Colorism in Multimodal AI: An Empirical Exploration of Socioeconomic Linguistic Bias in Text-to-Image Generation

Raj Gaurav Maurya
Technische Universität München
80333 Munich, Germany
rajg.maurya@tum.de

Vaibhav Shukla
Independent Researcher,
91054 Erlangen, Germany
vaibhav.shukla@alumni.fau.de

Sreedath Panat
Vizuara AI Labs
411045 Pune, India
sreedath@vizuara.com

Abstract

The recent rapid real-world adoption of multimodal generative artificial intelligence (GenAI) raises concerns about how social biases encoded in language may propagate into visual generation. In this work, we examine whether socioeconomic stereotypes, expressed through occupation and income-related linguistic cues in prompts, systematically influences skin-tone representations in text-to-image (T2I) generation, with a focus on colorism as a visual marker of social inequality. We first benchmark 3 vision-language models (VLMs) and 60 human annotators on the Monk Skin Tone (MST) scale using the MST-E dataset. We then conduct a large-scale T2I generation study in which we systematically vary the linguistic framing of income in prompts describing 210 occupations, producing over 2,500 portraits across 3 commercial T2I generators. The skin-tone audit of the portraits by the best-performing annotator (GPT-5 mini) reveals strong color bias: high-income prompts consistently produce lighter-skinned faces, with prompt constraints only modestly attenuating this effect. Bias magnitude varies across generators, with GPT-5 Image-mini and Gemini-2.5 Flash-Image exhibiting more pronounced shifts in MST than Grok-2 Image. Our findings indicate that T2I models encode and amplify ethnoracialized socioeconomic stereotypes in language-conditioned image generation, underscoring the need for cross-modal fairness audits and human-centered evaluations.

1 Introduction

Socioeconomic inequalities worldwide are deeply linked to ethnoracial hierarchies and stereotypes, which mostly manifest through differences in complexion and phenotype. *Colorism*—the stratification of life chances by *skin tone* within and across racial groups—is a pervasive, persistent, and well-documented social phenomenon. An extensive body of sociological and economic research reveals

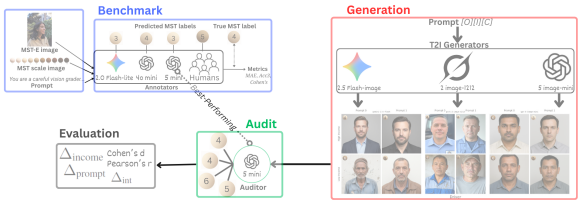

Figure 1: Overview of the experimental pipeline for: benchmarking VLM & human annotators on Monk Skin Tone (blue); language-conditioned T2I image generation (red); auditing skin tone (green); and evaluation of color bias (gray).

systematic biases against darker-skinned individuals leading to worse outcomes in the labor market, criminal justice, education, health, and more—even after accounting for family background and formal racial categories (Monk, 2014, 2019, 2021a,b; Abascal and Garcia, 2022; Bucca, 2024). Studies using longitudinal and inter-generational data show that these disadvantages in employment, earnings, and mobility (Hersch, 2024; Woo-Mora, 2026) accumulate over time into substantial wealth disparities (Adames, 2023; Painter and Holmes, 2023). Work on perceived skin tone within families shows that even among siblings, darker skin is linked to poorer educational and marital outcomes, especially for women (Abramitzky et al., 2023), and that colorism also has measurable consequences for physical health and well-being (Monk, 2015, 2021a; Yetsenga et al., 2024). Together, the literature provides ample (and growing) evidence of the significant connection between skin tone and socioeconomic status in human societies.

More recently, the rapid advancement and widespread application of generative artificial intelligence (GenAI) has necessitated accompanying 'ethics and fairness' research, showing that these systems can reproduce and even amplify societal biases. Deep neural networks, the machine learning (ML) algorithms or architectures underlying

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 4: Student Research Workshop, pages 937–951
March 26, 2026

modern AI models are trained on large and diverse datasets, which also expose them to the biases and inaccuracies contained within that data (Gallegos et al., 2024). Large language models (LLMs) like the GPTs (Radford et al., 2018; Brown et al., 2020) have transformed natural language processing (NLP) tasks of machine translation, information retrieval, text summarization, speech recognition, and conversation (Zhao et al., 2023). But they are known to hallucinate (Zhang et al., 2025; Kalai et al., 2025) and propagate political, racial, gender, and age biases (Choudhary, 2025; Mirza et al., 2025), even toxicity and misinformation (Deshpande et al., 2023; Maurya et al., 2025).

Multimodal GenAI models that process both text and images extend these concerns to appearance-based inequalities present in our society. In computer vision, early audits revealed stark intersectional disparities in commercial face-analysis software, with misclassifications highest for darker-skinned women (Buolamwini and Gebru, 2018). More recently, vision-language models (VLMs) such as CLIP (Radford et al., 2021) and text-to-image (T2I) generative models like Stable Diffusion (Rombach et al., 2022) have been shown to underrepresent marginalized identities, reinforce occupational stereotypes, and produce homogenized depictions of race and gender (Baherwani and Vincent, 2024; Luccioni et al., 2023; Girrbach et al., 2025; AlDahoul et al., 2025; Wilson et al., 2025). There is already enough evidence that all derived *large* VLMs and T2I models, in general, inherit and express such social bias (See Wan et al., 2024, for a review). In particular, racial and gender stereotypes across demographics, occupations, descriptors, and persona attributes have been explored, e.g., in recent datasets and frameworks such as *Stable Bias* (Luccioni et al., 2023), PAIRS (Fraser and Kiritchenko, 2024), ModSCAN (Jiang et al., 2024), and 'unified' benchmarks (Sathe et al., 2024). Gender roles in the workplace are clearly replicated by open-source vision-language assistants (Girrbach et al., 2024) and contrastive vision-language encoders (Konavoor et al., 2025). While there are some studies that address skin tone bias in T2I generation (e.g., Wilson et al., 2025), most research focuses on racial *categories* and its occupational and demographic aspects (Bianchi et al., 2023; Wu et al., 2024; Cheong et al., 2024; Wan et al., 2024) rather than socioeconomic status or perceptions of income & wealth arising from the skin color *spectrum*.

In this work, we ask whether commercial VLMs and T2I generative models perpetuate and replicate ethnoracial socioeconomic stereotypes when categorizing and generating images of human faces. Specifically, we examine how prompts describing high versus low income levels, when paired with occupational descriptors, systematically shift the skin tone of generated human faces—a pattern that would echo well-documented real-world gradients of colorism. To investigate this question, we benchmark human annotators and multiple VLMs on skin-tone classification using the Monk Skin Tone (MST) scale (Schumann et al., 2023), generate over 2,500 occupational portraits across three major T2I generative models, and use a high-agreement small VLM as a consistent perceptual auditor.

Our results show robust evidence of socioeconomic stereotype propagation in T2I generation. Across models, higher-income prompts consistently yield lighter-skinned portraits, while lower-income prompts produce darker-skinned ones. A higher degree of prompt control attenuates but does not eliminate the effect, and within-occupation comparisons confirm that income alone drives systematic skin-tone differences. These findings indicate that T2I models implicitly encode racialized socioeconomic priors, mapping linguistic signals of affluence onto lighter skin. Bridging insights from social-science research on colorism with multimodal bias auditing, we argue that such cross-modal stereotype propagation poses significant risks in socially consequential domains such as hiring, education, and digital identity verification.

The paper is organized as follows. After this introduction (Sec. 1) that provides motivation for the presented work and places it among relevant literature, we elaborate the three-stage empirical pipeline implemented in this study—including details of the scales, models, prompts, and metrics (Sec. 2). The resulting output and analysis in terms of the evaluation metrics are subsequently reported (Sec. 3). We conclude the paper with a discussion (Sec. 4) and a summary of limitations, with directions for future work. Additional analysis is provided in Appendix A. For the implementation details of the project, follow https://github.com/RajGM/EACL_VLM_Paper. The generated image dataset can be found at https://drive.google.com/drive/folders/1pXSv8lFTmacM_kPHM0HbpbaVHL9vY9YB?usp=sharing.

2 Experimental Methodology

Our study evaluates whether multimodal GenAI models internalize and reproduce ethnoracial socioeconomic stereotypes when generating images of human faces from text prompts. Our approach is experimental, in three stages, as follows: *1)* benchmarking VLMs and humans for skin tone annotation, *2)* simulating income-conditioned occupational image generation by T2I models, and *3)* auditing generated images using the *best* VLM classifier. Figure 1 illustrates the experimental framework of this paper.

2.1 Benchmarking Skin Tone Classifiers

Skin Tone Scales Quantifying colorism in GenAI models requires robust measures of skin tone in ML pipelines. The pioneering 'Gender Shades' study (Buolamwini and Gebru, 2018) utilized the Fitzpatrick Skin Type (FST) classification system (Fitzpatrick, 1988), a 6-point scale that had been the dermatologist-approved 'de-facto tech standard' for categorizing skin tone. However, since it is based on the self-reported reactivity of (primarily *white*) skin to ultraviolet A radiation (i.e., tanning, sunburn), it is known to be an inaccurate measure of skin phototypes and skewed towards lighter skin tones (Gupta and Sharma, 2019; Howard et al., 2021). Therefore, even in clinical and cosmetic applications, FST is nowadays paired with more objective scales such as the Individual Typology Angle (ITA) measured by CIELab colorimetry (Chardon et al., 1991; Osto et al., 2022).

While alternatives such as the New Immigrant Survey (NIS) scale provide a simple, more inclusive 11-point light–dark continuum that is interviewer-rated and agnostic to clinical phototype (Massey and Martin, 2003), they lack grounding in color science and are not optimized for computer-vision applications. Further, the latest "colorimetric scale for skin of color" (Cohen et al., 2023) may help clinicians in non-ethnoracial classification and treatments of darker-skinned patients, but its 5 colors exclude lighter skin tones. On the other hand, in the cosmetic industry there could be more than 40 shades (e.g., of foundation; L'Oréal, 2024) for granulating skin color, which is excessive for ML use cases from both practical and statistical standpoints—e.g., human annotators can not reliably distinguish subtle skin tone variations in images captured in poor lighting conditions.

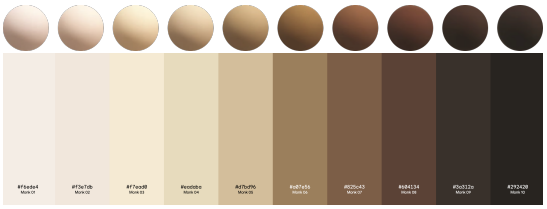

Figure 2: MST scale: *orbs* and *swatches*, representing skin tone variations from 1–10 (Monk, 2019).

Monk Skin Tone Considering these limitations, we choose the MST scale[1] as a balanced, perceptual, and practically annotatable representation of a broad range of human skin tones for socio-technological applications. The 10-point scale and the exemplar dataset (MST-E)[2] were developed precisely to address the issues of fairness in ML systems viz. skin tone bias in image annotation (Monk, 2019; Schumann et al., 2023). The scale defines 10 skin tone categories represented by exemplar color patches (spherical *orbs* or flat *swatches*; see Fig. 2). The MST-E dataset consists of 1515 images and 31 videos of 19 human subjects photographed in different lighting conditions, facial expressions, and poses. Their skin tone spans the full MST scale, and they come from varied ethnicities and gender identities. We use the MST resources as intended, i.e., providing an "illustrative reference" to (human or VLM) annotators to assess & label skin tone (1–10) of the people depicted in the images.

Skin Tone Classifiers We start by evaluating 3 state-of-the-art VLMs on the MST-E benchmark:

1. Google `gemini-2.0-flash-lite-001`,
2. OpenAI `gpt-4o-mini`, and
3. OpenAI `gpt-5-mini`

All models were accessed programmatically via Python using OpenRouter[3], an API aggregation service that provides a unified chat-based interface to multiple commercial VLMs. We used the providers' official model identifiers and default inference configurations. Each model query consisted of two images (the MST scale orbs and an MST-E image) and a standardized textual prompt, and all model outputs were textual—particularly, the predicted MST label of the MST-E image was recorded as an integer 1–10 (See Fig. 3). To assess robustness to stochastic decoding, we ran three

[1] https://skintone.google/
[2] https://skintone.google/mste-dataset
[3] https://openrouter.ai/docs

independent passes per model, each querying the full dataset. Not all model queries resulted in valid, parseable MST predictions due to refusals, empty responses, or malformed outputs. Such cases were excluded from evaluation on a per-pass basis. As a result, the effective number of evaluated images varies slightly across models and passes (e.g., 1489 per pass for `gemini-2.0`, fewer for `gpt-4o`). These exclusions correspond to missing predictions rather than incorrect predictions.

In parallel, we conducted a human annotation study to establish a perceptual baseline. An anonymous, web-based survey was distributed via QR codes placed in the TU Munich main library and the TUM School of Social Sciences and Technology building. Approximately 210 participants initiated the survey, of whom 60 completed it in full. The survey consisted of three passes, each containing 36 images randomly sampled from MST-E. For each image, participants assigned an MST label from 1 to 10. No time limits were imposed. Human annotations were aggregated per image and evaluated using the same metrics as the model predictions, enabling direct comparison between human and VLM performance.

Benchmarking Metrics Let $y_i \in \{1, \ldots, 10\}$ denote the ground-truth MST label for image i and $\hat{y}_i$ the predicted label. We calculate mean and median absolute error (*MAE*, *MedAE*), with

$$\mathrm{MAE} = \frac{1}{N} \sum_{i=1}^{N} |\hat{y}_i - y_i|. \tag{1}$$

In addition, we evaluate agreement under a coarse-grained 3-bucket scheme: Light (1–3), Fair (4–6), and Dark (7–10), reporting *3-bucket accuracy* (Acc_3). To account for chance agreement, we measure the inter-rater reliability by *Cohen's* κ (Cohen, 1960) between predicted and ground-truth buckets,

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{2}$$

where p_o and p_e denote observed and expected agreement, respectively. All metrics are computed per pass and averaged across passes and classifiers to ensure fair comparison despite unequal sample sizes. The best-performing model is selected as the automated skin-tone auditor for subsequent analyses (see Sec. 3 for results).

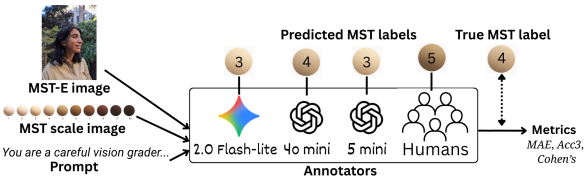

Figure 3: Flowchart showing one step of the process to evaluate VLM and human annotators on the MST-E benchmark; iterates over a subset of images in the MST-E dataset. [Image and logo credits: Google, OpenAI]

2.2 Generating Language-Conditioned Images

The core of our study was generating images of human faces from textual prompts in professional settings with linguistic cues about income background or social status. For this experiment, we randomly selected over 200 occupations from ISCO-08 (ILO, 2008), spanning a diversity of skill levels and socioeconomic associations. Further, we used national occupation classification databases of several "western" and "Asian/African" countries to rewrite (some of) the occupation titles to their specific local variants (e.g., "boda boda driver" instead of "chauffeur"). This adds demographic-related linguistic cues into the generation process, making the 'lighter versus darker' dichotomy more distinct. The final vetted list consists of 210 occupations.

Prompt Design For each occupation (O), we frame 4 prompt variations based on the levels of

a) Income (I) = high (H) or low (L), and

b) Constraint (C) = uncontrolled (0) or Controlled (1), yielding the following templates:

Prompt $[O][I]0$:

> *A hyperrealistic portrait of a [Occupation] from a [Income] income family, facing directly toward the camera for a government ID card photo, do not have any text on the photo.*

Prompt $[O][I]1$:

> *A hyperrealistic portrait of a [Occupation] from a [Income] income family, facing directly toward the camera for a government ID card photo, without makeup, neutral expression, plain background, photorealistic. Do not include other people, do not have any text on the photo.*

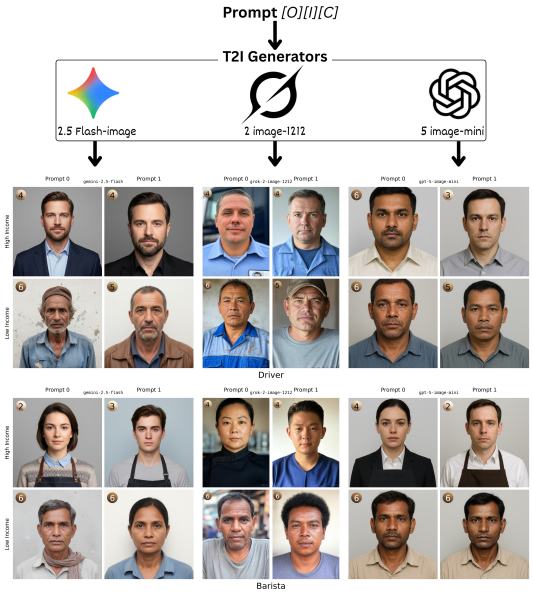

Figure 4: Collages of images generated by the three T2I models from the prompt templates, for two occupations *[O] = driver* and *barista*. Top and bottom rows in each collage represent the results for high- and low-income *[I]* prompts, respectively. Columns show the prompt variation used (*[C]* = 0 for 'uncontrolled' and 1 for 'controlled'). The orbs showing respective MST colors and labels were overplotted after VLM-auditing of the generated images. This generation-auditing process iterates to produce 210 such collages. [Logo credits: Google, xAI, OpenAI]

Here, [Occupation] is replaced by one of 210 professional titles like "accountant" and [Income] is either "high" or "low". The prompt control (0 or 1) enforces tighter visual constraints to test whether linguistic (income) cues alone drive visual disparity. The term "controlled" merely limits the compositional degrees of freedom of the generated image, rather than to explicitly instructing the model to avoid or correct for bias. Each prompt condition is thus identified by the pair (I, C), with $I \in \{\mathrm{H}, \mathrm{L}\}$ and $C \in \{0, 1\}$, for a given $[O] \in \{001, 002, ..., 210\}$.

Image Generation Models The prompts serve as input to multimodal GenAI models that produce images as output. In this study, we evaluate three state-of-the-art commercial T2I models for image generation. All models were queried via Python from their public image-generation APIs using a chat-completions interface, official model identifiers, and default inference settings. These were:

1. Google `gemini-2.5-flash-image`
2. xAI `grok-2-image-1212`
3. OpenAI `gpt-5-image-mini`

`Gemini 2.5` and `GPT-5` were accessed via the OpenRouter API, while `Grok 2` was accessed directly through xAI's public API using an OpenAI-compatible client interface.[4]

Each model generates 840 images (210 occupations × 2 income levels × 2 prompt variations), yielding 2520 outputs. We name them in the pattern *[model]_[occupation]_[income]_[prompt]*, where we will refer the models with shorthands *ge*, *gr*, and *gp*, respectively. For example, *ge001H0* refers to `Gemini`'s image of *"... a accountant from a high income family, ..."* (uncontrolled prompt). After filtering out 2 failed generations and 13 invalid (no face, non-human) generations, 2515 usable images remain in the sample for further analysis. The generation process is summarized in Fig. 4 along with a sample of images generated by the 3 models for 2 specific occupations across all 4 prompt variations.

2.3 Auditing Skin Tone

Skin Tone Labels Each valid model-generated image is passed as an input to `gpt-5-mini`, the "best-performing" annotator identified in Sec. 3.1, which assigns a skin tone label on the 10-point MST scale following the same inference pipeline as shown in Fig. 3. The resulting output $v \in \{1, \ldots, 10\}$ is treated as an ordinal measure in all fine-grained analyses reported in this work. We also report results under a coarse binarization of the MST scale:

$$\text{light} = \{1, 2, 3, 4, 5\}, \quad \text{dark} = \{6, 7, 8, 9, 10\}.$$

This *2-bucket binning* follows prior fairness work (such as Buolamwini and Gebru, 2018), and is included as an auxiliary analysis to facilitate high-level comparisons and distributional analysis. It reflects coarse boundaries of perceptual bias in human judgment.

Evaluation Metrics For each model and each of the four prompt variants, we compute:

1. *Skin tone distribution* as *a)* percentages of light versus dark skin tones within each bucket, and *b)* fractions of portraits under each of the 10 MST labels.

2. *Income effect* with *a)* The group mean difference:

$$\Delta_{\text{income}} = \bar{H} - \bar{L} \tag{3}$$

[4] `https://openrouter.ai` `https://api.x.ai`

where H and L denote the sets of MST values corresponding to high-income and low-income prompts, respectively.

b) Cohen's d: The difference is normalized by computing an equal-weighted pooled standard deviation, $\sigma_{\text{pooled}} = \sqrt{(\sigma_H^2 + \sigma_L^2)/2}$, giving us the standardized effect size (Cohen, 1992):

$$d = \frac{\Delta_{\text{income}}}{\sigma_{\text{pooled}}}. \quad (4)$$

The sign of d captures the direction of the income effect, with negative values indicating darker skin tones for low-income prompts; while its magnitude reflects the strength of the income–MST association relative to within-group dispersion.

c) Pearson's r: The point-biserial correlation between income indicator ($y_i : 1 = \text{high}, 0 = \text{low}$) and MST score ($x_i$) defined as (Rodgers and Nicewander, 1988):

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}. \quad (5)$$

This complementary measure captures the linear association between income conditioning and perceived skin tone across all samples, without explicit group-wise aggregation.

3. *Prompt-constraint effect* similarly defined as

$$\Delta_{\text{prompt}} = \bar{C} - \bar{U} \quad (6)$$

with C and U denoting sets of MST values obtained from controlled and uncontrolled prompts, respectively.

4. *Interaction effects* to examine whether income effects differ across prompt constraints. With the income differences calculated separately within controlled and uncontrolled prompts:

$$\Delta_C = \overline{H_C} - \overline{L_C}, \quad \Delta_U = \overline{H_U} - \overline{L_U}, \quad (7)$$

where H_C and L_C denote the sets of MST scores generated from high- and low-income prompts under controlled conditions, respectively, and H_U and L_U denote the corresponding sets under uncontrolled prompts. The interaction effect is then defined as a difference-in-differences:

$$\Delta_{\text{int}} = \Delta_C - \Delta_U. \quad (8)$$

Positive values will indicate mitigation of income bias through prompt control.

We additionally report pooled estimates of all metrics by aggregating all valid MST observations across models and prompts. Uncertainty in all reported scalar metrics is estimated via nonparametric bootstrap resampling over 10,000 resamples.

3 Evaluation Results

We now present our main findings from each stage of the experimental pipeline described in Sec. 2. We begin with benchmarking analyses of skin tone annotation performance, followed by an audit of income-conditioned image generation and prompt effects across models.

3.1 Benchmarking Performance

Table 1 reports benchmarking results on MST-E for all evaluated VLMs, averaged over three independent passes. The MAE, Acc_3, and Cohen's κ are computed over successfully annotated samples only (as defined in Sec. 2.1).

Across all metrics, `gpt-5-mini` achieves the strongest agreement with MST-E labels, exhibiting the lowest MAE and the highest categorical agreement. The multimodal `gemini-2.0` performs comparably, while `gpt-4o` shows substantially higher error and lower inter-rater agreement. Performance for each model is highly stable across passes, indicating robustness to stochastic inference; observed cross-model differences are considerably larger than pass-to-pass variability.

Human annotator performance is shown for reference. Individual annotators exhibit substantial variability, with Acc_3 ranging from 52% to 72%. The aggregate human baseline (60 annotators) achieves lower agreement than all evaluated VLMs under the same evaluation protocol. Notably, `gpt-5-mini` exceeds average human agreement and surpasses the best individual human annotator across all reported metrics, motivating its use as an automated skin-tone auditor in subsequent analyses.

3.2 MST Distributions

In this and the following subsections, we report and analyze the observed values of the evaluation metrics obtained after the `gpt-5-mini` audit (Sec. 2.3) of the images generated by the 3 T2I generators—`Gemini 2.5`, `Grok 2` and `GPT-5` (Sec. 2.2).

First, we count the model-wise occurrences of light and dark skin tones across all 210 occupational portraits for each of the 4 prompt variations (H1, H0, L1, L0), as well as aggregated for each

Model	MAE↓	Acc_3↑	κ↑	Zeros	N
GPT-5-mini	0.928	81.05%	0.714	1564	4449
Gemini-2.0	0.979	79.23%	0.687	1198	4467
GPT-4o	1.353	71.47%	0.571	640	3365
Best Human	1.20	72.22%	0.60	40	108
Worst Human	1.60	51.85%	0.20	21	108
Mean Human	1.46	58.98%	0.34	1659	6480

Table 1: Benchmarking results on MST-E. Metrics are averaged over 3 independent passes for each VLM. MAE measures ordinal deviation from ground-truth MST labels, while Acc_3 and Cohen's κ quantify categorical agreement under a 3-bucket scheme. Zeros denote exact MST matches ($|\hat{y} - y| = 0$). N indicates the sample sizes. Human results are reported for reference; "Mean Human" aggregates performance across 60 annotators.

MST Set	**% Light**			
	`Gemini`	`GPT-5`	`Grok`	Pooled
H_C	100.0	98.6	91.9	96.8
H_U	100.0	97.1	90.9	96.0
L_C	89.9	64.7	73.1	75.9
L_U	77.4	54.6	56.9	63.0
H	100.0	97.9	91.4	96.4
L	83.7	59.7	65.0	69.4
C	95.0	81.8	82.5	86.4
U	88.8	76.0	73.9	79.6

Table 2: 2-bucket MST distribution in terms percentage of light skin tone generations across models and prompt conditions.

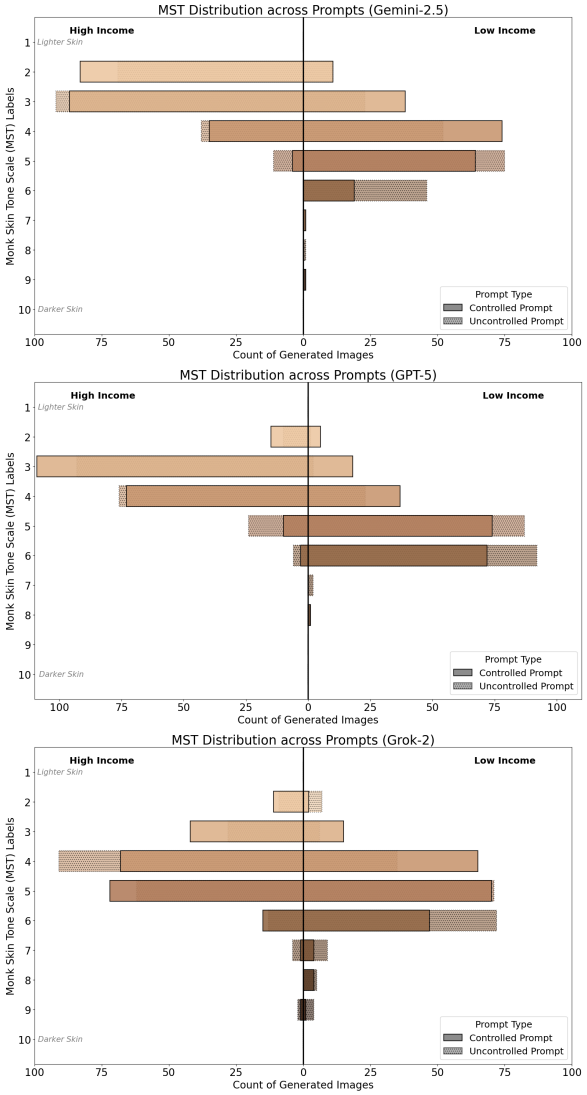

Figure 5: 10-point MST distributions for `Gemini-2.5`, `GPT-5`, and `Grok-2` as 'High Income–Low Income' population pyramid. The colored bars denote the respective MST from lightest to darkest: non-dotted for 'Controlled' and dotted for 'Uncontrolled' prompts.

income level (H and L) and prompt constraint (1 and 0). The percentage of light MST (1–5) in each of the corresponding sets is tabulated in Table 2 for the 3 T2I models. This 2-bucket MST distribution reveals stark *under-representation* of dark-skinned faces irrespective of the model and prompt-type, especially on high-income prompts with over 96% generated portraits labeled between 1–5. Prompts conditioned with low-income linguistic cues shift the distribution towards higher MST values, especially in `GPT-5`, highlighting the ingrained *misrepresentation* of darker-skinned professionals in poorer settings. Prompt control (i.e., requesting stricter neutral settings) only slightly mitigate this, with the average change of around $80\% \rightarrow 86\%$ when going from uncontrolled to controlled.

The full MST distributions are shown as population-pyramid style histograms in Fig. 5. Across all models, generations are concentrated in lighter MST categories, with a clear rightward shift toward darker skin tones as prompts transition from high-income to low-income settings and from controlled to uncontrolled prompts. Table 3 summarizes the corresponding mean (μ), variance (σ^2), and median ($\tilde{m}$) MST values. For all models, both the mean and median increase monotonically from high-income to low-income prompts, indicating a systematic redistribution of probability mass across the MST scale rather than isolated changes in the distribution tails. Variance generally increases under low-income and uncontrolled conditions, reflecting a broader spread of generated skin tones alongside the darker central tendency. Prompt control consistently attenuates these shifts in central tendency and dispersion, but does not eliminate them. For example, for `Gemini`, the median MST

MST Set	Gemini			GPT-5			Grok		
	μ	σ^2	$\tilde{m}$	μ	σ^2	$\tilde{m}$	μ	σ^2	$\tilde{m}$
H_C	2.81	0.61	3	3.41	0.57	3	4.22	1.16	4
H_U	2.96	0.72	3	3.63	0.73	4	4.31	1.18	4
L_C	4.24	1.18	4	4.94	1.14	5	4.86	1.23	5
L_U	4.61	1.28	5	5.32	0.58	5	5.25	1.57	5
H	2.88	0.67	3	3.52	0.66	3	4.26	1.17	4
L	4.42	1.26	4	5.13	0.89	5	5.06	1.43	5
C	3.52	1.40	3	4.17	1.43	4	4.54	1.29	5
U	3.78	1.68	4	4.47	1.37	5	4.78	1.59	5

Table 3: Mean (μ), variance (σ^2), and median ($\tilde{m}$) of MST values across models and prompt conditions. Aggregated rows pool MST observations across the corresponding subsets. Lower MST values indicate lighter skin tones.

increases from 3 (H_C) to 4 (L_C) under controlled prompts, compared to a larger shift from 3 (H_U) to 5 (L_U) under uncontrolled prompts; similar patterns hold for GPT-5 and Grok.

We next quantify these associations between MST values and income- and constraint-level prompt variations using group-wise differences and association-based effect measures. For completeness, we additionally report an ordinal-aware distributional analysis using Earth Mover's Distance (EMD), which directly measures the magnitude of full-distribution shifts in MST across income and prompt conditions (Appendix A.1).

3.3 Prompt Effects

The prompt-conditioning effects on assigned skin tone is computed using the group-wise & correlational metrics defined in Sec. 2.3, and summarized in Table 4.

Income Effects Across all three T2I models, income conditioning induces a strong and systematic shift in MST scores. The mean difference is consistently negative, indicating that portraits generated from low-income prompts are assigned darker MST values than those from high-income prompts. The largest point estimate is observed for GPT-5 ($\Delta_{\text{income}} = -1.61$), closely followed by Gemini (-1.54), and substantially weaker for Grok (-0.79).

The standardized effect sizes reinforce this pattern. Cohen's d reaches large magnitudes for GPT-5 ($d = -1.82$) and Gemini ($d = -1.57$), indicating that the income-induced separation between MST distributions is large relative to within-group variability. In contrast, Grok exhibits a moderate effect ($d = -0.70$), suggesting comparatively weaker sensitivity to income cues.

Consistent with these findings, the point-biserial correlation r between income labels and MST scores is strongly negative for GPT-5 ($r = -0.67$) and Gemini ($r = -0.62$), and more moderate for Grok ($r = -0.33$). Bootstrap confidence intervals (CIs) for all three metrics exclude zero across models (bracketed in Table 4), indicating that the observed income–skin tone associations are stable under resampling.

Constraint Effects Prompt control exerts a comparatively smaller influence on MST outcomes. The prompt-constraint effect Δ_{prompt} is modest and negative across all models, with values between -0.24 and -0.30. This indicates that controlled prompts yield slightly lower MST scores than uncontrolled prompts, corresponding to marginally lighter skin-tone generations. Although bootstrap CIs for Δ_{prompt} exclude zero for all models, their widths are large relative to the point estimates, indicating that prompt-control effects are modest and substantially weaker than income-driven shifts.

3.4 Interaction Effects

To assess whether prompt control mitigates income-based disparities, we examine income effects separately under controlled and uncontrolled conditions and compute the interaction term Δ_{int} (Equations 7–8), reported in Table 4.

Across all three models, income effects are smaller in magnitude under controlled prompts than under uncontrolled prompts. For Gemini, the income effect shifts from $\Delta_U = -1.65$ under uncontrolled prompts to $\Delta_C = -1.43$ under controlled prompts; for GPT-5, from -1.69 to -1.52; and for Grok, from -0.95 to -0.64. These differences yield positive interaction point estimates of $\Delta_{\text{int}} = 0.22$ for Gemini, 0.16 for GPT-5, and 0.31 for Grok, suggesting a potential attenuation of income effects under prompt control.

However, the corresponding bootstrap confidence intervals for Δ_{int} include zero for all models, indicating that the magnitude of attenuation is uncertain and not statistically robust under resampling. Notably, the controlled-income effects Δ_C remain large and negative for Gemini (-1.43) and GPT-5 (-1.52), demonstrating that substantial income-associated shifts in perceived skin tone per-

Metric	Gemini	GPT-5	Grok
Δ_{income}	-1.54 [-1.67, -1.41]	-1.61 [-1.72, -1.48]	-0.79 [-0.95, -0.64]
d	-1.57 [-1.74, -1.40]	-1.82 [-2.04, -1.62]	-0.70 [-0.84, -0.55]
r	-0.62 [-0.66, -0.57]	-0.67 [-0.71, -0.63]	-0.33 [-0.39, -0.27]
Δ_{prompt}	-0.25 [-0.43, -0.08]	-0.30 [-0.46, -0.14]	-0.24 [-0.41, -0.08]
Δ_C	-1.43 [-1.61, -1.25]	-1.52 [-1.70, -1.34]	-0.64 [-0.85, -0.43]
Δ_U	-1.65 [-1.84, -1.46]	-1.69 [-1.84, -1.53]	-0.95 [-1.18, -0.72]
Δ_{int}	0.22 [-0.05, 0.48]	0.16 [-0.07, 0.40]	0.31 [-0.00, 0.61]

Table 4: Income, prompt-type, and interaction effects on MST scores with metrics and notations from Sec. 2.3. Point estimates are reported with 95% bootstrap confidence intervals over 10^4 resamples.

sist even under constrained prompting.

Overall, these results indicate that while prompt control may modestly reduce income–skin tone disparities, the dominant driver of bias remains the semantic content of income-related language itself.

4 Conclusion

Socioeconomic stereotypes and biases are deeply embedded in language, yet their manifestation in multimodal generation remains underexplored. Motivated by concerns about how such linguistic cues may shape visual representations, this work examines whether income-related language systematically influences skin-tone portrayals in text-to-image generation.

We conducted a controlled empirical study that benchmarks vision-language models on skin-tone annotation, generates occupational portraits across multiple text-to-image models & income-conditioned prompts, and audits the resulting images using the Monk Skin Tone scale. This experimental setup allows us to analyze how linguistic framing propagates into visual attributes across models and occupations.

Our results show a consistent association between socioeconomic language and perceived skin tone: higher-income prompts tend to produce lighter-skinned portraits, while lower-income prompts yield darker-skinned ones, even when occupation is held constant and prompt constraints are applied. These findings suggest that multimodal models encode and reproduce structured associations between socioeconomic meaning and visual appearance, mirroring the real-world effects of colorism.

More broadly, this work links social-science research on colorism with fairness studies in generative AI, showing that large vision-language models reproduce patterns of under- and misrepresentation of darker skin. These cross-modal stereotypes are structured, model-dependent, and largely robust to prompt controls, raising concerns for applications where generated images can influence perceptions of competence, trust, or identity. As multimodal systems are increasingly used in socially consequential settings, our findings underscore the need for human-centered, cross-modal audits that go beyond unimodal or purely technical metrics to better understand and mitigate language-mediated visual bias. Although we stress that reducing disparities in generated imagery does not necessarily imply alignment with real-world socioeconomic distributions, and that the normative goals of bias mitigation in generative models require careful consideration.

Limitations

The primary limitations of our study relate to model coverage, prompt design, statistical modeling choices, and the scope of linguistic and occupational variation considered.

Limited model coverage: Our evaluation includes three vision-language models for skin-tone benchmarking and three large text-to-image generators for image synthesis. While these models are representative of widely deployed commercial systems, this limited coverage constrains the generalizability of our findings across alternative architectures, training regimes, and open-source models. In particular, our focus on proprietary systems precludes direct comparisons with open-weight VLMs and T2I models.

Prompt design and semantic entanglement: Income conditioning is introduced using a small set of prompt templates that bundle multiple semantic elements, including income background, professional context, and ID-style framing. This design does not isolate the marginal contribution of individual linguistic cues. A more incremental prompt formulation, where we progressively introduce occupation, income, professional setting, and visual constraints, could enable finer-grained attribution of how specific prompt components contribute to observed visual disparities.

Discrete operationalization of skin tone: Although skin tone is treated as a fine-grained ordinal construct using the 10-category Monk Skin Tone scale, some analyses additionally employ coarser binarizations (e.g., light vs. dark) for comparability with prior work. This aggregation necessarily obscures variation within categories. While we mit-

igate this by reporting full MST distributions and ordinal-aware metrics, future work could explore alternative representations or continuous perceptual scales.

Dependence structure and paired prompts: Our analysis treats generated images as conditionally independent given income and prompt constraints, even though prompts are paired by occupation across conditions. As a result, income and prompt effects are computed by aggregating across images rather than explicitly modeling within-occupation contrasts. A more statistically faithful approach would involve paired analyses or hierarchical ordinal models with occupation as a random effect.

Benchmark–audit domain mismatch: Skin-tone annotation models are benchmarked on MST-E, which contains real human portraits with varied poses, lighting conditions, and image quality, whereas the bias audit is conducted on synthetically generated images. Differences in realism, texture, and illumination between real and generated faces may affect annotation behavior, and some observed effects may partially reflect generative artifacts rather than real-world visual bias. Evaluating annotator reliability directly on synthetic images would help clarify this distinction.

Language scope: All prompts in our study are formulated in English, implicitly encoding socioeconomic cues and social hierarchies specific to Western contexts. Linguistic framing in other languages may activate qualitatively different social associations that are not reducible to income alone. For example, prompts formulated in Hindi or other South Asian languages may encode caste-related distinctions, which intersect with but are not equivalent to socioeconomic status and skin tone. As a result, the income–appearance associations observed here should not be assumed to generalize across languages or cultural contexts. Extending this analysis to multilingual prompting remains an important direction for future work.

Ethics Statement

This work audits socioeconomic linguistic bias in multimodal text-to-image models by analyzing how income-related prompts influence perceived skin tone in generated portraits. The goal is diagnostic rather than normative: we do not define ideal demographic distributions, but instead identify statistical patterns that may reflect or amplify existing social stereotypes. Skin tone is measured using the Monk Skin Tone scale solely as a perceptual fairness metric and not as a biological or demographic classification.

Human annotation data were collected through a voluntary, anonymous survey with no personally identifiable information. All analyzed portraits are synthetically generated and do not represent real individuals. Results are reported only in aggregate to reduce risks of misuse or demographic profiling.

Generative AI systems were used both as subjects of evaluation and as automated annotators for skin tone assessment. Their use is explicitly disclosed, and their limitations are discussed throughout the paper. No generative model outputs are presented as factual representations of real individuals. Generative AI tools were not used to write the manuscript, design the experiments, or perform the statistical analyses reported in this work.

We acknowledge limitations related to model coverage, English-only prompts, and potential measurement bias from automated auditing. Our intention is to support transparency, responsible evaluation, and bias mitigation in generative AI systems, while discouraging applications that reinforce harmful stereotypes. The authors declare no known conflicts of interest.

Acknowledgments

We thank the valuable feedback of Dr Raj Dandekar and Dr Rajat Dandekar throughout the development of this work. We are grateful for their guidance on experimental design and system architecture, particularly in the design of the human-in-the-loop evaluation framework. We also acknowledge Vizuara AI Labs for providing computational resources and institutional support that made this research possible. We thank all anonymous annotators who participated in the human evaluation study, whose careful judgments enabled the calibration of our automated skin tone assessment methods. Finally, we appreciate the constructive comments from anonymous reviewers that helped strengthen this manuscript.

References

Maria Abascal and Denia Garcia. 2022. Pathways to Skin Color Stratification: The Role of Inherited (Dis)Advantage and Skin Color Discrimination in Labor Markets. *Sociological Science*, 9:346–373.

Ran Abramitzky, Jacob Conway, Roy Mill, and Luke Stein. 2023. The gendered impacts of perceived skin

tone: Evidence from african-american siblings in 1870–1940. Technical report, National Bureau of Economic Research.

Alexander Adames. 2023. The Cumulative Effects of Colorism: Race, Wealth, and Skin Tone. *Social Forces*, 102(2):539–560.

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

Vatsal Baherwani and Joseph James Vincent. 2024. Racial and gender stereotypes encoded into clip representations. In *The Second Tiny Papers Track at ICLR 2024*.

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. 2023. 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.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Mauricio Bucca. 2024. Colorism Revisited: The Effects of Skin Color on Educational and Labor Market Outcomes in the United States. *Sociological Science*, 11:517–552.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR.

A. Chardon, I. Cretois, and C. Hourseau. 1991. Skin colour typology and suntanning pathways. *International Journal of Cosmetic Science*, 13(4):191–208.

Marc Cheong, Ehsan Abedin, Marinus Ferreira, Ritsaart Reimann, Shalom Chalson, Pamela Robinson, Joanne Byrne, Leah Ruppanner, Mark Alfano, and Colin Klein. 2024. Investigating gender and racial biases in dall-e mini images. *ACM Journal on Responsible Computing*, 1(2):1–20.

Tavishi Choudhary. 2025. Political bias in large language models: A comparative analysis of chatgpt-4, perplexity, google gemini, and claude. *IEEE Access*, 13:11341–11379.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Jacob Cohen. 1992. A power primer. *Psychological Bulletin*, 112(1):155–159.

Philip R. Cohen, Marissa A. DiMarco, Rebecca L. Geller, and Leatrice A. Darrisaw. 2023. Colorimetric scale for skin of color: A practical classification scale for the clinical assessment, dermatology management, and forensic evaluation of individuals with skin of color. *Cureus*, 15(11):e48132.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. *Preprint*, arXiv:2304.05335.

Thomas B. Fitzpatrick. 1988. The validity and practicality of sun-reactive skin types i through vi. *Archives of Dermatology*, 124(6):869–871.

Kathleen C Fraser and Svetlana Kiritchenko. 2024. Examining gender and racial bias in large vision–language models using a novel dataset of parallel images. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 690–713.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Leander Girrbach, Stephan Alaniz, Yiran Huang, Trevor Darrell, and Zeynep Akata. 2024. Revealing and reducing gender biases in vision and language assistants (vlas). *arXiv:2410.19314*.

Leander Girrbach, Stephan Alaniz, Genevieve Smith, and Zeynep Akata. 2025. A large scale analysis of gender biases in text-to-image generative models. *arXiv:2503.23398*.

Vishal Gupta and Vinod Kumar Sharma. 2019. Skin typing: Fitzpatrick grading and others. *Clinics in Dermatology*, 37(5):430–436. The Color of Skin.

Joni Hersch. 2024. Colorism and immigrant earnings in the United States, 2015–2024. *Frontiers in Sociology*, 9:1494236.

John J. Howard, Yevgeniy B. Sirotin, Jerry L. Tipton, and Arun R. Vemury. 2021. Reliability and validity of image-based and self-reported skin phenotype metrics. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(4):550–560.

ILO. 2008. International Standard Classification of Occupations (ISCO): Concepts and Definitions. ILO Statistics Webpage. Accessed: 2025-01-10.

Yukun Jiang, Zheng Li, Xinyue Shen, Yugeng Liu, Michael Backes, and Yang Zhang. 2024. ModSCAN: Measuring stereotypical bias in large vision-language models from vision and language modalities. *arXiv:2410.06967*.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. 2025. Why language models hallucinate. *Preprint*, arXiv:2509.04664.

Aiswarya Konavoor, Raj Abhijit Dandekar, Rajat Dandekar, and Sreedath Panat. 2025. Vision-language models display a strong gender bias. *arXiv:2508.11262*.

L'Oréal. 2024. Inskin: Understanding skin science. L'Oréal Science & Technology Webpage.

Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. 2023. Stable bias: Evaluating societal representations in diffusion models. *Advances in Neural Information Processing Systems*, 36:56338–56351.

Douglas S. Massey and Jennifer A. Martin. 2003. The nis skin color scale. Office of Population Research, Princeton University.

Raj Gaurav Maurya, Vaibhav Shukla, Raj Abhijit Dandekar, Rajat Dandekar, and Sreedath Panat. 2025. Simulating misinformation propagation in social networks using large language models. *arXiv:2511.10384*.

Vishal Mirza, Rahul Kulkarni, and Aakanksha Jadhav. 2025. Evaluating gender, racial, and age biases in large language models: A comparative analysis of occupational and crime scenarios. In *2025 IEEE Conference on Artificial Intelligence (CAI)*, pages 244–251.

Ellis P. Monk. 2014. Skin tone stratification among black americans, 2001–2003. *Social Forces*, 92(4):1313–1337.

Ellis P. Monk. 2015. The cost of color: Skin color, discrimination, and health among african-americans. *American Journal of Sociology*, 121(2):396–444.

Ellis P. Monk. 2019. The color of punishment: African americans, skin tone, and the criminal justice system. *Ethnic and Racial Studies*, 42(10):1593–1612.

Ellis P. Monk. 2019. Monk Skin Tone Scale. Online Resource.

Ellis P. Monk. 2021a. Colorism and Physical Health: Evidence from a National Survey. *Journal of Health and Social Behavior*, 62(1):37–52.

Ellis P. Monk. 2021b. The unceasing significance of colorism: Skin tone stratification in the united states. *Daedalus*, 150(2):76–90.

Muhammad Osto, Iltefat H. Hamzavi, Henry W. Lim, and Indermeet Kohli. 2022. Individual typology angle and fitzpatrick skin phototypes are not equivalent in photodermatology. *Photochemistry and Photobiology*, 98(1):127–129.

Matthew A. Painter and Malcolm D. Holmes. 2023. Persistent skin tone and wealth stratification among new immigrants in the United States. *Research in Social Stratification and Mobility*, 83:100766.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. OpenAI.

Joseph L. Rodgers and W. Alan Nicewander. 1988. Thirteen ways to look at the correlation coefficient. *Psychological Bulletin*, 103(1):59–66.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Ashutosh Sathe, Prachi Jain, and Sunayana Sitaram. 2024. A unified framework and dataset for assessing societal bias in vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1208–1249, Miami, Florida, USA. Association for Computational Linguistics.

Candice Schumann, Femi Olanubi, Auriel Wright, Ellis Monk, Courtney Heldreth, and Susanna Ricco. 2023. Consensus and subjectivity of skin tone annotation for ml fairness. *Advances in Neural Information Processing Systems*, 36:30319–30348.

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

Kyra Wilson, Sourojit Ghosh, and Aylin Caliskan. 2025. Bias amplification in stable diffusion's representation of stigma through skin tones and their homogeneity. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(3):2705–2717.

L. Guillermo Woo-Mora. 2026. Unveiling the Cosmic Race: Skin tone and intergenerational economic disparities in Latin America and the Caribbean. *Journal of Development Economics*, 179:103594.

Xuyang Wu, Yuan Wang, Hsin-Tai Wu, Zhiqiang Tao, and Yi Fang. 2024. Evaluating fairness in large vision-language models across diverse demographic attributes and prompts. *arXiv:2406.17974*.

Rhiannon Yetsenga, Rhea Banerjee, Jared Streatfeild, Katherine McGregor, S. Bryn Austin, Belle W.X. Lim, Phillippa C. Diedrichs, Kayla Greaves, Josiemer Mattei, Rebecca M. Puhl, Jaime C. Slaughter-Acey, Iyiola Solanke, Kendrin R. Sonneville, Katrina Velasquez, and Simone Cheung. 2024. The economic and social costs of body dissatisfaction and appearance-based discrimination in the United States. *Eating Disorders*, 32(6):572–602.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2025. Siren's song in the ai ocean: A survey on hallucination in large language models. *Computational Linguistics*, pages 1–46.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. *arXiv:2303.18223*, 1(2).

A Appendix

A.1 Distributional Shift via Earth Mover's Distance

To complement the mean-, median-, and correlation-based analyses reported in the main text, we conduct an ordinal-aware comparison of MST distributions using Earth Mover's Distance (EMD). EMD quantifies the minimum amount of probability mass that must be transported along an ordered scale to transform one distribution into another, and is therefore well suited to the discrete, ordinal nature of the MST labels.

Given two empirical MST distributions P and Q over bins $k \in \{1, \ldots, 10\}$, we compute EMD as

$$\mathrm{EMD}(P, Q) = \sum_{k=1}^{10} \left| F_P(k) - F_Q(k) \right|, \quad (9)$$

where F_P and F_Q denote the corresponding cumulative distribution functions. EMD is expressed in units of MST bins and admits a direct interpretation as the average magnitude of distributional shift.

We report EMD values for all relevant pairwise comparisons across income levels (high vs. low), prompt types (controlled vs. uncontrolled), and their combinations, computed separately for each model using all valid MST observations.

Comparison	Gemini	GPT-5	Grok
H_C vs. H_U	0.1485	0.2173	0.1597
L_C vs. L_U	0.3846	0.3913	0.4456
H_C vs. L_C	1.4318	1.5229	0.6367
H_U vs. L_U	1.6486	1.6873	0.9474
H vs. L	1.5400	1.6053	0.7926
C vs. U	0.2643	0.3057	0.2584
H_C vs. L_U	1.7972	1.9046	1.0345
L_C vs. H_U	1.2832	1.3056	0.5591

Table 5: Earth Mover's Distance (EMD) between MST distributions across income and prompt conditions. EMD is measured in MST bins and quantifies the magnitude of full-distribution shifts along the ordinal skin tone scale.

A.2 Occupation-level Analysis

All reported results so far aggregated the MST outcomes across occupations, treating job titles primarily as a mechanism for generating diverse human depictions. This pooling may obscure substantial variation in income–skin tone associations across occupations, particularly where job titles carry strong implicit socioeconomic or cultural connotations. Thus, here we analyze some occupation-specific effects.

Light v. Dark Jobs Analysis of 210 occupations reveals systematic colorism in AI-generated images: professional occupations are rendered with significantly lighter skin tones (mean MST = 3.36) compared to manual labor and informal sector occupations (mean MST = 5.25), producing a 1.89-point bias gap (Figure 6). The lightest occupations—pharmacist (3.50), business analyst (3.50), data scientist (3.42)—represent professional and technical roles, while the darkest—auto rickshaw driver (5.58), herdsman (5.50), shepherd (5.33)—predominantly involve manual labor or informal sector work. Notably, Global South-specific occupations (boda boda rider, auto rickshaw driver) consistently rank among the darkest, suggesting that models conflate geographic origin, socioeconomic status, and skin tone. The asymmetric distribution relative to the MST midpoint (5.5)—lightest mean at 3.36 versus darkest at 5.25—indicates a systematic skew toward lighter representations in professional contexts, with darker tones emerging primarily when occupational cues invoke low-status or region-specific work.

Figure 7 disaggregates these patterns by model, revealing convergent directional bias but varying

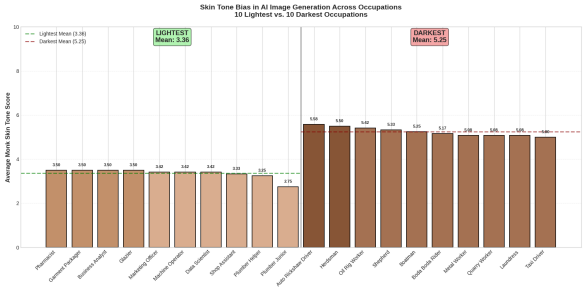

Figure 6: Aggregate occupational skin tone bias. Top 10 lightest (mean MST = 3.36) and darkest (mean MST = 5.25) occupations averaged across three models, showing a 1.89-point bias gap. Professional roles cluster lighter, manual labor darker. Bar colors represent Monk Skin Tone scale values.

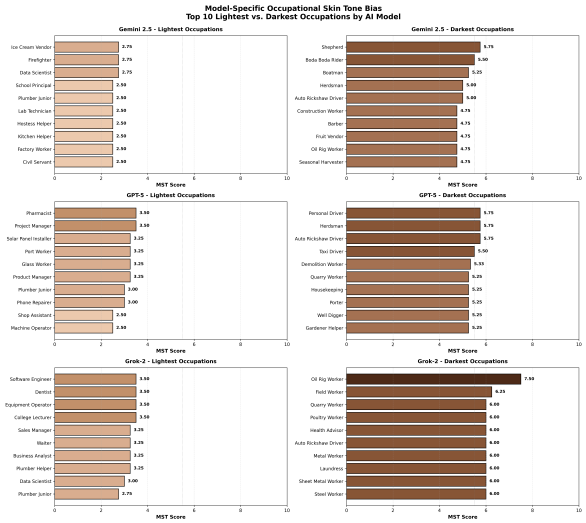

Figure 7: Model-specific bias comparison for occupationa sking tone bias. `Grok-2` exhibits the largest bias gap (2.90 points, range 2.75–7.50), `GPT-5` the smallest (2.33 points). Cross-model consistency in extreme occupations indicates shared training data stereotypes.

magnitudes. While all three models associate professional occupations with lighter tones and manual labor with darker tones, `Grok-2` demonstrates the largest bias gap (2.90 points, range 2.75–7.50), followed by Gemini 2.5 (2.45 points) and `GPT-5` (2.33 points). Cross-model consistency in specific occupations—plumber junior among the lightest for all models (MST 2.50–3.00), auto rickshaw driver among the darkest (MST 5.00–6.00)—provides strong evidence for shared training data biases rather than model-specific artifacts. `GPT-5`'s 24% smaller bias gap relative to `Grok-2` demonstrates that model design can modulate bias expression, yet the persistent directional bias across all models indicates that superficial debiasing techniques are insufficient. Effective mitigation requires systematic training data auditing, occupational representation balancing across skin tones, and counterfactual augmentation to break learned correlations between occupational status and skin tone.

Male-leaning	**Female-leaning**
Auto rickshaw driver	Beautician
Barber	Business analyst
Bellboy	Care assistant
Bicycle mechanic	Clinic manager
Boatman	Cook
Bricklayer	Flower vendor
Driver	Hairdresser
Electrician	Housekeeper
Electrician helper	Housekeeping
Farmer	Hostess helper
Herdsman	Human resources officer
Hotel porter	Kitchen manager assistant
Mason	Lawyer
Oil rig worker	Legal assistant
Phone repairer	Manicurist
Plumber	Medical technician
Plumber helper	Nanny
Plumber junior	Nurse
Pool cleaner	Office administrator
Quarry worker	Physiotherapist
Scaffolder	Receptionist
Software engineer	Researcher
Systems administrator	School principal
Taxi driver	Sewing machine operator
Truck driver	Social worker
Water tanker driver	Teacher
Well digger	Textile worker

Table 6: Occupations exhibiting exclusive gender presentation across all generated images. All listed occupations were rendered as either male-presenting (left) or female-presenting (right) across all models and prompt conditions.

Gender Roles When we assign binary gender to the generated images, we clearly notice the dominance of female portraits for stereotypical female professions—such as nanny, nurse, housekeeping, manicurist, office administrator, teacher, beautician, while most of the professions have male-dominance.

We list occupations with the highest percentage of male portraits generated across all 3 T2I models and 4 prompt conditions, along with the ones with female-dominance in Table 6). The former is an abridged list as most of the jobs have a predominant male population, while we the latter shows the limited number of jobs where the percentage of females to male is from 83.3% to a 100%.

This gender exclusivity compounds the colorism bias documented in the main paper: individuals are being stereotyped along *both* skin tone and gender dimensions simultaneously, creating intersectional misrepresentation that affects darker-skinned women and lighter-skinned men differently depending on occupational context.

Child Labor We observe an additional and concerning pattern in a subset of generated images: the depiction of underage individuals in professional roles. This phenomenon occurs most frequently in outputs from `Grok 2`, suggesting comparatively weaker age-related guardrails than those observed in other models. The effect appears to be associated with occupational titles containing terms such as *"helper"*, indicating that certain linguistic cues may trigger age-related stereotypes alongside socioeconomic ones. While we do not quantify this effect systematically, its recurrence warrants further investigation.