A Counterfactual-style Diagnostic Framework for Spurious Correlations in Text-to-Image Models

Anonymous authors
Paper under double-blind review

Abstract

Text-to-image diffusion models often encode correlations between demographic prompts and non-demographic attributes, some of which may be expected (e.g., gray hair with older age) while others may raise fairness concerns (e.g., cultural markers appearing only for certain ethnicities). Existing analyses of such correlations have been largely qualitative. In this work, we present a counterfactual-style diagnostic framework for stress-testing diffusion models. Inspired by stress-testing approaches (e.g., Veitch et al.), our method uses image-conditioned generation to approximately preserve facial features while systematically varying demographic variables in prompts (gender, ethnicity, age). This setup enables controlled observation of how non-demographic attributes (e.g., facial hair, accessories, hairstyles) shift under demographic changes. We introduce Counterfactual-style Invariance (CIV), along with positive and negative variance metrics (PCV, NCV), to quantify attribute stability and directional changes. Applying this framework across multiple text-to-image models reveals pervasive, prompt-dependent entanglements—for example, bushy eyebrows co-occur in 62.5% of generations with "Middle Eastern" prompts, and Black hair is amplified in 64.8% of "East Asian" generations. These findings show that generative models can amplify or introduce associations between the demographic variables and observed attributes. This highlights the need for systematic diagnostic evaluations to better understand and mitigate fairness risks in text-to-image generation.

1 Introduction

Text-to-image diffusion models such as Stable Diffusion have revolutionized image generation, producing diverse and high-fidelity visuals from natural language prompts. However, alongside their impressive capabilities, these models often encode correlations between demographic variables in prompts (e.g., gender, ethnicity, age) and other attributes, both demographic and non-demographic, such as hair texture, facial hair, and accessories. Some of these associations may be expected (e.g., gray hair with older age), while others raise fairness concerns (e.g., cultural markers appearing only for certain ethnicities). Such patterns (or *spurious correlations*) can inadvertently reinforce stereotypes, emphasizing the need for systematic evaluation of how these models handle demographic information.

A growing body of work has begun to audit and quantify bias in text-to-image generation. Early efforts often relied on visual inspection, while more recent approaches have incorporated automated classifiers to measure demographic representation or compare generated distributions with real-world data (Naik & Nushi, 2023; Luccioni et al., 2023). However, most existing evaluations remain limited to measuring the demographic composition of generated subjects (e.g., proportions of genders or ethnicities), without systematically analyzing the additional attributes that appear alongside those subjects, such as hairstyles, accessories, or facial features, and whether these co-occurrences reflect expected patterns (e.g., gray hair with older age) or potentially stereotypical associations (e.g., headscarves appearing only for certain ethnicities).

We address this gap by analyzing how demographic variables in text prompts influence the presence of both demographic and non-demographic attributes in generated images. Drawing inspiration from counterfactual fairness analysis (Kusner et al., 2017; Veitch et al., 2021), we introduce a *counterfactual-style diagnostic*

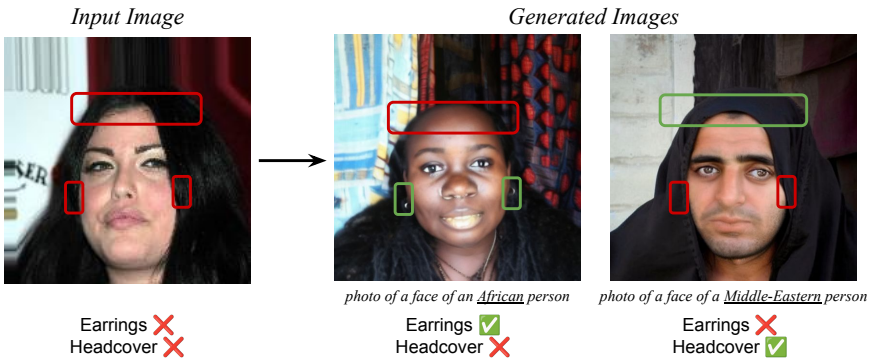

Figure 1: An illustration of spurious correlations in text-to-image generation. Images generated for certain ethnicities systematically exhibit *earrings* and *headcovers* attributes, showcasing unintended associations that may reinforce stereotypical representations.

framework which uses *image-conditioned generation* to approximately preserve a person's facial features while systematically varying demographic terms in prompts. This controlled setup isolates the impact of demographic descriptors, enabling us to observe how other attributes shift when only the demographic context changes.

To quantify these effects, we propose three complementary metrics. *Counterfactual-style Invariance (CIV)* measures the extent to which attributes remain unchanged when a demographic descriptor is altered. *Positive Counterfactual-style Variance (PCV)* captures the degree to which attributes are added or amplified following a demographic change, while *Negative Counterfactual-style Variance (NCV)* captures the extent to which they are removed or diminished. Together, these metrics move beyond aggregate demographic counts to provide a structured, quantitative assessment of prompt-induced changes, enabling us to distinguish between attribute invariance, amplification, and suppression.

We apply our framework to facial image generation, enabling fine-grained analysis of both demographic and non-demographic attributes. Using over 120,000 generations across multiple prompts and state-of-the-art diffusion models, we uncover persistent and prompt-dependent attribute entanglements. For example, *bushy eyebrows* appear in over 62.5% of images generated from "Middle Eastern" prompts, while *Black hair* is amplified in over 64.8% of "East Asian" generations. These patterns suggest that generative models can amplify or introduce correlations between demographic variables and other attributes, some of which may reflect stereotypes. Not all correlations revealed by our framework are inherently problematic. Certain demographic attribute links are expected and contextually appropriate, for example, elderly prompts producing wrinkles or gray hair. However, others raise fairness concerns. A consistent tendency to associate particular ethnicities with specific hairstyles or accessories, or to depict women with added makeup and jewelry by default, risks reinforcing limiting and stereotypical portrayals (Refer Figure 1). While determining whether a correlation is contextually appropriate or stereotype-reinforcing may involve normative judgment, systematically identifying and quantifying such patterns is a necessary first step toward informed discussions and mitigation strategies in generative AI.

Our contributions are threefold: (1) we present a counterfactual-style diagnostic framework for systematically analyzing how text-to-image models handle demographic prompts, (2) we introduce quantitative metrics that disentangle attribute invariance from prompt-induced additions and removals, allowing nuanced fairness evaluations, and, (3) we conduct a qualitative study showing how these metrics reveal both expected and spurious correlations, highlighting risks of stereotype reinforcement in state-of-the-art diffusion models. By moving beyond demographic counts to attribute-level diagnostics, our framework offers a new lens for fairness analysis in generative AI. It enables more targeted bias detection, supports principled discussions on mitigation, and complements existing evaluation methods by addressing a previously underexplored dimension: the systematic co-occurrence of demographic and non-demographic attributes in generated content.

2 Related Work

Spurious Correlations and Counterfactual-Style Analysis: The literature on spurious correlations has predominantly focused on predictive frameworks. Veitch et al. (2021) established a formal definition of counterfactual invariance in predictive settings, demonstrating how strategically designed stress tests that "poke" the system can effectively uncover spurious correlations. Subsequent research has extended these methodologies across various predictive domains (Feder et al., 2022). Within the generative paradigm, Fang et al. (2024) examined how personal attributes influence biography generation by analyzing pairs of identical biographies that differ solely in specific personal attributes of interest. While spurious correlations have been discussed in various vision-language contexts (Liusie et al., 2022; Kim et al., 2023), they remain, to the best of our knowledge, unexplored within the specific domain of text-to-image generation.

Fairness in Text-to-Image Generation: Recent investigations have systematically evaluated bias in text-to-image systems. Luccioni et al. (2023) conducted a comprehensive analysis of how gender and ethnicity markers manifest in images generated from profession-related prompts, identifying systematic representational disparities. Fraser et al. (2023) revealed racial implications in images generated from socially charged prompts such as "lawyer" and "felon," highlighting the propagation of harmful stereotypes. Bianchi et al. (2023) provided qualitative evidence showing how prompts invoking socioeconomic status, social structures, and occupations become entangled with sexist, racist, and heteronormative signals in the outputs. Naik & Nushi (2023) further documented the predominance of young adults (aged 18–40) in generated images, underscoring age-related representational imbalances. Collectively, these studies illustrate the multifaceted nature of bias in text-to-image systems and emphasize the need for more rigorous analytical frameworks.

Facial Attributes in Text-to-Image Generation: While studies have examined bias in text-to-image models in relation to specific professions or cultural concepts, few have concentrated exclusively on facial image generation, a critical domain given its sensitivity to demographic representation. Rosenberg et al. (2023) conducted an interesting investigation into face generation within text-to-image models, employing SEGA (Yang et al., 2022) to produce demography-specific outputs and systematically analyze the inherent limitations of contemporary text-to-image systems.

3 Counterfactual-Style Diagnostics for Spurious Correlations

We use the term *spurious* to refer to unintended or non-essential associations between demographic variables and other attributes, associations that may reinforce undesirable stereotypes or raise fairness concerns (e.g., earrings or head coverings predominantly appearing for specific ethnicities). This does not imply that such correlations are false; rather, our diagnostic framework aims to surface and quantify them for further analysis.

Counterfactual-Style Observational Framework: Unlike predictive tasks, where spurious correlations have been extensively studied under causal frameworks (Veitch et al., 2021), the generative setting lacks tools for systematic analysis. While causal inference frameworks (Pearl, 2009; Pearl et al., 2016) provide valuable conceptual grounding, we emphasize that our approach does not assume a structural causal model (SCM) for text-to-image diffusion models. Instead, we adopt an empirical stress-testing approach inspired by Veitch et al. (2021), where inputs are systematically varied to probe the stability of model behavior.

Mathematically, we formalize our setting as follows. Let x denote a demographic variable in the prompt (e.g., gender, ethnicity, age). Let U denote the set of input images, and let $u \in U$denote an individual input image which serves as the conditioning variable for the generative model. The generated image is denoted by $y = Y_x(u)$ where $Y_x(u)$ denotes the model's output under prompt x and input u. Each generated image y is associated with an attribute vector $A(y)$ that characterizes its salient features. For each image, we annotate n attributes such that $A(y) = \{a_1^y, a_2^y, \ldots, a_n^y\}$ with $a_i^y \in \{0, 1\}$ for $i = 1, \ldots, n$. These attributes include both demographic and non-demographic attributes (e.g., facial hair, hairstyles, accessories, gender, age).

Analyzing relationships solely between the attributes in the generated image $A(y)$ does not provide a controlled framework for studying the impact of prompt variations. To systematically influence input components, we replace the random noise typically used in text-to-image pipelines with a specific input image u, which serves as the conditioning variable. This input image has an associated attribute vector $A(u)$.

Analogous to $A(y)$, $A(u) = \{a_1^u, a_2^u, \ldots, a_n^u\}$ with $a_i^u \in \{0, 1\}$ for $i = 1, \ldots, n$. This conditioning ensures that high-level facial features remain approximately fixed, enabling a controlled analysis of how demographic prompt variations influence attribute outcomes.

To summarize, in our setup, the input image u is fixed to preserve facial features, while the demographic variable x in the prompt is systematically varied. This allows us to observe changes in attributes $A(y)$ with control over existing attribute distribution $A(u)$. We note that this process does not estimate causal effects but rather diagnoses representational stability and attribute entanglement.

Measuring Counterfactual-Style Invariance (CIV): Based on this framework, we formulate a measure of counterfactual invariance. The objective of counterfactual invariance is to capture the impact of changes in the input u for changes in $Y_x(u)$. Specifically, while holding the prompt x fixed, we modify the input image u, thereby altering its corresponding attributes $A(u)$. Subsequently, analyzing $A(y)$ relative to $A(u)$ yields insights into attribute associations and correlations. The counterfactual invariance for a given attribute i is computed as:

$$CIV_x(i) = \frac{|\{u \in U : a_i^u = 0 \wedge a_i^y = 0\}| + |\{u \in U : a_i^u = 1 \wedge a_i^y = 1\}|}{|\{u \in U : a_i^u = 0\}| + |\{u \in U : a_i^u = 1\}|} \tag{1}$$

where $|\cdot|$ denotes the cardinality (number of samples) of the set. Intuitively, $CIV_x(i)$ quantifies the degree to which the ith attribute is preserved through the generation process given x.

To further enhance the evaluation and understand the direction of the attribute shifts, we define Positive Counterfactual-style Variance (PCV) and Negative Counterfactual-style Variance (NCV). This decomposition elucidates whether an attribute is more likely to be *enhanced* or *diminished*. Formally, these quantities are defined as:

$$PCV_x(i) = \frac{|\{u \in U : a_i^u = 0 \wedge a_i^y = 1\}|}{|\{u \in U : a_i^u = 0\}|} \tag{2}$$

$$NCV_x(i) = \frac{|\{u \in U : a_i^u = 1 \wedge a_i^y = 0\}|}{|\{u \in U : a_i^u = 1\}|} \tag{3}$$

For instance, if an input image u depicts a male face without facial hair, and under the prompt 'photo of a Middle Eastern person', the generated output y adds sideburns, this is captured as a positive shift in PCV for the 'sideburns' attribute. Together, these metrics provide a structured, interpretable view of how attributes respond to demographic variations in prompts.

While PCV and NCV separately measure the proportion of attribute additions and suppressions relative to their respective source groups, their values are not directly comparable due to different denominators. To address this, we also define normalized variants (PCV') and (NCV'), where both are computed relative to the total number of samples as,

$$PCV_x(i)' = \frac{|\{u \in U : a_i^u = 0 \wedge a_i^y = 1\}|}{|\{u \in U : a_i^u = 0\}| + |\{u \in U : a_i^u = 1\}|} \tag{4}$$

$$NCV_x(i)' = \frac{|\{u \in U : a_i^u = 1 \wedge a_i^y = 0\}|}{|\{u \in U : a_i^u = 0\}| + |\{u \in U : a_i^u = 1\}|} \tag{5}$$

These normalized metrics provide a balanced perspective on how frequently attributes are added or suppressed across the entire dataset, complementing the original definitions of PCV and NCV. To further analyze attribute dependencies and control for confounding variables, we also compute conditional co-occurrence matrices on both source and generated images, which complement CIV/PCV/NCV by revealing multi-attribute entanglements.

Relationship with Fairness: Fairness considerations arise when certain attributes consistently shift or entangle with demographic variables in ways that may reinforce stereotypes. Our metrics serve as diagnostic tools to identify fairness-relevant associations. We treat these evaluations as stress tests, revealing where model outputs may reflect biased or undesirable correlations, thus motivating further mitigation strategies.

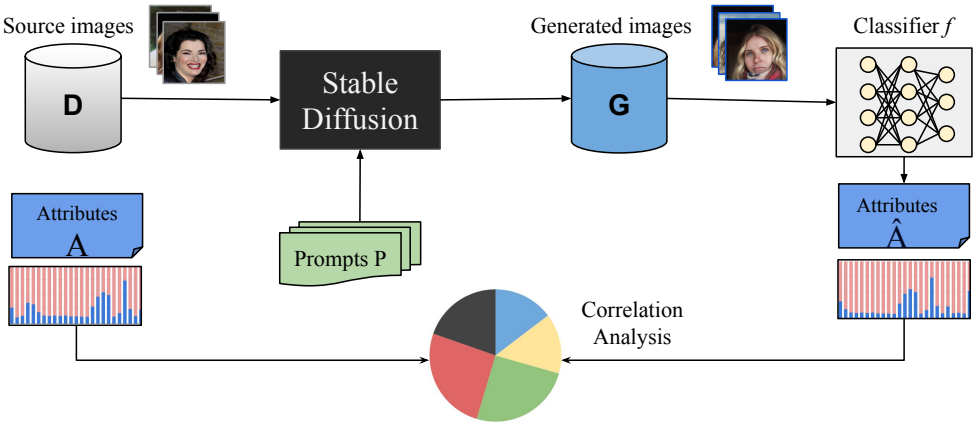

Figure 2: The proposed pipeline utilizes the image-to-image capabilities of the Stable Diffusion model to analyze the images generated with respect to different demographic and non-demographic attributes.

4 Experiment Design: Face Generation

The study of spurious correlations in text-to-image generation requires a domain with rich and varied features. Facial imagery serves as an ideal testbed where datasets exist that annotate up to 40 distinct facial characteristics (Liu et al., 2018). This multidimensional attribute space provides an excellent framework for probing correlational patterns and assessing their implications, especially with regard to fairness concerns identified in prior work. In this section, we outline the key components of our experimental setup, including model selection, prompt engineering, dataset curation, and attribute classification methods. Figure 2 shows an overview of the complete experimental pipeline.

4.1 Model and Prompt Selection

Our analysis employs three state-of-the-art diffusion-based text-to-image generative models—Stable Diffusion v1.5 (SDv1.5), Stable Diffusion v2 (SDv2), and Stable Diffusion XL (SDXL)—for both qualitative and quantitative analyses. Due to computational constraints, additional models including Midjourney, DALL-E 3, LLAMA, and Stable Diffusion 3 are utilized exclusively for qualitative assessment.

For quantitative experimentation, we meticulously designed 24 prompts spanning three fundamental demographic dimensions: age, gender, and ethnicity. Each prompt adheres to a standardized template: “The photo of a face of a $\langle d \rangle$ person" or “The photo of a face of a $\langle d \rangle$," where the descriptor $\langle d \rangle$ is selected from demographic-specific lexical sets. For age, we employ {young, middle-aged, old, infant, child, young adult, adult, elder}; for gender, {male, female}; and for ethnicity, {Indian, White, Caucasian, Latino, Hispanic, African, African American, Asian, Black, East Asian, Middle Eastern, Southeast Asian}, with an additional prompt for mixed ethnicity. We further include a control prompt without any demographic descriptor ($\langle d \rangle$) to establish a baseline for generation in the absence of explicit demographic signaling. The deliberate inclusion of overlapping descriptors across demographic categories enables cross-categorical pattern analysis, enhancing the robustness of our findings.

4.2 Dataset and Attributes Selection

To comprehensively examine spurious correlations in facial attribute generation, we focus on two attribute categories, informed by their prevalence in existing literature (Liu et al., 2018; Kärkkäinen & Joo, 2019).

Non-Demographic Attributes encompass appearance-related characteristics, with particular emphasis on hair features and accessories. For this analytical dimension, we utilize the CelebA dataset (Liu et al., 2018) as our primary corpus. This dataset comprises facial images annotated with 40 binary attributes. Given the inherently subjective nature of certain attributes (e.g., “attractive,” “chubby”), and the potential subjectivity

Table 1: The non-demographic (top) and demographic (bottom) attributes considered for the experiments.

Category	Attributes
Facial Hair	Bushy Eyebrows, Goatee, Mustache, No Beard, Sideburns
Hair Color	Black Hair, Blond Hair, Gray Hair
Hair Style	Bald, Bangs, Straight Hair
Accessories	Eyeglasses, Wearing Earrings, Wearing Hat
Gender	Male, Female
Ethnicity	Indian, Latino Hispanic, White, Black, East Asian, Southeast Asian, Middle Eastern
Age	0-19 (young), 20-29 (young adult), 30-39 (adult), 40-49 (adult/middle-aged), 50-59 (middle-aged), 60+ (old/elder)

Figure 3: Some samples from the CelebA dataset used as conditioning input for generation, along with the attributes present in them before generation.

in others (e.g., "big nose," "oval face," "receding hairline"), we restrict our analysis to attributes characterized by high objectivity and low annotator variability, as enumerated in Table 1.

From the CelebA dataset, we extract a carefully balanced subset of 2,000 images from the test partition (examples shown in Figure 3), ensuring a minimum representation of 200 positive samples for each attribute under investigation. To enhance analytical robustness, we supplement this with an additional CelebA subset and 2,000 samples from the Labeled Faces in the Wild (LFWA) dataset (Huang et al., 2008) for comparative quantitative analysis.

Demographic Attributes comprise gender, ethnicity, and age designations. Since these dimensions are not annotated in the CelebA dataset, we utilize the well-established FairFace classifier (Kärkkäinen & Joo, 2019) for their prediction. We exclude skin color from our analysis due to the absence of skin-tone annotations in CelebA and the observed inaccuracies in automated skin-tone classification, frequently compromised by illumination variability. While facial images inherently exhibit multiple attributes simultaneously, resulting in attribute distribution imbalances, our experimental methodology focuses on *change* in attributes rather than absolute frequencies, thereby mitigating concerns regarding observational skew.

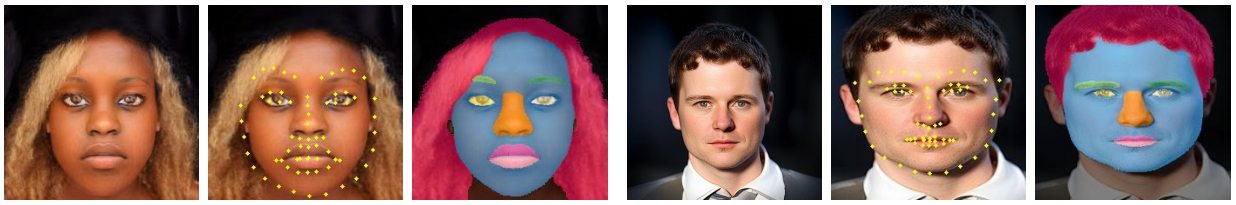

Figure 4: FaceXfomer detecting facial landmarks and face regions on images generated using SDv1.5 and SDXL models, highlighting its zero-shot capabilities for facial analysis.

4.3 Attribute Classification

Attribute classification for generated images requires specialized classifiers for both demographic and non-demographic characteristics. For demographic attributes, we utilize the pre-trained FairFace classifier (Kärkkäinen & Joo, 2019), widely acknowledged for its robustness in demographic classification tasks. For non-demographic attributes, we employ FaceXformer (Narayan et al., 2024), a unified transformer architecture designed for facial analysis. To validate this classifier's efficacy, we evaluated its performance against ground-truth labels from the CelebA data subset, achieving an accuracy of 93.89% and an F1-score of 49.03%. Comparative experimentation with the zero-shot CLIP model (Radford et al., 2021) for non-demographic attribute prediction yielded substantially inferior results (accuracy: 80.39%, F1-score: 49.03%), supporting our classifier selection. Given FaceXformer's multifunctional capabilities, including face parsing and facial landmark detection, we conducted manual validation of its performance on generated images using these auxiliary tasks, with representative examples illustrated in Figure 4.

User Study: To further corroborate the classifier's efficacy, we conducted a comprehensive user study involving 550 randomly sampled images generated using the SDv1.5 model across various prompts. Each user was asked three questions corresponding to each image. The first required the user to mark if a given image consisted of an attribute from the set of non-demographic attributes. Next, they were asked if the generated image corresponded to the prompt used to generate it. The users were provided with three options (a) Yes, (b) No, and (c) Not Sure. Further, they were asked to assess the quality of the image on a scale of 1 to 5, where the scale was defined as 1-not a face image or 2-face image, but most facial features are distorted. 3-face image, but some facial features are distorted. 4-face image with only the eye region distorted. 5- face image with clear facial features.

Based on the user responses, we observed an average agreement of 84.4% between the user ratings and the model predictions across all the attributes, demonstrating the reliability of the pre-trained classifier utilized for annotating attributes on the generated images. Using the user's response on whether a given image matched the specified demographic prompt, we observe that 80.34% of the image responses received a resounding yes, while in 11.38% of the cases, the users were not sure. In less than 10% of the cases, the models were found to be generating images that were not faithful to the provided prompts. The users also rated the average quality of the image as 3.88 out of 5, which is indicative of a reasonably high-quality face image. The average confidence of the 68 users in the study came out to be 3.85 out of 5, indicating high confidence in their responses.

Based on the user study, we can conclude that (a) the generated face images are of reasonable quality, (b) the images indeed reflect the demographic properties introduced through the prompt, and lastly, (c) the attribute classification may be safely scaled through the attribute classifier employed in the study.

Implementation Details

We use three stable diffusion models since they are open-source and free to use. We use the Stable Diffusion models, namely SDv1.5, SD2, and SDXL. All three models are popular and have high downloads on the Huggingface platform. All images were resized to 512x512 before being provided as input to the model. The *guidance scale* is set to 7.5, while the *strength* parameter is set to 0.75 for all generation processes. To ensure that no other factors are changed, the input images and model parameters are kept consistent for all the experiments, and we only change the descriptor (demographic) in our prompts. For qualitative analysis, we utilize Midjourney (mid, 2023), LLAMA3.1, and Dalle3 via ChatGPT-4o. All experiments are conducted on two 32GB GPUs on the NVIDIA DGX station.

5 Results and Analysis

This section presents both qualitative and quantitative results using the CelebA subset as described in Section 4.2. The results for non-demographic attributes are showcased in Tables 2 and 3, and the results for demographic attributes are shown in Tables 4 and 5.

Table 2: Counterfactual Invariance (CIV) of the SD1.5, SD2, and SDXL models on some of the non-demographic variables (%). Lower CIV indicates stronger susceptibility to prompt-induced attribute shifts. For example, bushy eyebrows under 'Middle Eastern' prompts show very low invariance, revealing high attribute entanglement.

Prompts	Attribute	SD1.5	SDv2	SDXL
A photo of a face of an African person.	Black_Hair	58.65	68.25	62.25
A photo of a face of an African American person.	Black_Hair	56.20	61.30	64.65
A photo of a face of an Asian person.	Black_Hair	62.00	34.30	69.95
A photo of a face of a Black person.	Black_Hair	58.05	56.60	59.05
A photo of a face of a East Asian person.	Black_Hair	63.30	35.20	70.25
A photo of a face of a Southeast Asian person.	Black_Hair	62.60	44.75	68.45
A photo of a face of a person with mixed ethnicity.	Black_Hair	68.40	41.20	66.55
A photo of a face of a Latino person.	Bushy_Eyebrows	71.70	24.60	40.75
A photo of a face of a Middle Eastern person.	Bushy_Eyebrows	37.45	19.15	22.95
A photo of a face of an old person.	Gray_Hair	44.35	42.70	70.85
A photo of a face of an elder.	Gray_Hair	48.15	49.10	73.45
A photo of a face of a Middle Eastern person.	Sideburns	68.85	51.80	85.55
A photo of a face of an Asian person.	Straight_Hair	49.15	47.00	49.65
A photo of a face of a East Asian person.	Straight_Hair	56.10	46.40	45.85
A photo of a face of a Southeast Asian person.	Straight_Hair	60.55	71.30	59.05
A photo of a face of an African American person.	Wearing_Earrings	73.75	85.55	82.65
A photo of a face of a Black person.	Wearing_Earrings	75.15	85.50	84.75
A photo of a face of a Middle Eastern person.	Wearing_Hat	63.40	65.30	87.80
A photo of a face of a Indian person.	Wearing_Hat	78.10	65.55	90.70

5.1 Non-Demographic Correlations

We computed the counterfactual invariance (CIV) for various non-demographic attributes, as detailed in Table 2. The PCV, NCV, PCV', NCV', as well as the specific counts for attributes being added and removed have been provided in Table 3. These are limited to the SDv1.5 model for readability, and the remaining results are provided as supplementary material online.

Hairstyles: For the attribute straight hair, SDv1.5 exhibits particularly low Counterfactual-style Invariance (CIV) for Asian-related prompts. CIV is only 49.15% for Asian and 56.1% for East Asian prompts, indicating frequent attribute changes when these demographic variables are varied. Southeast Asian prompts show slightly higher invariance (60.55%), but the attribute remains less stable than for many other attributes examined. Across models, this pattern of instability is consistent. SDv2 shows comparable or slightly lower CIV values (47.00% for Asian, 46.40% for East Asian), while SDXL also records low invariance (49.65% and 45.85%, respectively). This cross-model consistency suggests that straight hair is a persistently entangled attribute under Asian-related prompts. Directional metrics for SDv1.5 highlight this instability. For Asian prompts, PCV (50.76%) and NCV (51.90%) are nearly balanced, indicating frequent shifts in both directions. Yet when normalized across all samples, PCV(46.75%) exceeds NCV(4.10%), and raw counts (935 additions vs. 82 removals) reveal that in absolute terms, straight hair is more often added during generation. Similar patterns, such as low invariance, high instability, and an absolute tendency toward addition, are observed for East Asian and Southeast Asian prompts. Thus, straight hair under Asian-related prompts illustrates a case where low invariance coincides with prompt-dependent attribute amplification, providing clear evidence of representational entanglement in these models.

Hair Color: We observe strong prompt-dependent correlations between hair color and demographic variables. The attribute *black hair* consistently appears in generated faces for African-American and Asian prompts across all models, reflecting low CIV and a tendency toward demographic–attribute entanglement during generation. For *gray hair*, all models show a strong association with elderly prompts (Figure 5(h)). SDv1.5 and SDv2 record low CIV (44.35% and 42.70%, respectively), indicating that this attribute frequently shifts under age-related prompt variations. Directional metrics further reveal that these shifts predominantly involve attribute addition, with Positive Counterfactual Variance (PCV) reaching 56.85% for SDv1.5 and 59.08% for SDv2. This suggests that generative models amplify gray hair when age prompts are introduced, reinforcing age-associated patterns that, while expected, still warrant careful consideration in fairness analyses.

Facial Hair: Distinct patterns are observed in the generation of facial hair attributes across demographic variables. The attribute *bushy eyebrows* shows a strong association with Middle Eastern prompts, reflected in low CIV across SDv1.5, SDv2, and SDXL (37.45%, 19.15%, and 22.95%, respectively). Directional metrics indicate that this association is primarily driven by attribute addition, with Positive Counterfactual Variance (PCV) reaching 64.11% in SDv1.5 (refer Figure 5(d)) and similarly high values in SDv2 (84.32%)

Table 3: Columns 3-7 report metrics for the SDv1.5 models for non-demographic variables (%). Columns 8-9 report the absolute count of samples for which a particular attribute was added and removed, respectively.

Prompts	Attribute	CIV	PCV	NCV	PCV'	NCV'	0 → 1	1 → 0
A photo of a face of an African person.	Black_Hair	58.65	36.11	64.50	29.45	11.90	589	238
A photo of a face of an African American person.	Black_Hair	56.20	39.12	64.50	31.90	11.90	638	238
A photo of a face of an Asian person.	Black_Hair	62.00	30.72	70.19	25.05	12.95	501	259
A photo of a face of a Black person.	Black_Hair	58.05	36.11	67.75	29.45	12.50	589	250
A photo of a face of a East Asian person.	Black_Hair	63.30	28.14	74.53	22.95	13.75	459	275
A photo of a face of a Southeast Asian person.	Black_Hair	62.60	29.55	72.09	24.10	13.30	482	266
A photo of a face of a person with mixed ethnicity.	Black_Hair	68.40	20.29	81.57	16.55	15.05	331	301
A photo of a face of a Latino person.	Bushy_Eyebrows	71.70	25.84	75.00	24.55	3.75	491	75
A photo of a face of a Middle Eastern person.	Bushy_Eyebrows	37.45	64.11	33.00	60.90	1.65	1218	33
A photo of a face of an old person.	Gray_Hair	44.35	56.85	41.88	52.30	3.35	1046	67
A photo of a face of an elder.	Gray_Hair	48.15	52.99	38.75	48.75	3.10	975	62
A photo of a face of a Middle Eastern person.	Sideburns	68.85	25.10	74.09	22.00	9.15	440	183
A photo of a face of an Asian person.	Straight_Hair	49.15	50.76	51.90	46.75	4.10	935	82
A photo of a face of a East Asian person.	Straight_Hair	56.10	42.51	60.13	39.15	4.75	783	95
A photo of a face of a Southeast Asian person.	Straight_Hair	60.55	36.26	76.58	33.40	6.05	668	121
A photo of a face of an African American person.	Wearing_Earrings	73.75	16.42	85.87	14.10	12.15	282	243
A photo of a face of a Black person.	Wearing_Earrings	75.15	14.74	86.22	12.65	12.20	253	244
A photo of a face of a Middle Eastern person.	Wearing_Hat	63.40	32.81	67.74	29.25	7.35	585	147
A photo of a face of a Indian person.	Wearing_Hat	78.10	14.25	84.79	12.70	9.20	254	184

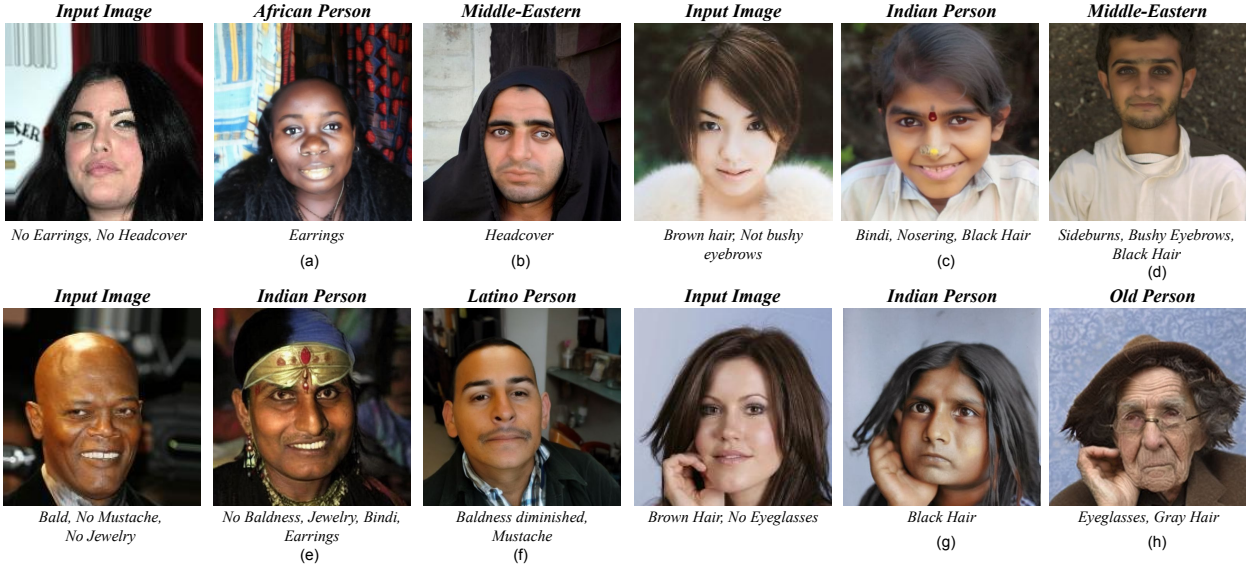

Figure 5: Demonstration of the various biases observed in the Stable Diffusion (SDv1.5) model observed quantitatively and qualitatively corresponding to different non-demographic attributes such as earrings, head cover, and jewelry.

and SDXL (81.00%), based on extended results. Indian, Hispanic, and Latino prompts also display an amplified presence of bushy eyebrows, whereas the attribute shows minimal correlation with age but remains strongly linked to male-coded images. For *sideburns*, SDv2 shows a notable association with Middle Eastern prompts, with a CIV of 51.80%—much lower than for other ethnicities such as Caucasian (80.55%) and Indian (79.75%), according to extended results. These observations highlight how specific facial hair attributes become disproportionately entangled with certain demographic prompts across models.

Accessories: We analyze three accessory attributes: *eyeglasses*, *hats*, and *earrings*. No strong demographic correlations were observed for eyeglasses. In contrast, hat-wearing exhibits clear demographic dependencies. For Middle Eastern prompts, SDv1.5 records a Counterfactual-style Invariance (CIV) of 63.40%, which is lower than the values for other ethnicities such as Indian (78.10%) and Caucasian (approximately 80% based on extended results). This lower invariance reflects a stronger entanglement between hats and Middle Eastern prompts (refer Figure 5(b)). In SDv2, the pattern persists with Middle Eastern prompts (CIV 65.30%) and extends to Indian prompts (CIV 65.55%), both showing reduced invariance compared to other groups. Generations conditioned on older age prompts also exhibit increased variation in hat presence, suggesting that the models associate hats more strongly with old age.

Table 4: Counterfactual Invariance (CIV) of the SD1.5, SD2, and SDXL models on some of the demographic variables (%).

Prompts	Attribute	SD1.5	SD2	SDXL
A photo of a face of a White person.	Age: 20-29	53.70	46.25	64.10
A photo of a face of a Caucasian person.	Age: 20-29	62.60	49.60	63.65
A photo of a face of an African American person.	Age: 20-29	55.30	49.50	59.25
A photo of a face of an Asian person.	Age: 20-29	56.60	52.45	60.25
A photo of a face of a Latino person.	Age: 20-29	61.35	51.75	57.70
A photo of a face of a Hispanic person.	Age: 50-59	72.90	85.50	85.95
A photo of a face of a Latino person.	Age: 50-59	77.10	87.05	87.50
A photo of a face of a Indian person.	Age: 40-49	78.95	64.40	82.45
A photo of a face of a female person.	Ethnicity: White	64.75	63.30	67.00
A photo of a face of a female person.	Ethnicity: East Asian	76.90	93.65	78.80
A photo of a face of a male person.	Ethnicity: White	69.35	54.40	71.35
A photo of a face of an infant.	Ethnicity: East Asian	84.15	84.95	84.40
A photo of a face of a young adult.	Ethnicity: East Asian	92.40	92.15	85.65
A photo of a face of an elder.	Ethnicity: Southeast Asian	84.00	54.25	25.30
A photo of a face of a young person.	Gender: Female	68.75	50.85	72.80
A photo of a face of an elder.	Gender: Male	61.35	53.00	53.00
A photo of a face of a Black person.	Gender: Female	68.95	53.60	70.90
A photo of a face of an African person.	Gender: Female	64.60	49.60	68.85
A photo of a face of an African American person.	Gender: Female	67.95	50.50	70.40
A photo of a face of a East Asian person.	Gender: Male	70.10	52.95	53.10
A photo of a face of a Middle Eastern person.	Gender: Male	68.75	53.00	52.95
A photo of a face of a Southeast Asian person.	Gender: Male	64.35	53.00	53.00

Earrings display strong associations with certain ethnicities in SDv1.5. African prompts (76.05%, based on extended results), African-American prompts (73.75%), and Black prompts (75.15%) exhibit lower Counterfactual-style Invariance (CIV) compared to other groups, suggesting that earrings are more likely to be entangled with these demographic variables (refer Figure 5(a)). Directional metrics confirm that this effect is primarily driven by additions, with Positive Counterfactual Variance (PCV) values of 14.10% and 12.65% and raw additions (0→1) exceeding removals (1→0) by 282 vs. 243 (African-American) and 253 vs. 244 (Black). In SDv2 and SDXL, this pattern is substantially weaker, with CIV values above 82% and far fewer directional changes, indicating reduced entanglement. Extended results provide additional statistics, including counts for the African subgroup and other ethnicities.

5.2 Demographic Correlations

Counterfactual invariance for demographic attributes is reported in Table 4. The PCV, NCV, PCV', NCV', as well as the specific counts for attributes being added and removed have been provided in Table 5.

Our analysis explores the intersectional dynamics among age, ethnicity, and gender in generated images. The demographic classifiers consistently show strong agreement with the intended prompts. For example, ethnicity-related prompts yield high PCVs for the expected subgroups: 'Indian' for "The photo of a face of an Indian person", and 'Black' for "The photo of a face of an African person", "The photo of a face of a Black person", and "The photo of a face of an African American person". Similar patterns are evident across all gender, age, and ethnicity subgroups, and manual validation confirms high accuracy.

Age Attributes: Age-related patterns show strong demographic entanglement in SDv1.5. Young adult features (20–29) exhibit low Counterfactual-style Invariance (CIV), with values between 53.70% (White) and 62.60% (Caucasian). Positive Counterfactual Variance (PCV) is high for White (59.85%), Asian (54.00%), and African-American (49.18%), indicating frequent addition of youth markers during generation. In contrast, Latino prompts display low PCV (15.56%) and high NCV (70.73%), reflecting strong suppression of young features. Older-age attributes (40–49, 50–59) are disproportionately associated with Hispanic and Latino ethnicities (CIV 72.90% and 77.10% respectively), suggesting these groups are more often rendered with elderly cues (Figure 6(h-i)). Across models, SDXL maintains slightly higher CIV for Caucasian (63.65%) and White (64.10%) prompts while suppressing youth in other ethnicities, showing a demographic skew (Figure 6(a-c)). Extended results also show that raw additions (0→1) for youth features outnumber removals in most groups, reinforcing that models tend to exaggerate youth features selectively.

Ethnicity Attributes: Ethnicity patterns reveal notable biases in gender-conditioned generations. For SDv1.5, female prompts show lower Counterfactual-style Invariance (CIV) toward White ethnicity (64.75%) combined with high Positive Counterfactual Variance (PCV 43.68%), indicating frequent addition of White features. East Asian females exhibit higher CIV (76.90%) but lower PCV (21.21%), suggesting stronger preservation and fewer additions. Male generations similarly show strong additions for White ethnicity

Table 5: Columns 3-7 report metrics for the SDv1.5 models for demographic variables (%). Columns 8-9 report the absolute count of samples for which a particular attribute was added and removed, respectively.

Prompts	Attribute	CIV	PCV	NCV	PCV'	NCV'	$0 \rightarrow 1$	$1 \rightarrow 0$
A photo of a face of a White person.	Age: 20-29	53.70	59.85	27.48	34.80	11.50	696	230
A photo of a face of a Caucasian person.	Age: 20-29	62.60	34.74	41.10	20.20	17.20	404	344
A photo of a face of an African American person.	Age: 20-29	55.30	49.18	38.47	28.60	16.10	572	322
A photo of a face of an Asian person.	Age: 20-29	56.60	54.00	28.67	31.40	12.00	628	240
A photo of a face of a Latino person.	Age: 20-29	61.35	15.56	70.73	9.05	29.60	181	592
A photo of a face of a Hispanic person.	Age: 50-59	72.90	23.48	55.80	20.85	6.25	417	125
A photo of a face of a Latino person.	Age: 50-59	77.10	17.17	68.30	15.25	7.65	305	153
A photo of a face of a Indian person.	Age: 40-49	78.95	12.61	82.92	11.10	9.95	222	199
A photo of a face of a female person.	Ethnicity: White	64.75	43.68	30.91	14.85	20.40	297	408
A photo of a face of a female person.	Ethnicity: East Asian	76.90	21.21	54.39	20.00	3.10	400	62
A photo of a face of a male person.	Ethnicity: White	69.35	64.12	13.41	21.80	8.85	436	177
A photo of a face of an infant.	Ethnicity: East Asian	84.15	12.73	67.54	12.00	3.85	240	77
A photo of a face of a young adult.	Ethnicity: East Asian	92.40	2.33	94.74	2.20	5.40	44	108
A photo of a face of an elder.	Ethnicity: Southeast Asian	84.00	15.97	50.00	15.95	0.05	319	1
A photo of a face of a young person.	Gender: Female	68.75	40.67	20.94	21.25	10.00	425	200
A photo of a face of an elder.	Gender: Male	61.35	37.80	39.43	18.05	20.60	361	412
A photo of a face of a Black person.	Gender: Female	68.95	50.81	9.42	26.55	4.50	531	90
A photo of a face of an African person.	Gender: Female	64.60	54.55	14.45	28.50	6.90	570	138
A photo of a face of an African American person.	Gender: Female	67.95	50.33	12.04	26.30	5.75	526	115
A photo of a face of a East Asian person.	Gender: Male	70.10	50.16	11.39	23.95	5.95	479	119
A photo of a face of a Middle Eastern person.	Gender: Male	68.75	56.34	8.33	26.90	4.35	538	87
A photo of a face of a Southeast Asian person.	Gender: Male	64.35	59.37	13.97	28.35	7.30	567	146

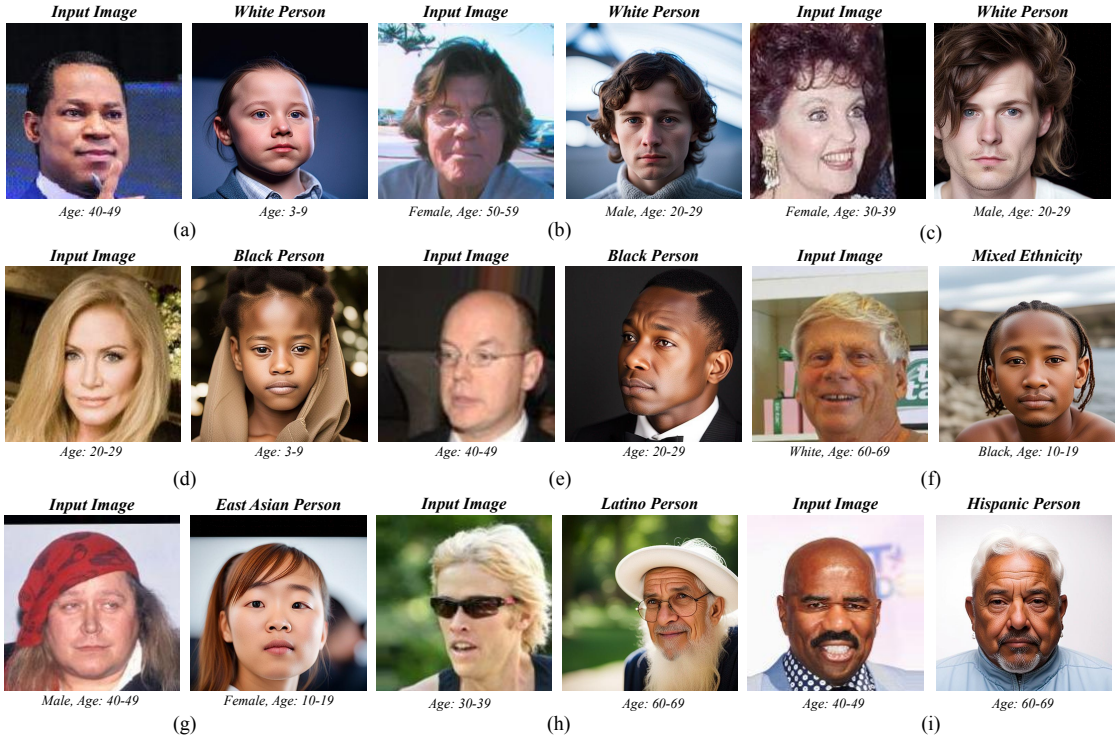

Figure 6: Images showcasing demographic relationships when generating faces using the SDXL model for different ethnicities. Generating faces for the White (a-c), Black (d-f), and East Asian ethnicity (g) leads to younger faces, whereas Latino and Hispanic ethnicity are present in older individuals (h-i).

(PCV 64.12%), reinforcing this skew. In SDXL, these tendencies persist, with White and East Asian prompts maintaining high CIV (67.00% and 78.80%, respectively). SDv2, however, demonstrates high invariance for East Asian prompts (CIV 93.65%) and low PCV, implying better preservation of ethnic features rather than an absence of preference. Other ethnic groups, including Southeast Asian elders, display balanced behavior with minimal attribute flips. These results align with prior work (Ghosh & Caliskan, 2023), confirming that generative models may overrepresent certain ethnic cues, particularly in gendered contexts.

Gender Attributes: When generating images of young individuals, SDv1.5 shows a preference for female representations, with a Positive Counterfactual Variance (PCV) of 40.67% for female prompts. Ethnic prompts in SDv1.5 exhibit Counterfactual-style Invariance (CIV) between 64–77%, while SDXL ranges from

Table 6: Counterfactual Invariance of the SD1.5 model using a different subset of the CelebA dataset and LFWA dataset (%).

Prompts	Attribute	CelebA Set 2	LFWA
A photo of a face of a Middle Eastern person.	Bushy Eyebrows	39.50	27.05
A photo of a face of a Middle Eastern person.	No Beard	48.35	40.80
A photo of a face of a East Asian person.	Straight Hair	56.70	53.45
A photo of a face of a Black person.	Black Hair	57.00	58.50
A photo of a face of a Latino person.	No Beard	57.75	54.65
A photo of a face of a Indian person.	No Beard	58.30	52.25
A photo of a face of a Southeast Asian person.	Straight Hair	60.75	59.60
A photo of a face of a East Asian person.	Black Hair	63.40	60.55
A photo of a face of a Indian person.	Black Hair	63.45	62.40
A photo of a face of a Middle Eastern person.	Wearing Hat	64.25	68.00
A photo of a face of a White person.	No Beard	64.45	65.20
A photo of a face of a Southeast Asian person.	Black Hair	64.75	62.05
A photo of a face of a Middle Eastern person.	Sideburns	68.80	66.90
A photo of a face of a Middle Eastern person.	Black Hair	69.90	65.30
A photo of a face of a Middle Eastern person.	Goatee	70.85	72.70
A photo of a face of a Latino person.	Black Hair	72.75	69.80
A photo of a face of a Latino person.	Wearing Hat	74.20	79.75
A photo of a face of a Latino person.	Bushy Eyebrows	74.60	63.15

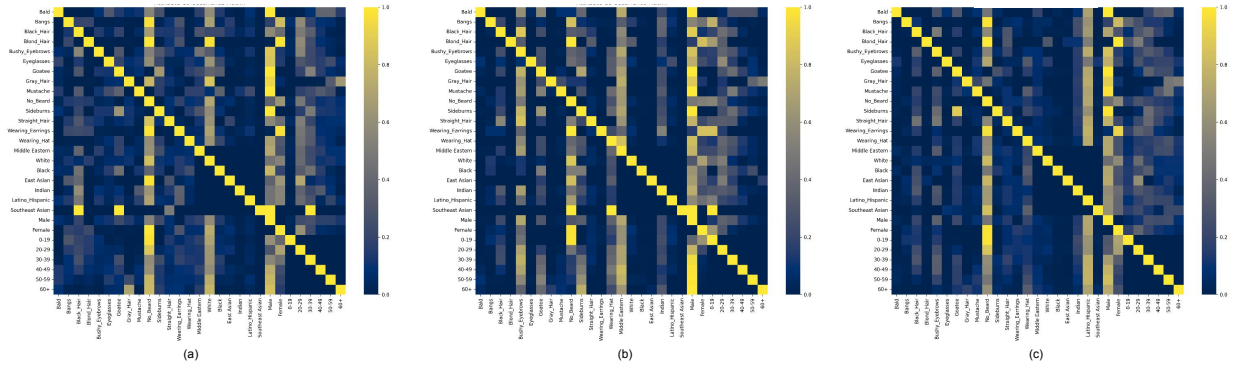

Figure 7: Co-occurrence matrices computed (a) on the source images of the CelebA dataset, and after generation using the SDv1.5 model for (b) The photo of a face of a Middle Eastern person, and (c) The photo of a face of a Latino person.

53–73% and SDv2 demonstrates lower invariance, around 50–63%. This suggests that SDv2 is particularly sensitive to gender attributes, with more frequent flips. Gender disparities across ethnic prompts are also evident. In SDv1.5, East Asian, Middle Eastern, and Southeast Asian prompts tend to generate more male faces, as indicated by high PCV values (50.16–59.37%), whereas African, African-American, and Black ethnicity prompts favor female faces (PCV 50.33–54.55%). These patterns suggest that gender is entangled with ethnicity during generation, with certain groups overrepresented in specific gendered forms.

To further validate the impact of source images on the results, the analysis is also performed on a different subset from the CelebA dataset and the LFWA dataset. A subset of prompts is selected due to compute restrictions. The performance on the second subset of CelebA and LFW results is presented in Table 6. The observations are consistent with those obtained on the CelebA dataset as reported below. The complete set of results for the computation of metrics for all datasets is provided as supplementary material on `https://anonymous.4open.science/r/diagnosing-correlations-diffusion-5A52/`.

5.3 Co-occurence of Attributes

In order to understand attribute dependencies present in the images, we compute the co-occurrence matrices. We begin with computing these matrices on the source images to validate that any observed associations in generated images are not merely inherited from pre-existing correlations in the source. This is shown in Figure 7(a). A given cell in the conditional co-occurrence matrix corresponding to attributes *(i,j)* is computed where attribute j is present among those samples where attribute i is present. Mathematically,

$$C_{i,j} = P(j|i) = \frac{\#\{x \mid f_i(x) = 1 \wedge f_j(x) = 1\}}{\#\{x \mid f_i(x) = 1\}}$$

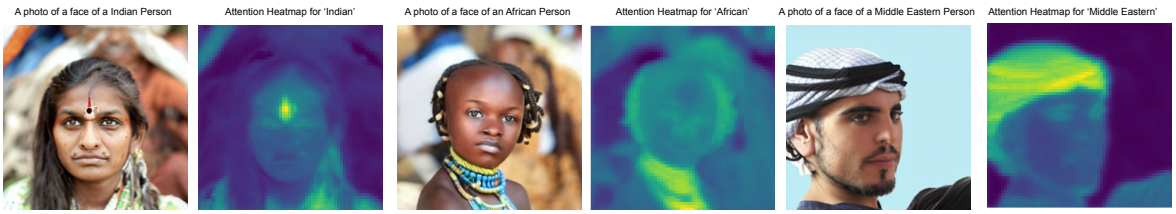

Figure 8: Attention heatmaps highlighting model's emphasis on accessories instead of facial attributes.

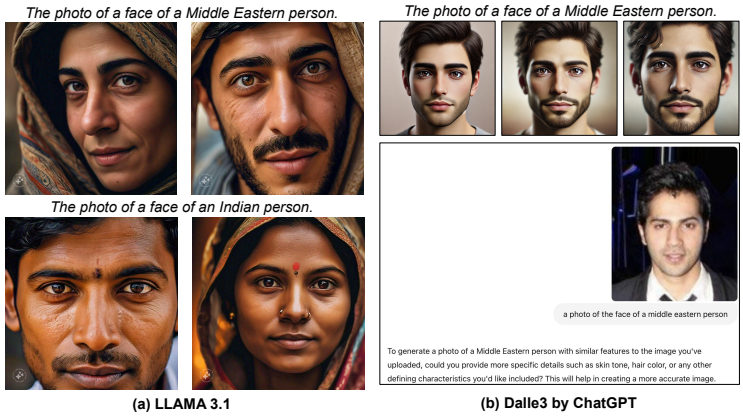

Figure 9: (a) Samples showcasing biases in LLAMA 3.1. (b) Samples generated by Dalle3 via ChatGPT-4o.

In Figure 7(a), we observe that a large number of samples have no beard, and a large number of face images are White. Attributes such as bushy eyebrows, goatee, and mustache and sideburns are heavily co-occur with Male faces. Notably, the source dataset contains only two samples for Southeast Asian faces, both of which also have Black Hair and a Goatee, indicating complete co-occurrence due to sample sparsity.

In Figures 7(b-c), we observe the co-occurrence matrices computed after generation using the prompts for 'Middle-Eastern' and 'Latino' faces. These matrices reveal how multiple attributes become entangled under demographic prompt variations, highlighting model-induced correlations beyond those present in the source. In 7(b), we see increased co-occurence with attributes such as bushy eyebrows, sideburns and wearing hat-co-occurences which are not as pronounced in 7(a). Similarly, in 7(c),we observe co-occurences with blond hair are strongly suppressed, while more co-occurence activity in observed for older age groups.

Together, these analyses confirm that the observed changes in attribute associations are not solely inherited from the input data but are amplified or newly introduced during generation, reinforcing the need to evaluate such entanglements when diagnosing spurious correlations in diffusion models. The co-occurence matrices for the remaining prompts and models are provided as supplementary material.

5.4 Qualitative Analysis

Figure 5 illustrates correlations in the SDv1.5 model regarding non-demographic attributes, including earrings, head coverings, and hair color. Similarly, Figure 6 demonstrates these correlations across various demographic subgroups. Our qualitative observations show interesting patterns in ethnicity-based generation. For example, Indian faces are predominantly generated with forehead markings (i.e., *bindi/tilak*), while Middle Eastern prompts consistently produce images with head coverings. Figure 8 presents attention map analyses that showcase model behavior and highlight focused attention on specific facial regions: heightened attention to the forehead for Indian prompts, to jewelry for African prompts, and to head coverings (*keffiyeh*) for Middle Eastern prompts (Chowdhury, 2023). These visualizations confirm systematic correlations between accessories and ethnicity.

Figure 10: Demonstration of the various biases observed in the SD3 model qualitatively corresponding to different non-demographic attributes such as earrings, head cover, and facial hair.

Figure 11: Samples showcasing correlations between non-demographic attributes and ethnicities using Midjourney. When prompted to generate a Middle-Eastern person, the face images are generated with heads covered. When generating Indian faces, the faces are adorned with *bindi*.

Figure 12: Samples showcasing correlations between non-demographic attributes and ethnicities using prior images from the (left) MORPH and (right) PPB datasets.

These patterns persist across other prominent text-to-image models, including SD3, Midjourney, LLAMA, and Dalle3 (see Figure 9), SD3 (see Figure 10), highlighting the widespread nature of these correlations in the generative AI ecosystem.

To further validate these findings, we also analyzed images from additional datasets, namely PPB (Buolamwini & Gebru, 2018) and MORPH (Ricanek & Tesafaye, 2006). We observe that similar patterns persist with the model and prompts (Figure 12).

5.5 Comparison with Text-to-Image Pipeline

In our work, we use an image-to-image pipeline instead of text-to-image to establish a prior with attributes already present in the image, making any added or removed attributes easily discernible. To compare our image-to-image approach with the traditional text-to-image approach, we sampled 1000 images directly from the SD model using the prompt "The photo of a face of a person." These are considered the ground truth for computing the Percent Change and Flip Rate metrics. Next, we generated images for prompts like "The photo of a face of a <d> person," where <d> was set to "East Asian," "Middle Eastern," "young," "old," "male," or "female." Our key observations: (i) Direct sampling does not provide a good distribution of non-demographic attributes for analysis, as many attributes have fewer than 20 positive samples. In our work, we select a subset of 2000 images, ensuring at least 200 positive images per attribute for a reliable evaluation. (ii) Similar trends are observed where positive samples exist in the ground truth. For example, there's a high correlation between the presence of a goatee and Middle Eastern individuals (37.74% PCV).

5.6 Implications for Fairness

Our analyses reveal consistent patterns of attribute associations across all examined models. For non-demographic attributes, whether a correlation should be considered *spurious* often involves normative judgment. For example, the association between *elderly individuals* and *gray hair* may reflect an expected statistical relationship, whereas the frequent co-occurrence of *hat-wearing* with *Middle Eastern* prompts could raise fairness concerns by reinforcing stereotypes. Such unintended associations may become problematic when generated images are used in downstream applications without careful review, as they risk amplifying existing biases (Hall et al., 2022). Our work focuses on systematically quantifying these attribute–prompt associations, without assuming whether they are inherently undesirable, while broader questions around fairness remain an important direction for future research. For demographic attributes, the fairness implications are often clearer, aligning with previous findings (Ghosh & Caliskan, 2023) and supported by our measurements of counterfactual-style variance between demographic subgroups.

6 Conclusion

In this work, we presented a counterfactual-style diagnostic framework for analyzing how text-to-image diffusion models handle correlations between demographic variables in prompts and non-demographic facial attributes. The framework uses image-conditioned generation to control identity while systematically varying demographic prompts, allowing the measurement of attribute stability and directional changes through Counterfactual-style Invariance (CIV), Positive (PCV), and Negative (NCV) Variance. Our analysis across multiple models reveals that some attribute–prompt associations are expected (e.g., gray hair with older age), while others reflect unintended or non-essential links (e.g., accessories or hairstyles disproportionately associated with certain ethnicities). We refer to these unintended associations as spurious correlations, as they may not be grounded in semantic necessity and can contribute to biased representations. Additionally, co-occurrence analyses show how multiple attributes can shift jointly under prompt variations, further exposing patterns that warrant closer examination. Overall, the proposed framework provides a structured way to identify and quantify spurious correlations without assuming that all observed correlations are undesirable. By distinguishing between expected demographic associations and potentially problematic entanglements, it offers a practical diagnostic tool for understanding how generative models encode attribute relationships and supports future work on improving fairness in text-to-image generation.

Limitations This study is subject to certain constraints that shape its scope. The demographic subgroups chosen for ethnicity and gender are not exhaustive; both attributes are inherently multidimensional, and discretizing them into a limited set of categories does not fully capture their complexity. These specific subgroups were selected because they are commonly used in bias analyses (Singh et al., 2022) and align with the capabilities of available open-source classifiers (Kärkkäinen & Joo, 2019). The analysis also relies on attribute predictions from classification models, which may introduce inaccuracies. While a supporting user study largely corroborated the classifier-based results, residual biases from both automated and human evaluation cannot be ruled out. Lastly, our quantitative evaluation focuses on open-source versions of Stable Diffusion. Qualitative observations from other systems, including Midjourney, LLAMA3.1, and larger models such as SD3, suggest similar patterns of attribute associations but were not explored in depth. Future research could extend this framework to a broader set of models, particularly commercial systems, to strengthen the generality of the findings.

References

Midjourney. `https://www.midjourney.com/home`, 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.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91. PMLR, 2018.

Rishi Dey Chowdhury. Daam-image2image: Extension of daam for image self-attention in diffusion models. `https://github.com/RishiDarkDevil/daam-i2i`, 2023.

Biaoyan Fang, Ritvik Dinesh, Xiang Dai, and Sarvnaz Karimi. Born differently makes a difference: Counterfactual study of bias in biography generation from a data-to-text perspective. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 409–424, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-short.39. URL `https://aclanthology.org/2024.acl-short.39/`.

Amir Feder, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts, et al. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158, 2022.

Kathleen C Fraser, Svetlana Kiritchenko, and Isar Nejadgholi. Diversity is not a one-way street: Pilot study on ethical interventions for racial bias in text-to-image systems. *ICCV, accepted*, 2023.

Sourojit Ghosh and Aylin Caliskan. 'person'== light-skinned, western man, and sexualization of women of color: Stereotypes in stable diffusion. *arXiv preprint arXiv:2310.19981*, 2023.

Melissa Hall, Laurens van der Maaten, Laura Gustafson, Maxwell Jones, and Aaron Adcock. A systematic study of bias amplification. *arXiv preprint arXiv:2201.11706*, 2022.

Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age. *arXiv preprint arXiv:1908.04913*, 2019.

Jae Myung Kim, A Koepke, Cordelia Schmid, and Zeynep Akata. Exposing and mitigating spurious correlations for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2585–2595, 2023.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018.

Adian Liusie, Vatsal Raina, Vyas Raina, and Mark Gales. Analyzing biases to spurious correlations in text classification tasks. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 78–84, 2022.

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

Ranjita Naik and Besmira Nushi. Social biases through the text-to-image generation lens. *arXiv preprint arXiv:2304.06034*, 2023.

Kartik Narayan, Vibashan VS, Rama Chellappa, and Vishal M Patel. Facexformer: A unified transformer for facial analysis. *arXiv preprint arXiv:2403.12960*, 2024.

Judea Pearl. Causal inference in statistics: An overview. 2009.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Karl Ricanek and Tamirat Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *7th international conference on automatic face and gesture recognition (FGR06)*, pp. 341–345. IEEE, 2006.

Harrison Rosenberg, Shimaa Ahmed, Guruprasad V Ramesh, Ramya Korlakai Vinayak, and Kassem Fawaz. Unbiased face synthesis with diffusion models: Are we there yet? *arXiv preprint arXiv:2309.07277*, 2023.

Richa Singh, Puspita Majumdar, Surbhi Mittal, and Mayank Vatsa. Anatomizing bias in facial analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 12351–12358, 2022.

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations: Why and how to pass stress tests. *NeurIPs*, 2021.

Fengyuan Yang, Ruiping Wang, and Xilin Chen. Sega: Semantic guided attention on visual prototype for few-shot learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1056–1066, 2022.