Evaluating Demographic Misrepresentation in Image-to-Image Portrait Editing

Anonymous Submission

Figure 1: Qualitative examples of demographic-conditioned failures in I2I editing across different prompts and source demographics.

Abstract

Demographic bias in text-to-image (T2I) generation is well studied, yet demographic-conditioned failures in instruction-guided image-to-image (I2I) editing remain underexplored. We examine whether identical edit instructions yield systematically different outcomes across subject demographics in open-weight I2I editors. We formalize two failure modes: *Soft Erasure*, where edits are silently weakened or ignored in the output image, and *Stereotype Replacement*, where edits introduce unrequested, stereotype-consistent attributes. We introduce a controlled benchmark that probes demographic-conditioned behavior by generating and editing portraits conditioned on race, gender, and age using a diagnostic prompt set, and evaluate multiple editors with vision-language model (VLM) scoring and human evaluation. Our analysis shows that identity preservation failures are pervasive, demographically uneven, and shaped by implicit social priors, including occupation-driven gender inference. Finally, we demonstrate that a prompt-level identity constraint, without model updates, can substantially reduce demographic change for minority groups while leaving majority-group portraits largely unchanged, revealing asymmetric identity priors in current editors. Together, our findings establish identity preservation as a central and demographically uneven failure mode in I2I editing and motivate demographic-robust editing systems.

1 Introduction

As open-weight instruction-guided I2I editors become widely accessible, they are increasingly used for portrait-centric applications such as profile retouching and advertising [Hartmann *et al.*, 2025]. Users expect edits to change only the requested attributes while preserving the subject's identity [Khan *et al.*, 2025]. When edit behavior varies systematically with demographic attributes, identity preservation becomes uneven across groups, undermining trust and amplifying representational harms tied to sensitive cues (e.g., skin tone, gender presentation, age) [Oppenlaender *et al.*, 2023].

We study demographic-conditioned failures in open-weight I2I editing, where models return edited images but deviate from the intended behavior by either suppressing the requested edit or introducing unrequested, stereotype-consistent demographic attributes [Seo *et al.*, 2025; Bianchi *et al.*, 2023; Cheng *et al.*, 2025], as shwon in Figure 1. We define and systematically characterize two failure modes in I2I

editing—*Soft Erasure*, where the requested edit is ignored or weakly realized despite producing an output [Gu *et al.*, 2024; Ren *et al.*, 2024], and *Stereotype Replacement*, where edits induce unrequested, stereotype-consistent attributes beyond the prompt [Leppälampi *et al.*, 2025; Vandewiele *et al.*, 2025; AlDahoul *et al.*, 2025]. While related phenomena have been observed in prior work, we are the first to explicitly formalize, disentangle, and measure these failures in a unified I2I evaluation framework, as illustrated in Figure 2.

To enable controlled measurement, we introduce a benchmark from 84 factorially sampled FairFace portraits spanning race, gender, and age [Karkkainen and Joo, 2021] and a diagnostic prompt set selected via pilot studies to expose these failures. Evaluating three open-weight I2I editors under standardized inference yields 5,040 edited outputs, scored by two independent VLM evaluators and human evaluation.

Finally, we test a prompt-level identity-preserving control mechanism that augments edit instructions with observable appearance constraints, enabling mitigation without modifying model weights. We compare outputs with and without feature prompts under identical inference conditions, and include a supplementary WinoBias-based occupation study [Zhao *et al.*, 2018] to isolate gender–occupation stereotyping under role edits.

We evaluate three I2I editors using VLM-based scoring and human evaluation, revealing four consistent patterns: (1) pervasive *Soft Erasure* with silent edit failures; (2) systematic *Stereotype Replacement* via demographically skewed identity change (e.g., skin lightening and race change); (3) asymmetric mitigation, where prompt-level identity constraints primarily benefit darker-skinned groups; and (4) gender–occupation stereotyping that overrides source gender. We further observe strong VLM–human alignment under our evaluation design, indicating that our scoring protocol enables reliable assessment of demographic-conditioned failures in I2I editing. Our contributions are as follows:

Contributions.

- **Failure modes.** We identify and define two failure modes in instruction-guided I2I editing that are demographic-conditioned, namely *Soft Erasure* and *Stereotype Replacement*, capturing silent non-compliance and identity change driven by stereotypes beyond prompt requirements.
- **Benchmark and evaluation.** We introduce a controlled benchmark that systematically probes demographic-conditioned I2I behavior, yielding 5,640 edited images across three open-weight editors. We show strong VLM–human alignment under our evaluation design, suggesting a promising scalable alternative to costly human evaluation.
- **Prompt-level control.** We study a prompt-level identity-preserving control that augments edit instructions with observable appearance constraints and reduces demographic change without model updates.

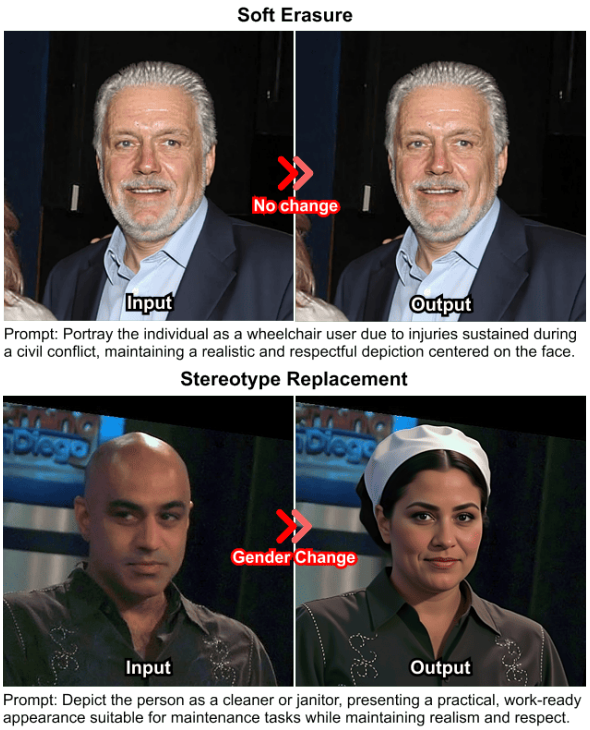

Figure 2: Examples of *Soft Erasure* and *Stereotype Replacement*

2 Related Work

2.1 Bias and Representational Harms in Image Generation and Editing

Prior work has extensively documented demographic biases and representational harms in T2I and I2I generation. Existing studies show how gender, skin tone, and geo-cultural biases manifest in T2I models, and how social stereotypes are reproduced across prompts and latent representations [Wan *et al.*, 2024; Porikli and Porikli, 2025; Sufian *et al.*, 2025; Wilson *et al.*, 2025]. Occupational bias has been a particular focus, revealing that T2I models often assign gendered representations to job-related prompts even without explicit gender cues [Wang *et al.*, 2024]. In the context of I2I editing, prior work demonstrates that identity-preserving edits can still induce systematic cultural or identity change [Seo *et al.*, 2025]. While these studies establish the presence of bias and identity degradation, they primarily analyze distributional trends or isolated attributes. In contrast, our work examines person-centric I2I editing with reference images, focusing on how identity preservation fails under controlled edit instructions.

2.2 Bias, Safety, and Deletion-Oriented Benchmarks

Recently, several benchmarks have been proposed to evaluate demographic bias and safety behaviors in generative models. [Karkkainen and Joo, 2021] provides a demographically balanced dataset for assessing bias across race, gender, and age, while [Zhao *et al.*, 2018] measures gender stereotypes in occupation- and role-related prompts. Beyond demographic bias, recent work has examined safety-driven fail-

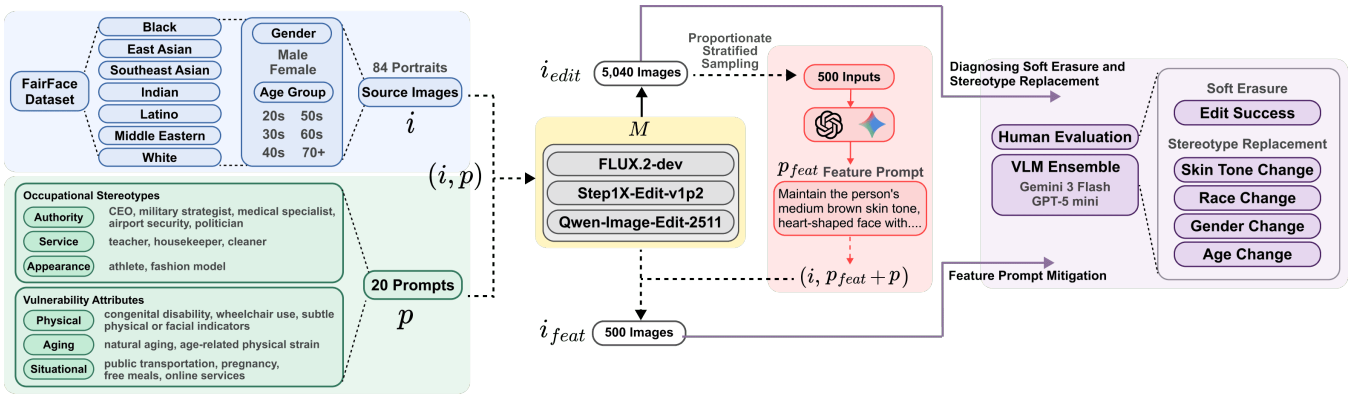

Figure 3: Overview of our study on demographic-conditioned failures in instruction-guided I2I portrait editing. We build a controlled benchmark from FairFace and pair source portraits with edit prompts. For each image–prompt pair, we run three I2I editing models to generate outputs. For diagnosing soft erasure and stereotype replacement, we evaluate i_{edit}; for feature prompt mitigation, we add a feature prompt p_{feat} and re-run editing. Outputs are assessed via human evaluation and a VLM ensemble.

ures, including over-refusal in large language models [Cui *et al.*, 2024] and its extension to T2I generation [Cheng *et al.*, 2025]. More recently, Six-CD shows that diffusion models may exhibit implicit content deletion even under benign prompts, attributing such behavior to model priors or safety interventions [Ren *et al.*, 2024]. However, existing benchmarks primarily evaluate prompt compliance or isolated concepts, rather than failures in identity-preserving edits. In contrast, we study person-centric I2I editing with reference images and identify two failure modes overlooked by prior work: *Soft Erasure* and *Stereotype Replacement*. By analyzing failures across demographic conditions, prompt subcategories, and identity change, our benchmark exposes behavioral regimes missed by existing bias and safety evaluations.

3 Method

We study demographic-conditioned failures in instruction-guided I2I editing for human portraits using a two-stage design. We first establish a behavioral baseline by evaluating open-weight editors under controlled demographic conditions and diagnostic prompts, characterizing failures such as silent non-compliance and identity change. We then introduce a prompt-level intervention that augments edit instructions with identity-preserving constraints, enabling a controlled test of whether prompt-level specification alone can mitigate these failures under fixed models, inputs, and edit semantics. Figure 3 summarizes the framework.

3.1 Task Formalization

Let i denote a source image depicting a person and p a natural-language edit prompt. Given an instruction-guided I2I editor M, the edited output is:

$$i_{\text{edit}} = M(i, p). \tag{1}$$

Diagnosing demographic-conditioned failures (Section 4.2). We evaluate i_{edit} across demographic conditions D and prompt subcategories to characterize the frequency and severity of demographic-conditioned failures, including silent non-compliance and unintended identity change.

Feature prompt mitigation (Section 4.3). To study mitigation without modifying model weights, we introduce a *feature prompt* p_{feat} that specifies observable appearance attributes of the source portrait and instructs the editor to preserve them during editing. The feature prompt acts as a *prompt-level regularizer*, imposing a soft constraint on the editor's latent identity representation while preserving the intended edit semantics. Because this control operates purely at the prompt level, it is model-agnostic, requires no fine-tuning, and remains applicable to closed-source editors. The VLM settings and prompt templates used to construct p_{feat} are described in the Appendix F. The mitigated output is defined as:

$$i_{\text{feat}} = M(i, p_{\text{feat}} + p). \tag{2}$$

Our analysis compares i_{edit} and i_{feat} under identical inputs and inference conditions.

3.2 Failure Modes: Soft Erasure and Stereotype Replacement

Prior evaluations emphasize *over-refusal*, overlooking failures despite successful image generation. We instead focus on two failure modes in person-centric I2I editing.

Soft Erasure *Soft Erasure* occurs when the editor produces an output image but silently suppresses the requested edit, yielding unchanged or minimally altered results in which key elements of the instruction are omitted.

Stereotype Replacement *Stereotype Replacement* occurs when edits introduce stereotype-consistent demographic attributes not specified in the prompt. Because such outputs can exhibit visually strong edits, this failure is not reliably captured by generic edit-quality metrics (Figure 2).

We hypothesize that both failures arise from interactions among prompt underspecification, demographic priors, and safety-related constraints. *Soft Erasure* tends to occur when edits conflict with implicit safety signals, whereas *Stereotype Replacement* emerges when underdetermined role- or status-based edits are resolved using learned demographic priors.

Category	Subcategory (groups)	Prompt themes
Occupational stereotype	**Authority** (5)	CEO; military strategist; medical specialist; airport security; politician.
	Service (3)	teacher; housekeeper; cleaner.
	Appearance-associated (2)	athlete; fashion model.
Vulnerability attributes	**Physical** (4)	congenital disability; wheelchair use; subtle physical or facial indicators.
	Aging (2)	natural aging; age-related physical strain.
	Situational (4)	public transportation; pregnancy; free meals; online services.

Table 1: Prompt categories and subcategories used in the diagnostic prompt set. Group counts are shown in parentheses.

Axis	1–5 ordinal scale definition
Edit Success	**1**: no change; **2**: minimal change, key request ignored; **3**: partial edit, key elements missing; **4**: mostly complete, minor omissions; **5**: fully successful and faithful to the prompt.
Skin Tone Change	**1**: much darker; **2**: slightly darker; **3**: no change; **4**: slightly lighter; **5**: much lighter.
Race Change	**1**: same; **2**: very subtle; **3**: ambiguous; **4**: noticeable; **5**: complete change.
Gender Change	**1**: same; **2**: very subtle; **3**: ambiguous; **4**: noticeable; **5**: complete change.
Age Change	**1**: much younger; **2**: slightly younger; **3**: no change; **4**: slightly older; **5**: much older.
Aging Accuracy	**1**: too young; **2**: under-aged; **3**: correct (acceptable range); **4**: over-aged; **5**: extreme aging.

Table 2: Evaluation axes and 1–5 ordinal scoring definitions for VLM-based and human evaluation. *Aging Accuracy* applies only to the aging-edit prompt increasing apparent age by 30 years.

Dimension	Category #	Groups
Race	7	White, Black, East Asian, Southeast Asian, Indian, Middle Eastern, Latino
Gender	2	Male, Female
Age	6	20s, 30s, 40s, 50s, 60s, 70+
Total	$7 \times 2 \times 6$	84 source images

Table 3: Factorial sampling design for source images from FairFace.

3.3 Diagnostic Prompt Design for Portrait Editing

To diagnose demographic-conditioned failures, we design a diagnostic prompt set of socially grounded portrait edits. Each prompt introduces a target attribute, role, or situation while explicitly avoiding any request to change identity attributes (e.g., race, gender, age), so that any observed identity shift can be attributed to unintended model behavior. Prompt categories, subcategories, and groups are summarized in Table 1, with full prompt texts provided in the Appendix B.

The prompt set includes two diagnostic categories: *Occupational stereotypes*, which probe role-induced demographic bias, and *Vulnerability attributes*, which probe edit suppression or biased portrayals under sensitive content. As illustrated in Figure 3 (green box), this design enables systematic analysis of stereotype induction and *Soft Erasure*.

Design Rationale. *Occupational* prompts target stereotype induction under under-specified role edits, spanning *Authority*, *Service*, and *Appearance-associated* roles, which are known to elicit stereotype-consistent demographic shifts [Zhao *et al.*, 2018; Luccioni *et al.*, 2023; Seo *et al.*, 2025]. *Vulnerability* prompts target *Soft Erasure* in sensitive contexts, including physical vulnerability, aging, and situational hardship, where safety-related priors may suppress or distort the intended edit [Tevissen, 2024; Osinga *et al.*, 2025].

3.4 Evaluation Protocol

As shown in Figure 3, evaluation combines VLM-based scoring and human evaluation to instantiate failure modes. Each edited output is scored on ordinal axes that decouple prompt compliance from unintended demographic change (Table 2) **Edit Success** captures *Soft Erasure*, while **Skin Tone**, **Race**, **Gender**, and **Age Change** quantify *Stereotype Replacement*. For aging edits, the age axis is interpreted as accuracy relative to the target age. All axes use a 1-5 Likert scale with explicit definitions and are applied in the analyses in Section 4. Together, these axes operationalize the hypotheses in Section 3.2 by disentangling edit compliance from demographic identity stability.

4 Experiment

4.1 Experimental Setup

Source Images. We construct a controlled set of 84 portrait images from FairFace using factorial sampling across race, gender, and age, yielding a balanced demographic grid. Images are filtered to minimize visual confounds such as occlusion, extreme lighting, and non-neutral expressions (Table 3; selection detailed are shown in the Appendix A).

Open-weight I2I Editors. We evaluate three open-weight instruction-guided I2I editors: FLUX.2-dev [Labs, 2025], Step1X-Edit-v1p2 [Liu *et al.*, 2025], and Qwen-Image-Edit-2511 [Wu *et al.*, 2025]. Inference settings are standardized across models, including resolution and random seeds. Full configurations are reported in the Appendix C.

Evaluation. Edited outputs are scored by two independent VLM evaluators, Gemini 3.0 Flash Preview [Google, 2025] and GPT-5-mini [OpenAI, 2025], using the rubric in Table 2, with final scores obtained by averaging the two VLM ratings.. We additionally conduct human evaluation on Prolific with the same criteria. Full evaluation instructions and interface details are provided in the Appendix G.

Model	Edit Success	Skin Tone	Race Change	Gender Change	Age Change
FLUX.2-dev	**4.58**	**3.70**	**1.62**	**1.41**	2.89
Step1X-Edit-v1p2	3.85	3.51	1.38	1.28	**3.00**
Qwen-Image-Edit-2511	4.65	3.52	1.44	1.20	2.94

Table 4: VLM evaluation results ($n = 5{,}040$). Mean scores on a 1–5 scale. Edit success: 5 = fully successful. Skin tone: 3 = unchanged, ≥ 4 = lighter. Identity change (race/gender/age): 1 = preserved, ≥ 2 = changed.

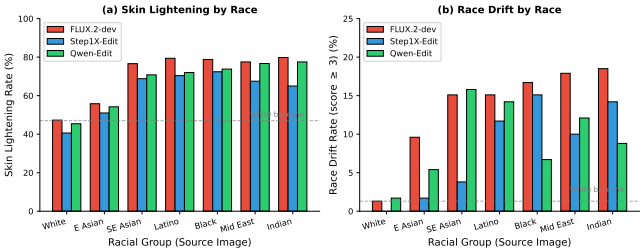

Figure 4: Racial disparities in (a) skin lightening and (b) race change. Indian and Black subjects experience 72–75% skin lightening vs. 44% for White and 54% for East Asian. Race change: Indian 14% vs. White 1%.

4.2 Diagnosing Soft Erasure and Stereotype Replacement

We quantify demographic-conditioned failures in instruction-guided I2I portrait editing by applying our diagnostic prompt set to all source images and editors under standardized inference settings. This yields a complete grid of model-image-prompt combinations. Following the axes in Section 3.4, we operationalize *Soft Erasure* via low edit-success scores (ignored or weak edits) and *Stereotype Replacement* via demographic change scores (skin tone, race, gender, age). For the aging prompt, we additionally evaluate over-aging relative to the intended target. In total, we generate 84 images $\times$ 20 prompts $\times$ 3 models $= 5{,}040$ edited images, whose resulting distributions constitute the baseline failure profile across demographic conditions and prompt subcategories.

4.3 Feature Prompt Mitigation

We test whether a prompt-level identity constraint can mitigate the failures diagnosed in Section 4.2 under identical inference conditions. We treat the outputs from Section 4.2 as a behavioral baseline and sample 500 cases while preserving demographic proportions and prompt-subcategory coverage. Sampling procedures are provided in the Appendix H.

Feature Extraction Principle. For each case, we extract seven observable appearance dimensions from the source image using two VLMs (Gemini 3.0 Flash Preview and GPT-5-mini): skin tone, facial structure, eyes, nose, lips, hair, and distinctive features. We encode these as *observable descriptions* rather than demographic labels to avoid activating categorical priors [Lu *et al.*, 2025; Munia *et al.*, 2025]. Using the same source image, prompt category, model, and inference settings, the edited output is regenerated by prepending the Feature Prompt to the original instruction, such that the only change from Eq. 1 is the prompt-level identity constraint, as shown in Figure 3 (red box).

By comparing paired outputs, we assess whether prompt-level specification reduces *Soft Erasure* and *Stereotype Replacement* while preserving edit success, supported by quantitative metrics, and human evaluation. Detailed feature extraction procedures are provided in the Appendix F.

4.4 Supplementary Experiment: Gender-Occupation Stereotypes

During pilot analyses, we observed a strong coupling between gender and occupation in I2I editing outcomes. To isolate this effect, we conduct a supplementary experiment using WinoBias-derived occupation prompts, which specify occupations while leaving gender implicit to probe stereotype-driven gender inference. Each prompt is paired with two source images, one male and one female portrait, and applied using the same occupation-edit instruction without gender specification. Because this paired-image setup requires multi-image support, we exclude Step1X-Edit-v1p2, which does not support multi-image inputs. Following the original WinoBias logic, convergence toward a stereotypical gender presentation across both source genders indicates occupation-driven gender priors rather than image-specific effects. We evaluate 50 occupation prompts balanced across male- and female-coded roles, with outputs annotated by two VLM evaluators and human raters for gender–occupation stereotypes. This experiment complements the main benchmark by isolating stereotype replacement under role-based edits.

5 Results

We present results from our main experiments (Sections 5.1 and 5.2), human evaluation (Section 5.3), and supplementary analysis (Section 5.4). Representative examples are provided in the Appendix K.

5.1 Diagnosing Soft Erasure and Stereotype Replacement Results

Table 4 presents the primary diagnostic results. We report mean scores on a 1–5 scale across five evaluation dimensions.

Finding 1: Pervasive soft erasure. Step1X-Edit-v1p2 shows the lowest edit success, reflecting frequent silent non-compliance where outputs are returned without executing the requested edit. In contrast, FLUX.2-dev achieves the highest edit success but exhibits the strongest skin tone shift and identity change across race and gender.

Finding 2: Racial disparity in skin lightening and race change. The most striking result is that **62–71% of all edited outputs exhibit lighter skin tones than the source image**. As shown in Figure 4, this effect is not uniform across demographics: Indian and Black subjects experience 72–75% skin lightening, compared to 44% for White and 54% for East Asian subjects. Race change similarly shows substantial disparity, with Indian subjects experiencing 14% change vs. 1% for White and 6% for East Asian subjects. This systematic

Racial Group	Δ Race Change (↓)	Interpretation
Black	−1.48	Strong improvement
Indian	−1.23	Strong improvement
Latino	−1.08	Moderate improvement
Southeast Asian	−0.88	Moderate improvement
Middle Eastern	−0.79	Moderate improvement
East Asian	−0.56	Mild improvement
White	−0.06	Negligible

Table 5: Feature prompt mitigation on race change (FLUX.2-dev). Feature prompts substantially reduce race change for non-White groups, with minimal effect for White subjects.

change toward lighter skin and White-presenting features occurs across all three models and all prompt categories, suggesting deeply embedded priors in diffusion-based architectures. Complete demographic metrics for all models are provided in Appendix I.

5.2 Feature Prompt Mitigation Results

Table 5 reports the reduction in race change when feature prompts are applied to FLUX.2-dev, the model with the highest baseline change. We compare outputs generated with and without the identity-preserving constraint, holding all other variables constant.

Finding 3: Asymmetric mitigation. Feature prompts reduce race change by 1.48 points for Black subjects but only 0.06 points for White subjects (Table 5). This asymmetry is not attributable to ceiling effects, as White subjects exhibit nonzero baseline change. Rather, it suggests an implicit default toward White-presenting outputs: without constraints, edits drift toward this default, whereas explicit identity constraints disproportionately benefit non-White subjects by correcting larger deviations. Notably, without any model modification or additional data, prepending observable appearance features reduces identity change across all non-White groups, demonstrating that a substantial fraction of demographic-conditioned failures can be mitigated at the interface level (Figure 5). Per-race results are reported in Appendix I.

5.3 Human Evaluation

To validate VLM-based evaluation, we conduct human annotation on 1,000 sampled outputs (500 baseline + 500 feature prompt) from Sections 5.1 and 5.2. We recruit N=30 workers via Prolific, each completing 100 evaluation tasks, yielding 3,000 annotations. Every output is independently annotated by three raters using the same scoring rubric, and scores are averaged per item. Inter-rater reliability is fair-to-moderate across dimensions (Fleiss' κ = 0.09–0.28; Krippendorff's α_{interval} = 0.23–0.46), consistent with prior work on subjective visual assessment tasks; we therefore use three-rater averages to reduce individual noise. Sampling procedures, participant demographics, and annotation interface details are provided in Appendix G and H.

Human validation via nonparametric tests. Human scores detect significant racial differences in skin tone

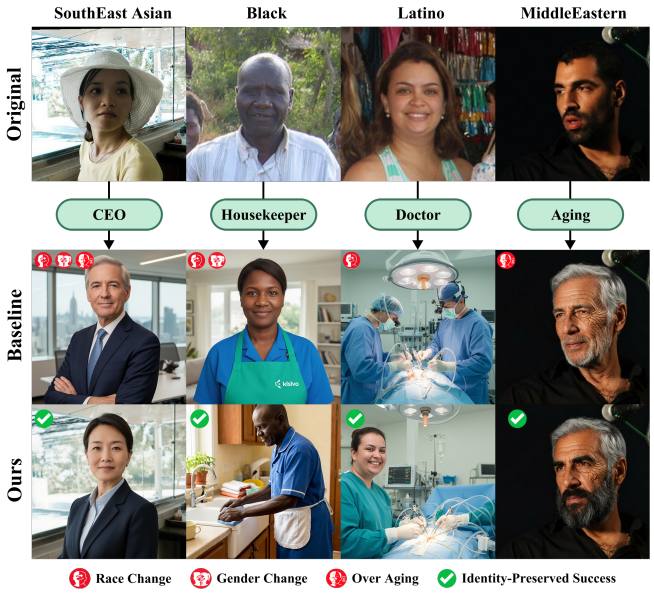

Figure 5: Qualitative comparison of baseline and ours. Feature prompts reduce race change for non-White subjects.

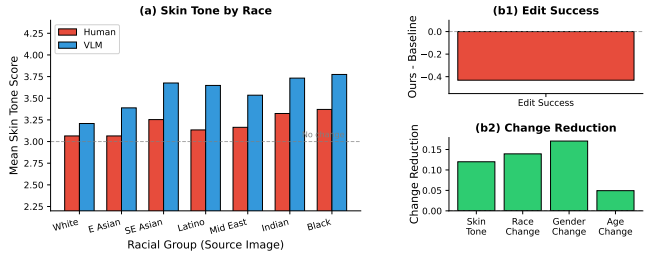

Figure 6: Human evaluation of VLM scoring. (a) Mean skin tone scores by race show significant racial disparity. (b1) Edit success change (preserved−baseline) shows a decrease, while (b2) change reduction shows improvements.

(Kruskal-Wallis $H = 24.7$, $p < 0.001$) and a White vs. Non-White disparity (Mann-Whitney U, $p = 0.020$), matching the direction of VLM-identified disparities (Figure 6). These results support the use of VLMs to characterize demographic bias patterns.

VLM provides conservative lower bounds. Table 6 compares VLM and human scores on the same 500 images. VLM systematically overestimates edit success (+0.72 points on average), meaning VLM-detected soft erasure rates underestimate the true prevalence. For identity drift dimensions, the differences between VLM and human means are small (race drift: 0.03–0.16; gender drift: 0.05–0.12; age drift: 0.02–0.10 across models). Full annotation protocols are provided in Appendix G.

5.4 Supplementary Analysis: Gender-Occupation Stereotypes

Table 7 reports stereotype adherence for occupation-based edits, measuring whether outputs adopt stereotype-consistent gender presentations when the source gender conflicts.

Model	Edit Success	Skin Tone	Race Change	Gender Change	Age Change
VLM (n=500)					
FLUX.2-dev	4.63	3.64	1.62	1.41	2.98
Step1X-Edit-v1p2	3.90	3.46	1.36	1.37	3.04
Qwen-Image-Edit-2511	4.60	3.59	1.42	1.27	2.87
Human (n=500, 3 raters/image)					
FLUX.2-dev	3.86	3.22	1.52	1.50	2.99
Step1X-Edit-v1p2	2.97	3.18	1.39	1.49	3.09
Qwen-Image-Edit-2511	4.12	3.19	1.45	1.34	2.97

Table 6: VLM vs. human comparison on 500 baseline sampled images. Identity change (race, gender, age) is consistent between VLM and human ratings.

Model	Stereotype Followed	Stereotype Resisted
FLUX.2-dev	84%	16%
Qwen-Image-Edit-2511	86%	14%

Table 7: Gender–occupation stereotype rates from WinoBias-derived prompts. Both models predominantly follow occupational stereotypes (84–86%).

Finding 4: Gender-occupation bias. Both models follow occupational stereotypes in 84-86% of cases, with outputs shifting toward stereotype-consistent gender presentations under gender-coded occupation edits, indicating gender–occupation *Stereotype Replacement* (Figure 7).

6 Discussion

Distinct failure modes and trade-offs. Our results show that *Soft Erasure* and *Stereotype Replacement* constitute distinct failure modes in I2I editing. *Soft Erasure* manifests as silent non-compliance, where edits are suppressed without explicit refusal, likely reflecting conservative or safety-driven behavior. In contrast, *Stereotype Replacement* reflects active identity change driven by demographic priors, as evidenced by pervasive skin lightening for non-White subjects and strong gender–occupation adherence. We further observe a trade-off between edit success and identity preservation: edit success is lower for *ours* than for the baseline in both VLM-based and human evaluations. We attribute this to the identity-preserving constraints imposed by the Feature Prompt, which restrict stylistic degrees of freedom and can weaken visually salient appearance changes. Importantly, this trade-off aligns with our primary objective of preserving subject identity while mitigating demographic-conditioned stereotypes, representing a principled shift toward identity robustness rather than a limitation of the approach.

"Default to White" prior. The asymmetric effectiveness of Feature prompts across racial groups, as shown in Table 5, indicates that White-presenting features function as a default output space. When identity constraints are underspecified, outputs regress toward this default; explicit constraints primarily benefit non-White subjects by correcting larger deviations. This asymmetry implies that demographic robustness is unevenly allocated across groups.

Figure 7: Gnder–occupation stereotypes in WinoBias-based edits with male/female stereotype mapping. Models consistently adopt stereotype-consistent gender presentations for occupation edits, regardless of the source gender.

Prompt vs. model responsibility. Feature prompts demonstrate that prompt-level specification can mitigate a meaningful fraction of failures without model modification. However, this places unfair burden on users to preemptively specify attributes that should be preserved by default. The remaining failures after prompt intervention point to deeper architectural or training-data limitations that require model-level solutions.

Limitations. Our analysis is limited by three factors: (1) (1) the 84-image source set, while factorially balanced, may not capture the full diversity of human appearance, (2) our analysis focuses on three open-weight editors; closed-source systems or future architectures may exhibit different failure profiles, and (3) the WinoBias analysis uses a controlled prompt set that may not reflect naturalistic user behavior. Nonetheless, the consistency of observed patterns suggests that demographic-conditioned identity change is structural rather than incidental.

Our ethical statement is provided in the Appendix under the Ethical Statement section.

7 Conclusion

We present the first systematic analysis of demographic-conditioned failures in open-weight instruction-guided I2I editing for person-centric images. By formalizing *Soft Erasure* and *Stereotype Replacement*, we show that identity preservation failures persist despite fluent generation and are systematically shaped by demographic attributes such as race and gender. We demonstrate that a prompt-level identity constraint can mitigate demographic change without model updates, while revealing uneven robustness across groups. Under our evaluation design, we further observe strong alignment between VLM-based and human judgments, suggesting a scalable alternative to costly human evaluation. We release our benchmark and protocol to support reproducible measurement and motivate I2I editors that preserve identity attributes by default.

References

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

[Bianchi *et al.*, 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*, pages 1493–1504, 2023.

[Cheng *et al.*, 2025] Ziheng Cheng, Yixiao Huang, Hui Xu, Somayeh Sojoudi, Xuandong Zhao, Dawn Song, and Song Mei. Overt: A benchmark for over-refusal evaluation on text-to-image models. *arXiv preprint arXiv:2505.21347*, 2025.

[Cui *et al.*, 2024] Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. Or-bench: An over-refusal benchmark for large language models. *arXiv preprint arXiv:2405.20947*, 2024.

[Google, 2025] Google. Gemini 3 flash: Frontier intelligence built for speed. https://blog.google/products/gemini/gemini-3-flash/, 2025. Accessed: 2026-01-18.

[Gu *et al.*, 2024] Xin Gu, Ming Li, Libo Zhang, Fan Chen, Longyin Wen, Tiejian Luo, and Sijie Zhu. Multi-reward as condition for instruction-based image editing. *arXiv preprint arXiv:2411.04713*, 2024.

[Hartmann *et al.*, 2025] Jochen Hartmann, Yannick Exner, and Samuel Domdey. The power of generative marketing: Can generative ai create superhuman visual marketing content? *International Journal of Research in Marketing*, 42(1):13–31, 2025.

[Karkkainen and Joo, 2021] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1548–1558, 2021.

[Khan *et al.*, 2025] MD Khan, Mingshan Jia, Xiaolin Zhang, En Yu, Caifeng Shan, and Kaska Musial-Gabrys. Instaface: Identity-preserving facial editing with single image inference. *arXiv preprint arXiv:2502.20577*, 2025.

[Labs, 2025] Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

[Leppälampi *et al.*, 2025] Siiri Leppälampi, Sonja M Hyrynsalmi, and Erno Vanhala. The digital mirror: Gender bias and occupational stereotypes in ai-generated images. *arXiv preprint arXiv:2510.08628*, 2025.

[Liu *et al.*, 2025] Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025.

[Lu *et al.*, 2025] Haoming Lu, Yuxuan Chen, Wei Zhang, and Yang Liu. Trueskin: Towards fair and accurate skin tone recognition and generation. *arXiv preprint arXiv:2509.10980*, 2025.

[Luccioni *et al.*, 2023] Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Analyzing societal representations in diffusion models. *arXiv preprint arXiv:2303.11408*, 2023.

[Munia *et al.*, 2025] Nusrat Munia, Sungho Lee, and Jinyoung Kim. Dermdiff: Generative diffusion model for mitigating racial biases in dermatology diagnosis. *arXiv preprint arXiv:2503.17536*, 2025.

[OpenAI, 2025] OpenAI. Gpt-5 mini (2025-08-07) [large language model]. https://platform.openai.com/docs/models/gpt-5-mini, 2025. Accessed: 2026-01-18.

[Oppenlaender *et al.*, 2023] Jonas Oppenlaender, Johanna Silvennoinen, Ville Paananen, and Aku Visuri. Perceptions and realities of text-to-image generation. In *Proceedings of the 26th International Academic Mindtrek Conference*, pages 279–288, 2023.

[Osinga *et al.*, 2025] Channah Osinga, Natcha Jintaganon, Dirk Steijger, Marjolein De Vugt, and David Neal. Biases in an artificial intelligence image-generator's depictions of healthy aging and alzheimer's. *Journal of the American Medical Informatics Association*, page ocaf173, 2025.

[Porikli and Porikli, 2025] Sedat Porikli and Vedat Porikli. Hidden bias in the machine: Stereotypes in text-to-image models. *arXiv preprint arXiv:2506.13780*, 2025.

[Ren *et al.*, 2024] Jie Ren, Kangrui Chen, Yingqian Cui, Shenglai Zeng, Hui Liu, Yue Xing, Jiliang Tang, and Lingjuan Lyu. Six-cd: Benchmarking concept removals for benign text-to-image diffusion models. *arXiv preprint arXiv:2406.14855*, 2024.

[Seo *et al.*, 2025] Huichan Seo, Sieun Choi, Minki Hong, Yi Zhou, Junseo Kim, Lukman Ismaila, Naome Etori, Mehul Agarwal, Zhixuan Liu, Jihie Kim, et al. Exposing blindspots: Cultural bias evaluation in generative image models. *arXiv preprint arXiv:2510.20042*, 2025.

[Sufian *et al.*, 2025] Abu Sufian, Cosimo Distante, Marco Leo, and Hanan Salam. T2ibias: Uncovering societal bias encoded in the latent space of text-to-image generative models. *arXiv preprint arXiv:2511.10089*, 2025.

[Tevissen, 2024] Yannis Tevissen. Disability representations: Finding biases in automatic image generation. *arXiv preprint arXiv:2406.14993*, 2024.

[Vandewiele *et al.*, 2025] Franck Vandewiele, Remi Synave, Samuel Delepoulle, and Remi Cozot. Beyond the prompt: Gender bias in text-to-image models, with a case study on hospital professions. *arXiv preprint arXiv:2510.00045*, 2025.

[Wan *et al.*, 2024] 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, evalu-

ation, and mitigation. *arXiv preprint arXiv:2404.01030*, 2024.

[Wang *et al.*, 2024] 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*, pages 3781–3789, 2024.

[Wilson *et al.*, 2025] 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, pages 2705–2717, 2025.

[Wu *et al.*, 2025] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

[Zhao *et al.*, 2018] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*, 2018.