g., 0.53 $\rightarrow$ 0.79 on FLUX.2-klein `Test-Desc`). So much of what makes our dataset useful isn't specific to our surrogate's own architecture. It carries over directly to a much simpler probe trained on the same data.

C.3 Additional Cross-Model Experiments

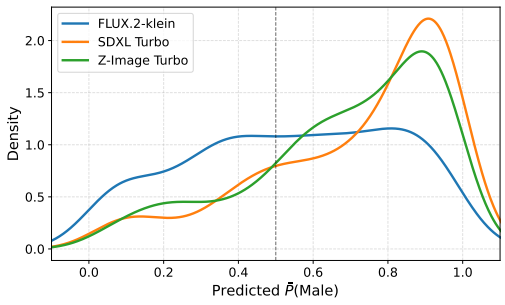
Concepts built around multiple people. The compositional explanation for neutral concepts (Section 5.2) motivates a smaller, targeted probe on FLUX.2-klein: ten prompts built around *couple* and *crowd*, two concepts that, like the neutral concepts in Section 5.2, place more than one person in the frame by construction rather than by incidental wording, e.g., "*a couple posing for an engagement photoshoot in a park*" and "*a crowd at an outdoor music festival*". Ground truth agrees with the surrogate's predictions here as well: every one of the ten prompts is classified Neutral, with predicted $P(\text{Male})$ ranging from 0.37 to 0.58, comfortably inside the neutral band (Figure 8). Both *couple* prompts and *crowd* prompts average almost exactly even (~ 0.50). This is a smaller, single-model check rather than the cross-model comparison used elsewhere in this section, but it reinforces the same explanation: the surrogate correctly predicts neutral specifically where the underlying scene

Table 8: Ablation over text-encoder scale: our full T5-XXL surrogate against smaller T5 Base / T5 Large LoRA variants, a variant without LoRA adaptation but still using T5-XXL and ADS-Native (the target model's frozen encoder trained on our curated data). KL Div. is the training objective (Equation (2)) evaluated on the test set. Best value per metric within this table, per model/test-set block, is in **bold**.

Target Model	Method	Acc. ↑	KL Div. ↓	TV Dist. ↓	Spearman ↑
Descriptive					
FLUX.2-klein	ADS-Native	0.7927	0.1533	0.1310	0.8659
	ADS-External w/o LoRA	0.7972	0.1484	0.1341	0.8619
	ADS-External (T5-Base)	0.8057	0.1394	0.1200	0.8775
	ADS-External (T5-Large)	0.8292	**0.1179**	0.1018	0.9021
	ADS-External (T5-XXL)	**0.8348**	0.1288	**0.0977**	**0.9073**
Z-Image Turbo	ADS-Native	0.8135	0.1919	0.1349	0.8329
	ADS-External w/o LoRA	0.8004	0.1979	0.1459	0.8240
	ADS-External (T5-Base)	0.8135	0.1746	0.1355	0.8448
	ADS-External (T5-Large)	0.8260	0.1710	0.1222	0.8579
	ADS-External (T5-XXL)	**0.8441**	**0.1506**	**0.1163**	**0.8657**
Templated					
FLUX.2-klein	ADS-Native	0.7355	**0.0746**	0.1187	0.8557
	ADS-External w/o LoRA	0.7020	0.0927	0.1340	0.8199
	ADS-External (T5-Base)	0.7200	0.0937	0.1332	0.8204
	ADS-External (T5-Large)	0.7310	0.0808	0.1211	0.8471
	ADS-External (T5-XXL)	**0.7505**	0.0784	**0.1138**	**0.8573**
SDXL Turbo	ADS-Native	**0.7880**	0.1144	0.1275	0.7588
	ADS-External w/o LoRA	0.7506	0.1289	0.1434	0.7172
	ADS-External (T5-Base)	0.7486	0.1350	0.1539	0.6838
	ADS-External (T5-Large)	0.7764	0.1149	0.1389	0.7281
	ADS-External (T5-XXL)	0.7698	**0.1031**	**0.1257**	**0.7815**
Z-Image Turbo	ADS-Native	0.6843	0.1553	0.1660	0.6797
	ADS-External w/o LoRA	0.6626	0.1646	0.1755	0.6459
	ADS-External (T5-Base)	0.6968	0.1620	0.1706	0.6500
	ADS-External (T5-Large)	0.6983	0.1460	0.1641	0.6910
	ADS-External (T5-XXL)	**0.7320**	**0.1325**	**0.1487**	**0.7286**



(a) Mean $\bar{P}$(Male) per model



(b) Predicted $\bar{P}$(Male) density per model

Figure 9: Aggregate bias level (left) and full distribution shape (right) per target model, across all the concepts (Templated).

itself contains a group of people and the context itself does not suggest the underlying gender distribution, rather than needing to resolve the gender of one ambiguous individual.

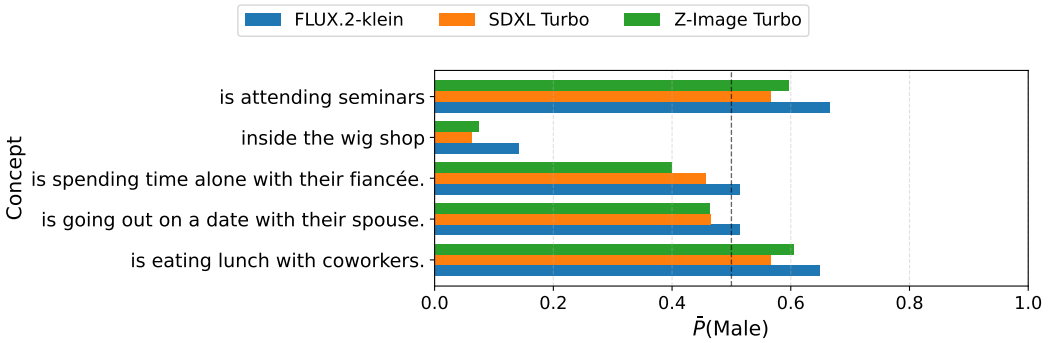


Figure 10: Concepts where FLUX.2-klein is the *most* male-biased of the three models, a genuine, if less frequent, counter-example to its general population-level tendency toward lower male bias.

FLUX.2-klein is systematically, but not uniformly, less male-biased. Figure 9 (left) shows the mean predicted $\bar{P}(\text{Male})$ across all concepts: 0.541 for FLUX.2-klein (ground truth 0.508), versus 0.712 for SDXL Turbo (0.719) and 0.682 for Z-Image Turbo (0.690). But this lower average isn't a uniform shift: Figure 9 (right) shows FLUX's density is much flatter across the whole $[0, 1]$ range, while SDXL and Z-Image both concentrate sharply around $\bar{P}(\text{Male}) \approx 0.9$.

However, this trend is not absolute, and does not hold uniformly at the level of individual concepts. Figure 10 shows five concepts (e.g., "*is attending seminars*", "*is eating lunch with coworkers*") where FLUX is in fact the *most* male-biased of the three models, a genuine, if less frequent, counter-example to its general tendency toward lower male bias. This confirms that "FLUX is less biased" is a population-level tendency with real, concept-specific exceptions, not a categorical property of the model itself.

C.4 Word-Level Bias Analysis

We repeat a version of the concept-level analysis (Section 5.2) at the level of individual words rather than templated concepts, using the long, free-form `Test-Desc` prompts. For every word appearing in at least 10 prompts and no more than 30% of all prompts, we compute the average predicted and ground-truth $P(\text{Male})$ over every prompt containing that word.

Words associated with each gender are semantically coherent and well-calibrated. On FLUX.2-klein `Test-Desc`, the words most associated with male-biased generations are physical/contextual descriptors such as *goatee*, *muscular*, *sideburns*, *mustache*, *necktie*, *bald*, and *basketball*; those most associated with female-biased generations include *feminine*, *mascara*, *eyeliner*, *eyeshadow*, *blush*, and *dresses* (Figure 11). Rankings by predicted and ground-truth average agree closely, giving an independent, word-level confirmation of surrogate fidelity beyond the prompt-level metrics of Section 4.2.

A small set of phrases flip bias direction entirely across models. We further identify phrases that are clearly male-associated (predicted $P(\text{Male}) \geq 0.65$) in at least one model *and* clearly female-associated in another (predicted $P(\text{Male}) \leq 0.35$). On `Test-Tmpl`, 18 such phrases exist, of which five are shown in Figure 12, and the pattern is strikingly one-sided: FLUX.2-klein is the clearly-female biased model and SDXL Turbo/Z-Image Turbo the clearly-male biased models for nearly every one of them (e.g., *credit analyst*: FLUX ground truth ≈ 0.10 vs. SDXL ≈ 0.90, Z-Image ≈ 0.80). This is a much sharper, word-level instance of the same model-dependent skew identified in Section 5.2, and pinpoints specific lexical triggers, not a diffuse, uniform difference, as major contributors to the aggregate gap between models.

Figure 13 shows the corresponding word ranking on the TIFA benchmark (Section 5.4).

C.5 Ground-Truth Validation of the Combinatorial Analysis

The combinatorial analysis of Section 5.3 (Figure 5) is computed entirely from the surrogate's own predictions, since exhaustively generating and labeling all 9,396 object–activity combinations would be prohibitively expensive under the traditional pipeline. To check that the combination rule it

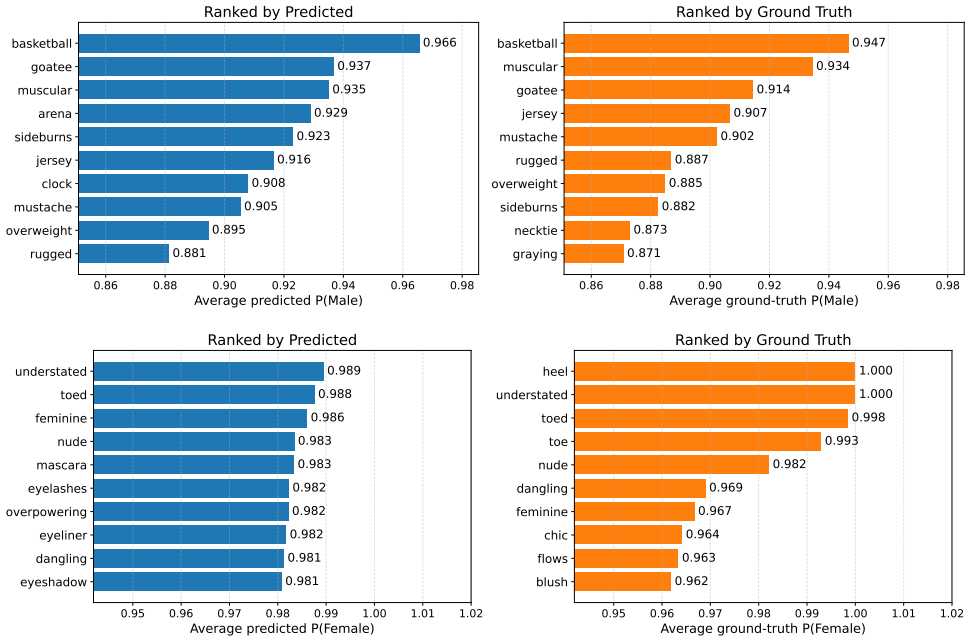


Figure 11: Top words by average predicted and ground-truth $P(\text{Male})$ over all FLUX.2-klein Descriptive-split prompts containing that word, for male-biased (top) and female-biased (bottom) generations. Close agreement between the predicted and ground-truth bars confirms surrogate fidelity down to the word level.

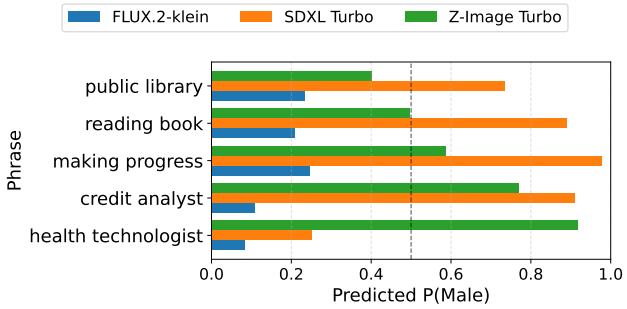


Figure 12: Phrases whose predicted gender association reverses sign across models: clearly male in one model, clearly female in another. FLUX.2-klein (blue) is the clearly-female model for nearly every flip phrase.

reveals also holds in reality, we validate it against real ground truth at a small scale. For each of the 3×3 (object bias class, activity bias class) cells, we classify every object and activity using *real* ground truth (FLUX.2-klein generations passed through the same detection-and-classification pipeline as Section 3.2, not the surrogate), construct every combined prompt for that cell, and, due to the cost of generating and labeling images at this scale, randomly sample only 10 prompts per cell (90 prompts total across all nine cells) rather than the full pool (ranging from 780 to 1,482 combinations per cell). We generate images for these 90 prompts, extract ground truth exactly as in Section 3.2, and compare the resulting outcome-class distribution against the surrogate's own prediction on the identical 10 prompts, cell by cell. Figure 14 shows both distributions side by side for every cell. The dominant outcome class agrees between prediction and ground truth in seven of the nine cells outright, is tied between two classes for the surrogate in an eighth, and disagrees in only one (Object: Female, Activity: Male), supporting that the combination-rule pattern identified through the surrogate is not an artifact of the surrogate itself.



(a) Top 5 male-biased words

(b) Top 5 female-biased words

Figure 13: Words most strongly associated with male- and female-biased predictions, pooled across all TIFA benchmark prompts (COCO, DrawBench, PartiPrompts, and PaintSkills combined).

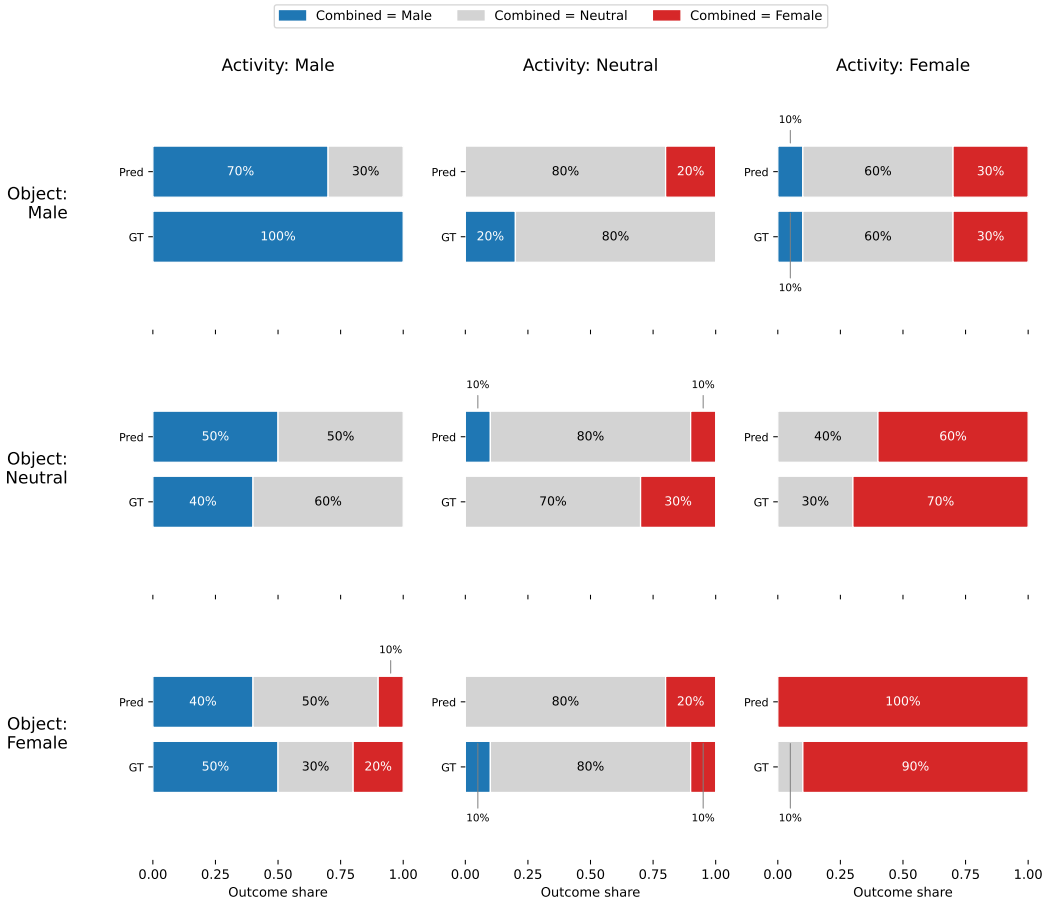


Figure 14: Predicted (top bar) vs. ground-truth (bottom bar) outcome-share breakdown for each of the nine object-class × activity-class cells, both computed on the same 10 real-FLUX-generation prompts per cell.

C.6 Ground-Truth Reliability: A Negative Result for DeepFace

We evaluated the surrogate's failure rate and flip catch rate (Section 5.5) on 1,200 steered generations (evenly split across intended directions), first using DeepFace to obtain $y_G(x)$, matching the



Figure 15: Example images where DeepFace and our YOLOv10 + InternVL-2 pipeline disagree on the perceived gender label (95/1,200 total disagreements); qualitatively, our pipeline is correct on most of these cases.

Table 9: Surrogate flip-diagnostics across ground-truth pipelines and correction stages, on the same 1,200 gender-steered generations.

Stage	**Failure rate** ↓	**Flip catch rate**	**#Flips (rate)** ↓
FairImagen, DeepFace labels	14%	33%	220/1,200 (18.33%)
FairImagen, our labels	8%	51%	163/1,200 (13.58%)
+ surrogate-guided reselection, our labels	5.92%	0%	71/1,200 (5.92%)

ground-truth convention of Fu et al. (2026). The failure rate of the surrogate, i.e., the cases where the surrogate's prediction disagreed with the ground truth label of the generated images was a reasonably low 14%, yet the flip catch rate was only 33%, an inconsistency, since a surrogate that agrees with the realized label 86% of the time overall should not simultaneously miss two-thirds of the specific cases where that label disagrees with the intended direction, unless the realized labels themselves are noisy. See Table 9. We tested this directly by re-labeling the same 1,200 images with our own detection-and-classification pipeline (YOLOv10 + InternVL-2, Section 3.2) and comparing the two label sets: they disagree on $95/1{,}200$ (7.9%) images. Qualitative inspection of these disagreements (Figure 15) suggests our pipeline is correct on most of them, which we further confirmed by manually validating the labels of all non-overlapping images and a subset of the overlapping ones. Thus, DeepFace, the classifier standard in this line of debiasing work, appears to mislabel gender at a rate high enough to materially distort flip-based diagnostics.

Recomputing both diagnostics with the corrected ground truth, the failure rate drops to 8% and the flip catch rate rises to 51%: most of the apparent unreliability in the surrogate's flip-anticipation was ground-truth noise, not a genuine limitation of f_ϕ itself.

C.7 Race Distributions

Figures 16 to 18 extend Figure 3's density-alignment comparison to the race instantiation of the surrogate (K=6): for each target model, every panel overlays the ground-truth density, ADS-Native's predicted density, and the Diffusion Activation Probe's density for one race category (`Test-Desc` and `Test-Tmpl` pooled where both splits exist; SDXL Turbo has `Test-Tmpl` only). Consistent with the race results in Table 2, ADS-Native's density tracks the shape of the ground-truth curve far more closely than the baseline across categories and models, matching the large gap in entropy correlation reported there. The baseline's predictions carry little information about the true distribution's shape, while ADS-Native's do.

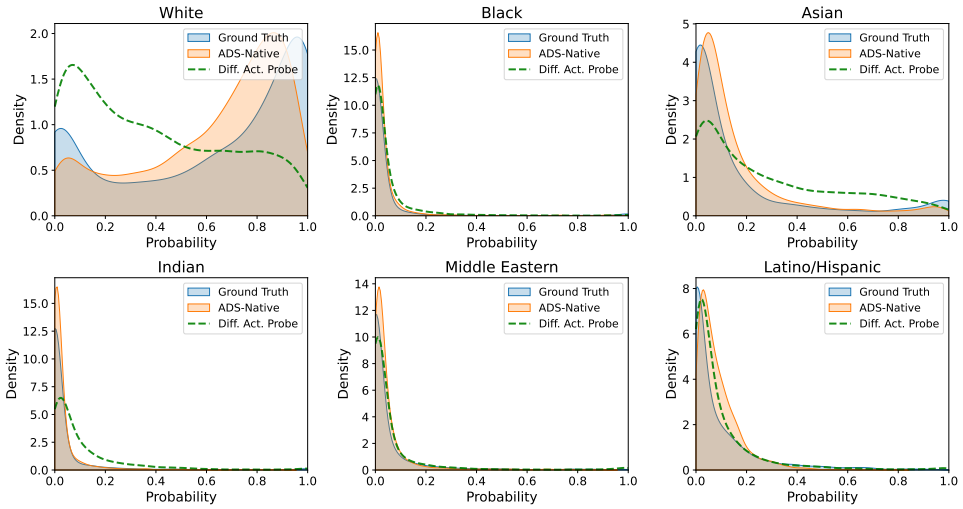


Figure 16: Predicted vs. ground-truth density alignment for the race instantiation of the surrogate (K=6) on FLUX.2-klein, analogous to Figure 3 for gender. Each panel additionally shows the Diff. Act. Probe baseline (dashed green), the strongest baseline in Table 1.

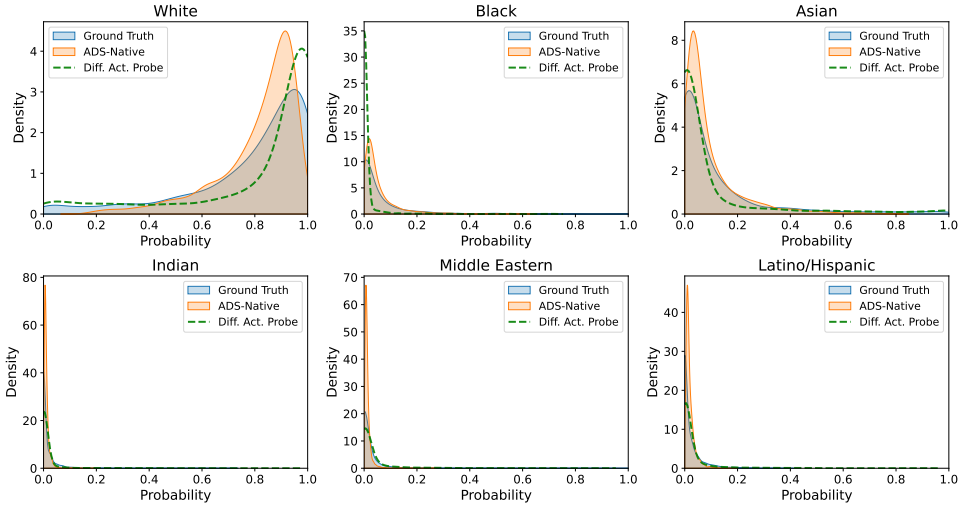


Figure 17: The same comparison as Figure 16, for SDXL Turbo. SDXL Turbo has no `Test-Desc` split, so this reflects `Test-Tmpl` only.

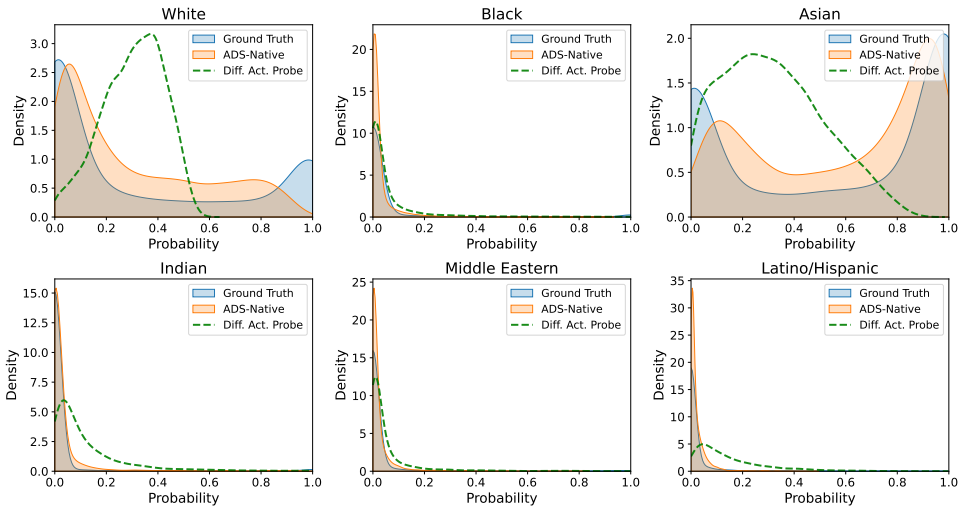


Figure 18: The same comparison as Figure 16, for Z-Image Turbo.