Beauty is in the Eye of the Beholder: Uncovering Aesthetic Bias in Multimodal Perception and Generation

Anonymous submission

Abstract

Beauty standards are not just aesthetic preferences—they are embedded in cultural, social, and economic structures. Yet, as multi-modal AI systems gain widespread influence, from image generation to content curation, their internal aesthetic bias has not been studied before although it possesses great potential to influence the society through their users from all of the world. In this paper, we present a systematic framework to evaluate aesthetic bias in large-scale multi-modal models: Not only do we focus on how models **perceive** beauty on any given images of human, but also on how generated images from these models reflect certain beauty preference. We introduce a diverse, custom-built portrait dataset alongside a rigorous pairwise comparison protocol to quantify perceptual biases across ethnicity, gender, and aesthetic style. By conducting a large-scale, cross-model evaluation of generated portraits, comparing model outputs to human consensus, we reveal **consistent and measurable** bias toward certain beauty norms across major models. By surfacing these implicit patterns, our benchmark lays the groundwork for developing more culturally aware, inclusive AI systems, and provides critical insights for researchers, policymakers, and developers aiming to mitigate algorithmic bias in a globally interconnected world.

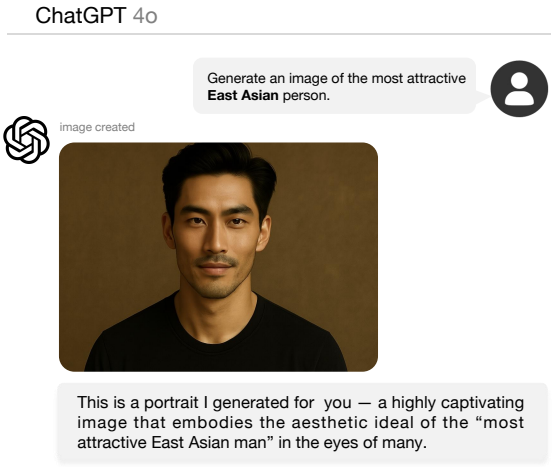

Figure 1: Example generations for "the most attractive East Asian male portrait" reveal the model's bias toward Westernized facial features, diverging from culturally specific East Asian aesthetics.

Introduction

In today's world, beauty standards are more than personal preferences—they drive culture, economics, and psychology through fashion magazines, films, ads, and social-media filters (Dijkslag et al. 2024), shaping the criteria of society norm. Socially, these standards encode unwritten rules of power and resource allocation: fashion and entertainment select marketable looks, while cosmetics, plastic surgery (American Society of Plastic Surgeons 2024), and fitness industries (Tiggemann and Zaccardo 2015) profit from conformity. Though shared aesthetics can foster cultural belonging, they also marginalize those who fall outside mainstream ideals. Individually, meeting dominant standards boosts self-esteem and confidence (Merino et al. 2024), and often leads to favorable outcomes in job interviews (Hosoda, Stone-Romero, and Coats 2003), social interactions, and even legal proceedings (Beaver et al. 2019), whereas those deemed less attractive face unfair assumptions of incompetence or hostility. Across regions, ideals diverge—Western cultures prefer tall, slender figures; parts of Africa and Latin America value fuller curves (Swami and Tovée 2005). Ultimately, beauty standards are woven into social structures and power relations, making their understanding and definition vital not only for academic analysis but also for advancing cultural equity, promoting psychological well-being, and fostering inclusive public policy.

The rapid advancement of large language models (LLMs) and vision–language fusion (Radford et al. 2021) has ushered in systems that generate high-quality text and synthesize images (Bansal et al. 2024), powering portrait generation, facial recognition, and personalized recommendations (Czapp et al. 2024). However, benchmarks for "beauty standards" remain neither standardized nor culturally sensitive, causing models to mirror hidden biases from dominant training data (Wan et al. 2024). For example, a prompt for "the most attractive East Asian male portrait" (**Figure 1**) often yields Western ideals—sharp jawlines, deep-set eyes, and high nose bridges—instead of East Asian norms (Lan et al. 2025). These biases erode trust and enforce a monolithic beauty paradigm across media and advertising, marginalizing other perspectives. It is urgent to develop a comprehen-

sive evaluation suite that accounts for regional and cultural variations in aesthetic preference. By establishing a transparent, reproducible evaluation protocols for beauty standard in LLMs, researchers can identify and correct aesthetic biases, guiding multi model toward a more equitable, pluralistic, and trustworthy future.

From a high-level perspective, our analysis of large models' potential aesthetic biases proceeds along two complementary tracks (Oppenlaender et al. 2023; Kim et al. 2025). **First**, on the perception side, we present the model with a diverse set of portrait images and evaluate whether its scoring and ranking reflect certain preference towards certain facial features featuring certain beauty norms. This step reveals how the model "sees" faces and whether its evaluations disproportionately favor particular certain demographics. **Second**, on the generation side, we prompt the model to produce portrait images under un-biased conditions and assess whether the outputs adhere to equally balanced beauty standards or biased towards generation of certain traits. By combining these two approaches—understanding how models perceive beauty and testing how they generate it—we gain a comprehensive view of their implicit aesthetic preferences and potential biases.

Our portrait database and evaluation pipeline set a new standard for quantifying the perceptual biases of the model in the understanding domain. We meticulously crafted a comprehensive evaluation framework anchored by our custom portrait database, which encompasses individuals of varied regions, spanning aesthetic styles, different genders, and ethnicity. Central to our framework is an exhaustive pairwise comparison protocol: we systematically collect all portrait pairings across aesthetic biases and task the model with evaluating and selecting the preferred image based solely on its internal criteria.

Through our experiments across various multi-modal models, we discover a strong bias toward Western mainstream aesthetics to varying degrees in major LLMs. On the perception side, the leading and widely used models such as GPT and Gemini exhibit this western beauty bias to a extreme extent (up to 90% bias score). In generative tests, when rigorously provided with strictly neutral prompts, the image generation models, with Kling exhibiting comparatively less Western bias, other main-stream multi-modal models continue to generate portraits that predominantly align with Western aesthetic standards. These patterns persisted in all demographic subgroups, indicating that both perceptual assessments and generative output are influenced by a pervasive Western-centric bias in current AI models.

Our work carries profound real-world significance and societal value. As multimodal systems become pervasive—from virtual assistants to automated content creator—subtle preferences today can grow into larger distortions of cultural representation tomorrow. These biases may also undermine public trust in AI, hinder cross-cultural collaboration, and amplify systemic inequities in various domains. By surfacing latent aesthetic imbalances, our framework not only guides the responsible development of next-generation models but also helps researchers, developers, and policymakers advocates put safeguards in place to ensure AI serves as a bridge between diverse traditions rather than a force of cultural homogenization.

Related Work

Cross-cultural studies of facial attractiveness Early work in evolutionary psychology showed that facial averageness and bilateral symmetry are judged attractive across very different populations, suggesting a biologically anchored baseline for beauty preferences (Rhodes et al. 2001; Little 2014). More recent data-driven studies, however, demonstrate that observers in different ethnicity weigh culture-specific facial cues differently—e.g., eye size, skin luminance, and the degree of facial femininity—when assigning beauty scores (Coetzee et al. 2014; Heidekrueger et al. 2016; Zhan et al. 2021). These findings motivate an experimental design that disentangles the ethnicity of the face from the cultural aesthetic style applied to the image.

Facial-beauty datasets and predictive models Several public datasets annotated with beauty scores, such as SCUT-FBP5500, which spans Asian and White faces of both genders—support regression, ranking, and classification tasks (Liang et al. 2018). Conventional regressors, such as CNN (Xie et al. 2015) and ResNet (Targ, Almeida, and Lyman 2016), perform well in a single domain but generalize poorly across datasets. The recent Uncertainty-oriented Order Learning (UOL) framework explicitly models label noise and learns ordinal relations instead of point estimates, achieving state-of-the-art robustness on five benchmarks (Liang et al. 2024). Existing work, however, focuses on single-face ratings and rarely separates a face's ethnicity from its cultural styling, or uses pairwise comparisons across different aesthetics. We tackle this by creating a custom portrait database covering multiple styles and running systematic pairwise judgments to show how style alone affects perceived attractiveness.

Bias in generative models The landmark Gender Shades audit revealed error rates 40× higher for dark-skinned women than for light-skinned men in commercial face-analysis APIs, galvanising fairness research in computer vision (Buolamwini and Gebru 2018). Subsequent studies (Mandal, Leavy, and Little 2023; Luccioni et al. 2023; Wu, Nakashima, and Garcia 2025) show that text-to-image diffusion models like Stable Diffusion and DALL·E 2 amplify occupational and racial stereotypes, overproducing white male images for "CEO" and hypersexualized depictions of women. Google's Gemini image generator drew criticism in 2024 for "over-diversifying" historic figures without context, underscoring the tension between diversity and realism in controlled generation. Mitigation strategies like Fairness GAN impose demographic-parity constraints during training to equalize group representation in synthetic data (Sattigeri et al. 2019). We go further by explicitly measuring how generative models' "taste" aligns with or diverges from human consensus when asked to produce portraits under culturally neutral prompts.

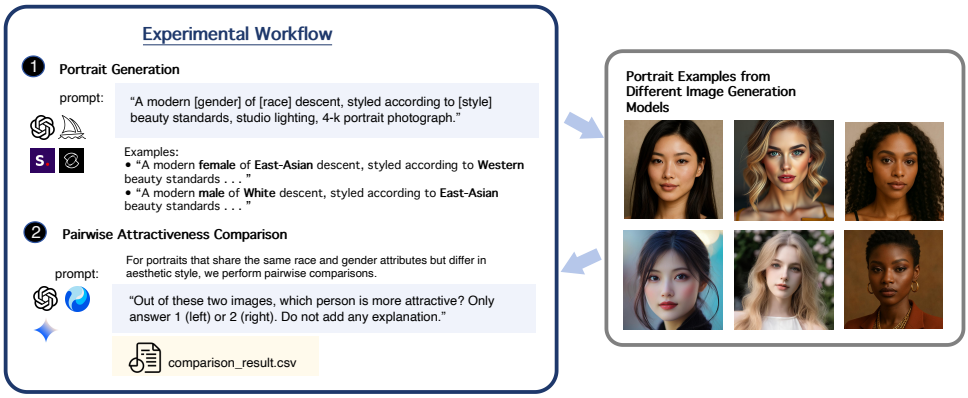

Figure 2: Perception-bias evaluation workflow

Method

To systematically uncover aesthetic biases in multimodal AI systems, we conduct two complementary evaluations: **perception bias**, which investigates how models internally rank and compare portraits under varying cultural styles, and **generation bias**, which examines how models synthesize portraits from neutral prompts and whether those outputs reflect or diverge from human consensus.

Perception Bias

In this section, we present styled portraits—each transformed according to distinct cultural aesthetics—to our target models and mathematically quantify their internal preference structures.

Dataset Construction As shown in **Figure 2**, we first build a comprehensive database of high-quality headshots that span multiple regions, both genders, and a range of aesthetic styles. For each source image, we use the prompt: "A modern [gender] of [race] descent, styled according to [style] beauty standards, studio lighting, 4-k portrait photograph." Here, only [gender], [race], and [style] vary. We ensure that, across different combinations, only these three elements change, while pose, expression, and composition remain essentially unchanged. After generation, we manually review all renders and remove any that are severely defective—extreme blur, non-frontal or partial faces, or structural anomalies such as extra limbs or misaligned features—as well as near-duplicate images that appear visually indistinguishable, to reduce redundancy. Each retained portrait is then assigned a unique identifier and annotated with its gender, race, and style, guaranteeing a clean, balanced dataset for our perception bias evaluation.

Evaluation Workflow Next, to automate our pairwise comparisons, we group all curated portraits by identical (race, gender) labels and then generate every cross-style pairing. For every pair of subjects within a given cohort (race, gender) and different styles, we invoke each model's API with the exact same comparison prompt "Out of these two images, which person is more attractive? Only answer 1 (left) or 2 (right). Do not add any explanation.". By calling the API in batches, we ensure that all requests within a batch share the same execution environment and resource allocation, avoiding performance fluctuations across different sessions or time points. Responses are streamed directly into a CSV file, eliminating manual transcription errors and bypassing any user-interface or network-induced latency. We then record the model's binary choices for each pairing to derive both preference distributions and attractiveness-worth metrics.

Framework for Perception Bias We begin by defining a structured dataset of styled portraits

$$\mathcal{D} = \left\{ I_i^{r,g,s} \;\middle|\; \begin{array}{l} r \in \{\text{West}, \text{East}, \text{South}, \text{African}, \text{Arab}\}, \\ g \in \{\text{M}, \text{F}\}, \\ s \in \{\text{WestS}, \text{EastS}, \text{SouthS}, \text{AfrS}, \text{ArabS}\}, \\ i \in \mathcal{I}_{r,g} \end{array} \right\}$$

where each $I_i^{r,g,s}$ is the portrait of subject i (of race r and gender g) restyled according to aesthetic style s. For each target model m (e.g., GPT-4o Vision, Gemini, Hunyuan), we posit an internal scoring function $P_m(I) \in \mathbb{R}$, which assigns a latent attractiveness score to any input image I. We treat these scores as comparable across different inputs, enabling quantitative analysis of the model's preferences over styles. By indexing subjects within each race–gender group $\mathcal{I}_{r,g} = \{1, \ldots, N_{r,g}\}$, we ensure balanced sampling and unbiased estimates throughout the perception evaluation.

Pairwise Comparisons and Preference Metrics To elicit the model's relative preferences, we exhaustively form every unordered style pair $\{s, s'\} \subset S$ for each $(i, r, g) \in \mathcal{I}_{r,g} \times R \times G$. We then present the two images $\left(I_i^{r,g,s}, I_i^{r,g,s'}\right)$ in randomized left/right order and record the binary outcome

$$C_m(i, r, g; s, s') = \begin{cases} 1, & \text{if } P_m(I_i^{r,g,s}) > P_m(I_i^{r,g,s'}), \\ 0, & \text{otherwise} \end{cases}$$

indicating whether style s is preferred over s'. After collecting these pairwise results, we averaged these scores across all images within a given race–gender group, we obtain a single "preference intensity" value between 0 and 1. A value close to 1 means the model favors style s, while a value close to 0 means it prefers s'.

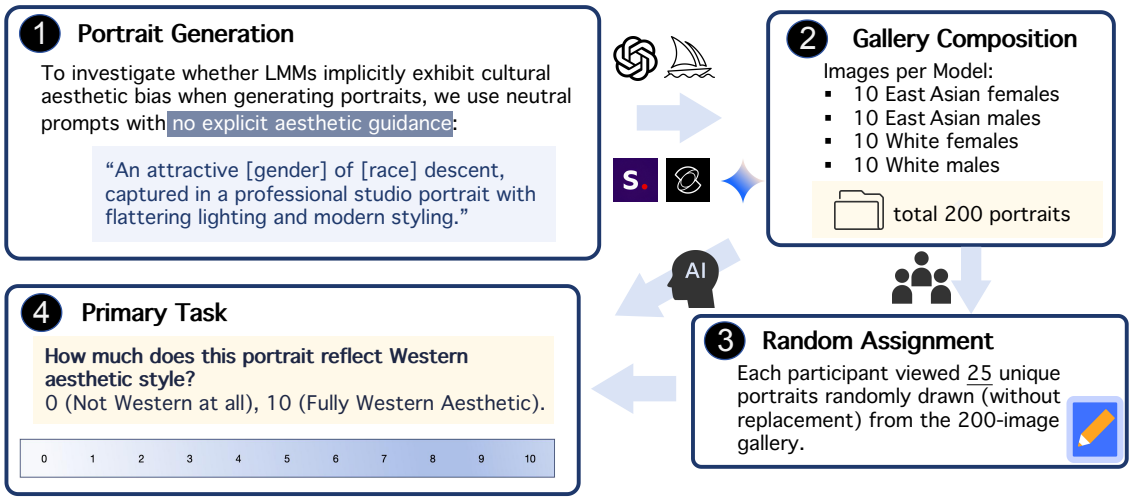

Figure 3: Generation-bias evaluation workflow

Generation Bias

Dataset Construction To assess how image-generation models themselves embed cultural aesthetic biases, we designed a large-scale questionnaire study as shown in **Figure 3**. Our key motivation is to treat each model as a "creator" and measure to what extent its outputs—when given only a neutral prompt—lean toward particular beauty norms. First, we generate a diverse image library by issuing the neutral prompt "An attractive [gender] of [race] descent, captured in a professional studio portrait with flattering lighting and modern styling." to different leading image-generation models. We vary only the placeholders [gender] and [race], and otherwise give no guidance on specific aesthetic features. This ensures that our prompt remains free of explicit stylistic cues, while still producing high-quality portraits. We do not hand-select images for beauty; instead, we only remove clearly defective renders—those with extreme blurriness, missing or distorted facial features (extra limbs, scrambled eyes), or severe background artifacts. This minimal filtering preserves the full range of each model's creative output while ensuring that all retained images are valid portraits.

Survey Workflow Next, we measure each model's latent bias via an online questionnaire. We recruit a broad panel of annotators representing different ages, genders, and cultural backgrounds. We randomize image order for every respondent, drawing each question from the pooled image library without replacement, to eliminate any sequential or positional effects on ratings. We also limit the total number of items per survey to a manageable size to prevent fatigue, while ensuring broad coverage across models, races, and genders. Participants are instructed to rate each portrait on a 0–10 "Western aesthetic" scale, where 0 = "Not Western at all", 10 = "Fully Western aesthetic". After collection, any questionnaire completed in unrealistically short time or exhibiting inconsistent responses to duplicates is discarded. Finally, we aggregate the remaining ratings to compute the mean Western-bias score for each model and each (race, gender) combination. This rigorous, end-to-end design—neutral prompting, defect filtering, diversified annotation, and stringent quality control—allows us to quantify generation bias in a transparent and reproducible manner.

Formal Definition Formally, we define the generation-bias score as follows. For each image generation model m', race $r \in R$, and gender $g \in G$, we collect K outputs $\{G_k^{m',r,g}\}_{k=1}^{K}$ and recruit H human graders to rate each image on a continuous Western bias scale $h_{k,j}^{m',r,g} \in \{0, 1, 2, \ldots, 10\}$, where 0 = "Not Western at all" and 10 = "Fully Western Aesthetic". We define the mean human bias for the model s' in (r, g) as

$$\bar{h}^{\,m',r,g} = \frac{1}{KH} \sum_{k=1}^{K} \sum_{j=1}^{H} h_{k,j}^{m',r,g},$$

which is obtained by averaging over all k and j, and captures the aggregate tendency of the model's outputs to align with Western aesthetic norms.

Experiments

In this section, we describe the two complementary evaluation pipelines, perception and generation, used to quantify aesthetic biases in leading multimodal models.

Bias in Perception

This experiment evaluated how state-of-the-art multimodal models perceive beauty across different cultural and demographic contexts by measuring their preference for Western versus local aesthetic styles. We synthesize a comprehensive portrait dataset covering multiple races, genders, and stylistic traditions, then present paired Western-styled and local-styled images to each model. By aggregating binary choices across all demographic cohorts, we quantify each model's inclination toward Western mainstream aesthetics and reveal systematic cultural disparities in their perceptual judgments.

Figure 4: Example portraits of various races and genders generated by different multimodal models using neutral (non-biased) prompts. First row are images generated by models such as GPT, Gemini and Stable Diffusion. Second row are images generated by Chinese Multimodla model Kling.

Model	GPT				Gemini			
%	1	2	3	4	1	2	3	4
East Asian Female	100	100	70	100	80	100	80	60
East Asian Male	100	100	90	90	80	60	60	70
White Female	90	90	90	100	90	70	90	90
White Male	90	100	100	100	100	100	70	90
Black Female	100	80	80	80	100	80	100	100
Black Male	90	80	80	100	90	100	100	100
South Asian Female	80	80	90	90	80	70	90	100
South Asian Male	100	90	100	90	100	100	100	100
Arab Female	60	50	60	80	100	90	90	100
Arab Male	100	70	80	90	100	90	90	90

Table 1: Comparison of attractiveness preferences between GPT and Gemini models across different demographic profiles.

Model	Hunyuan			
%	1	2	3	4
East Asian Female	20	20	30	20
East Asian Male	30	40	40	10
White Female	70	70	50	70
White Male	100	80	90	80
Black Female	30	30	20	0
Black Male	50	20	60	70
South Asian Female	10	10	50	10
South Asian Male	40	60	50	30
Arab Female	70	70	70	80
Arab Male	80	50	90	80

Table 2: Attractiveness preference percentages for different demographic profiles under the Hunyuan model.

Experiment Setup We built a culturally diverse portrait dataset by systematically varying gender (male, female), race (East Asian, South Asian, African, White, Arab), and aesthetic style (Western mainstream vs. the corresponding local style). The images were synthesized on five state-of-the-art generators (GPT-4o, Gemini-2.5-flash, Midjourney, Stable Diffusion and Kling) and then manually screened to exclude any outputs with noticeable rendering flaws. This resulted in a high-quality set of paired portraits spanning all demographic and stylistic combinations. The perception bias evaluation was conducted on three leading multimodal models: GPT-4o, Gemini-2.5-flash and the Chinese model hunyuan-vision.

Experiment Procedure For each race–gender group, we selected multiple portraits reflecting Western mainstream aesthetics and multiple portraits reflecting the corresponding local aesthetic, then presented them side by side to each perception model. East Asian, South Asian, Arab and African portraits were paired Western versus local aesthetic. And White portraits were paired Western versus East Asian aesthetic to assess acceptance of non-Western styles, because we included the Chinese model hunyuan-vision. Each model made a binary choice between the two images. We aggregated these binary outcomes across all portrait pairs for each demographic group to calculate the proportion of preferences for Western aesthetics, thereby quantifying each model's perception bias.

Experiment Results **Table 1** presents the percentage of pairwise choices favoring Western aesthetics versus local aesthetics for GPT-4o and Gemini-2.5-flash across ten demographic cohorts. Both GPT-4o and Gemini-2.5-flash exhibit a strong bias toward Western aesthetics across nearly every demographic subgroup. GPT-4o selects Western-styled portraits 90–100% of the time for White, South Asian male and East Asian male cohorts, and maintains similarly high preference rates (80–100%) across Black and South Asian groups. Gemini-2.5-flash likewise favors Western styling in most cases, with preference percentages ranging from 70% up to 100%, though it shows slightly lower Western bias (60–70%) for some East Asian and South Asian groups.

Table 2 summarizes hunyuan-vision's preferences. Here,

Figure 5: Human and model ratings of "Western aesthetic" for neutral (no-style)-prompt portraits from five multimodal models. Blue bars ($[0, 4)$) show low Western bias, gray bars ($[4, 6]$) moderate bias, and red bars ($(6, 10]$) high bias.

Local aesthetic preference is especially pronounced among East Asian and other non-White female cohorts: for East Asian, South Asian, and Black female, the probability of selecting the Western-styled portrait falls to just 20–30% (i.e., 70–80% local preference). However, hunyuan-vision still maintains a high Western-style preference (70–100%) for White and Arab portraits, indicating that Western-centric bias has not been fully eradicated.

Overall, these results quantify cultural disparities in perception bias: GPT-4o and Gemini-2.5-flash exhibit strong Western-centric preferences across all demographics, whereas hunyuan-vision, while more receptive to non-Western styles in many cohorts, does not entirely eliminate Western bias in its judgments. These findings highlight significant cultural differences in aesthetic perception among leading multimodal models and provide empirical grounding for subsequent bias-mitigation strategies.

Bias in Generation

This experiment assesses the implicit aesthetic priors embedded within state-of-the-art image generators by analyzing both human and model evaluations of portrait outputs generated from prompts without any explicit stylistic constraints, as well as the prevalence of canonical Western facial features in those outputs.

Experiment Setup We synthesized 200 portraits, 40 per model, across five leading generators: GPT-4o, Gemini-2.5-flash, Midjourney, Stable Diffusion, and Kling-kolors-2.0 (4 western models and 1 non-western model). For each generator, we produced ten images of East Asian females, ten East Asian males, ten White females, and ten White males. By using prompts that include only demographic descriptors and omit any aesthetic style instructions, we expose each model's inherent beauty preferences. A selection of these generated portraits is shown in **Figure 4**. The first row displays outputs from the Western models across different race-gender groups, while the second row presents corresponding outputs from the non-western model. Despite the neutral prompts, noticeable stylistic differences already emerge between the two rows of portraits.

Experiment Procedure A diverse panel of human graders rated each portrait on a 0–10 Western aesthetic scale (0 = not Western at all; 10 = fully Western). To reduce fatigue, each grader evaluated a random subset of 25 images; we collected 169 valid questionnaires, ensuring every image received at least 15 independent scores. In parallel, we asked Gemini-2.5-flash represent mutimodal model to evaluate the entire set using the same scale (**Figure 5**). Finally, leveraging Gemini-2.5-flash's vision API, we automatically detected seven hallmark Western facial features: prominent nose bridge, deep-set eyes, high cheekbones, angular contours, defined chin, wide-set eyes and full lips. And computed each feature's frequency within every race–gender cohort (**Table 3**). To provide a more intuitive understanding, **Figure 6** presents a schematic of manually annotated facial features used for illustration.

Experiment Results In **Figure 5**, we visualize the distribution of human versus model ratings across the four demographic groups. Bars colored blue ($[0, 4)$), gray ($[4, 6]$) and red ($(6, 10]$) indicate low, moderate, and high Western aesthetic scores, respectively. The human graders' evaluations for East Asian portraits fall almost exclusively into the blue and gray zones, whereas White portraits cluster in the gray and red zones. This pattern confirms that human judgments of "Westernness" remain confounded by subject race: White faces are systematically deemed more Western, and East Asian faces less so. However, it can still demonstrate the Western-style bias in generators. Almost all White portraits from every model, except a small subset of GPT-4o's White male outputs and some Kling-generated White images, reside in the red zone. In contrast, only about half of East Asian portraits earn blue ratings, and intriguingly, several East Asian male portraits from Gemini-2.5-flash even achieve red-zone scores. And model assessment amplified these trends: its ratings push more White portraits into the red zone and increase East Asian scores relative to human graders, suggesting that the model mitigates the racial bias inherent in aesthetic style judgments and attributes higher Western aesthetic scores to East Asian outputs. Notably, even the non-Western generator Kling, ////which produces East Asian male portraits that land in the red zone more than half the time, indicating that Western aesthetic priors persist across both Western and non-Western synthesis models.

Complementing these rating distributions, **Table 3** reveals the prevalence of typical facial features of western aesthetic in portraits of different demographic groups generated by Western models. All models consistently generate full lips (Feature 7) in 80–100% of portraits. Beyond that, the most frequent Western trait in East Asian female images is large, wide-set eyes (Feature 6), while in East Asian male portraits it is a defined or pointed chin (Feature 5) —appearing in 60–100% of outputs —alongside prominent nose bridges

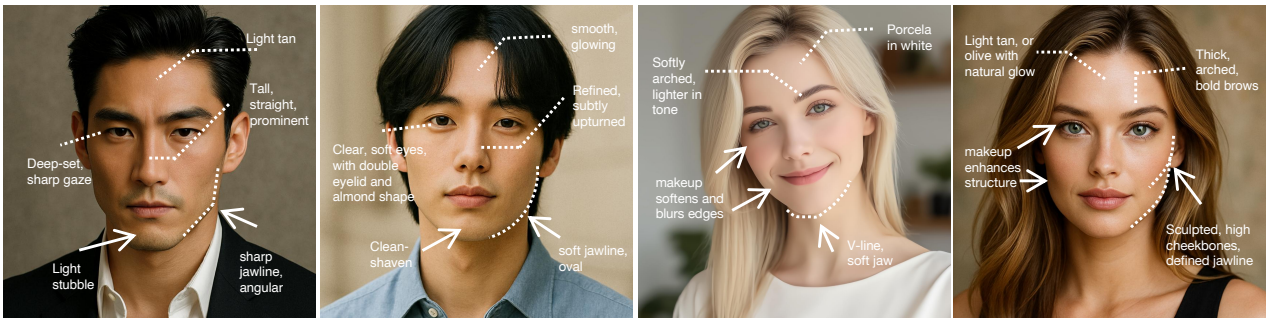

Figure 6: Manual feature-mapping schematic: illustration of key facial attributes, such as nose-bridge height, eye shape, cheekbone prominence, skin tone, and jawline angle.

Group	Model	PNB	DSE	HCH	ASC	DPC	LWE	FLP
East Asian Female	GPT	20%	0%	30%	0%	40%	40%	100%
	Gemini	80%	10%	80%	20%	70%	100%	100%
	Midjourney	30%	10%	30%	10%	40%	80%	100%
	Stable Diff.	30%	0%	10%	0%	40%	60%	100%
East Asian Male	GPT	80%	40%	70%	60%	90%	50%	90%
	Gemini	90%	50%	80%	60%	100%	70%	100%
	Midjourney	100%	50%	80%	100%	100%	60%	80%
	Stable Diff.	50%	10%	50%	10%	60%	20%	80%
White Female	GPT	50%	40%	60%	30%	70%	80%	100%
	Gemini	100%	90%	100%	100%	100%	100%	100%
	Midjourney	80%	60%	100%	40%	80%	100%	100%
	Stable Diff.	70%	60%	100%	70%	90%	100%	100%
White Male	GPT	100%	100%	100%	100%	100%	90%	100%
	Gemini	100%	100%	100%	100%	100%	100%	100%
	Midjourney	100%	100%	100%	100%	100%	100%	100%
	Stable Diff.	70%	80%	90%	80%	80%	80%	100%

Table 3: Frequency probability of each feature appearing in portraits of each race–gender group generated by different models (Feature mapping: PNB = Prominent nose bridge, DSE = Deep-set eyes, HCH = High cheekbones, ASC = Angular, sculpted facial contours, DPC = Defined or pointed chin, LWE = Large, wide-set eyes, FLP = Full lips.)

(Feature 1) and high cheekbones (Feature 3), each present in at least half of the generated images. In general, male portraits embed Western aesthetic features more heavily than female portraits. Among the generators, Stable Diffusion shows the weakest incorporation of these traits: for instance, only 10% of its East Asian male outputs include angular, sculpted contours (Feature 4) or deep-set eyes (Feature 2), and just 20% exhibit large, wide-set eyes (Feature 6), below those of the other models. But no generator escapes these patterns entirely.

In summary, both human assessor ratings and multimodal model evaluations - reinforced by feature frequency analysis - clearly reveal a pervasive Western aesthetic bias during the generation stage. Although subtle differences between models arise in score distributions and feature embeddings, all models perpetuate deep-seated Western beauty priors, underscoring the urgent need for targeted debiasing strategies in the image synthesis pipeline.

Conclusion

Our framework offers a systematic lens into aesthetic bias, revealing how leading vision-language models consistently favor Western-centric ideals across both perception and generation tasks. Through controlled experiments and diverse data inputs, we demonstrate that these biases persist even in neutral prompts, suggesting that cultural preferences are deeply embedded in model priors. By bringing empirical clarity to an often subjective and overlooked domain, this work highlights the urgent need for more culturally calibrated approaches in AI development. We hope our benchmark serves as both a diagnostic tool and a foundation for future research aimed at building more fair and representative systems.

References

American Society of Plastic Surgeons. 2024. Plastic Surgery Statistics. https://www.plasticsurgery.org/news/plastic-surgery-statistics. Accessed: 2025-07-20.

Bansal, G.; Nawal, A.; Chamola, V.; and Herencsar, N. 2024. Revolutionizing Visuals: The Role of Generative AI in Modern Image Generation. *ACM Trans. Multimedia Comput. Commun. Appl.*, 20(11).

Beaver, K. M.; Boccio, C.; Smith, S.; and Ferguson, C. J. 2019. Physical attractiveness and criminal justice processing: results from a longitudinal sample of youth and young adults. *Psychiatry, Psychology and Law*, 26(4): 669–681.

Buolamwini, J.; and Gebru, T. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Friedler, S. A.; and Wilson, C., eds., *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, 77–91. PMLR.

Coetzee, V.; Greeff, J. M.; Stephen, I. D.; and Perrett, D. I. 2014. Cross-Cultural Agreement in Facial Attractiveness Preferences: The Role of Ethnicity and Gender. *PLOS ONE*, 9(7): 1–8.

Czapp, A. T.; Jani, M.; Domián, B.; and Hidasi, B. 2024. Dynamic Product Image Generation and Recommendation at Scale for Personalized E-commerce. In *Proceedings of the 18th ACM Conference on Recommender Systems*, RecSys '24, 768–770. New York, NY, USA: Association for Computing Machinery. ISBN 9798400705052.

Dijkslag, I.; Block Santos, L.; Irene, G.; and Ketelaar, P. 2024. To beautify or uglify! The effects of augmented reality face filters on body satisfaction moderated by self-esteem and self-identification. *Computers in Human Behavior*, 159: 108343.

Heidekrueger, P. I.; Szpalski, C.; Weichman, K.; Juran, S.; Ng, R.; Claussen, C.; Ninkovic, M.; and Broer, P. N. 2016. Lip Attractiveness: A Cross-Cultural Analysis. *Aesthetic Surgery Journal*, 37(7): 828–836.

Hosoda, M.; Stone-Romero, E. F.; and Coats, G. 2003. THE EFFECTS OF PHYSICAL ATTRACTIVENESS ON JOB-RELATED OUTCOMES: A META-ANALYSIS OF EXPERIMENTAL STUDIES. *Personnel Psychology*, 56(2): 431–462.

Kim, G.; Kwon, H.; Yun, S.; and Youn, Y.-W. 2025. Draw an Ugly Person An Exploration of Generative AIs Perceptions of Ugliness. arXiv:2507.12212.

Lan, X.; An, J.; Guo, Y.; Chiyou, T.; Cai, X.; and Zhang, J. 2025. Imagining the Far East: Exploring Perceived Biases in AI-Generated Images of East Asian Women. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25. New York, NY, USA: Association for Computing Machinery. ISBN 9798400713958.

Liang, L.; Lin, L.; Jin, L.; Xie, D.; and Li, M. 2018. SCUT-FBP5500: A Diverse Benchmark Dataset for Multi-Paradigm Facial Beauty Prediction. *ICPR*.

Liang, X.; Liu, Z.; Lin, J.; Yang, X.; and Kumada, T. 2024. Uncertainty-oriented Order Learning for Facial Beauty Prediction. arXiv:2409.00603.

Little, A. C. 2014. Facial attractiveness. *WIREs Cognitive Science*, 5(6): 621–634.

Luccioni, A. S.; Akiki, C.; Mitchell, M.; and Jernite, Y. 2023. Stable bias: evaluating societal representations in diffusion models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23. Red Hook, NY, USA: Curran Associates Inc.

Mandal, A.; Leavy, S.; and Little, S. 2023. Multimodal Composite Association Score: Measuring Gender Bias in Generative Multimodal Models. arXiv:2304.13855.

Merino, M.; Tornero-Aguilera, J. F.; Rubio-Zarapuz, A.; Villanueva-Tobaldo, C. V.; Martín-Rodríguez, A.; and Clemente-Suárez, V. J. 2024. Body Perceptions and Psychological Well-Being: A Review of the Impact of Social Media and Physical Measurements on Self-Esteem and Mental Health with a Focus on Body Image Satisfaction and Its Relationship with Cultural and Gender Factors. *Healthcare*, 12(14).

Oppenlaender, J.; Silvennoinen, J.; Paananen, V.; and Visuri, A. 2023. Perceptions and Realities of Text-to-Image Generation. In *Proceedings of the 26th International Academic Mindtrek Conference*, Mindtrek '23, 279–288. New York, NY, USA: Association for Computing Machinery. ISBN 9798400708749.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*.

Rhodes, G.; Yoshikawa, S.; Clark, A.; Lee, K.; McKay, R.; and Akamatsu, S. 2001. Attractiveness of Facial Averageness and Symmetry in Non-Western Cultures: In Search of Biologically Based Standards of Beauty. *Perception*, 30(5): 611–625. PMID: 11430245.

Sattigeri, P.; Hoffman, S. C.; Chenthamarakshan, V.; and Varshney, K. R. 2019. Fairness GAN: Generating datasets with fairness properties using a generative adversarial network. *IBM Journal of Research and Development*, 63(4/5): 3:1–3:9.

Swami, V.; and Tovée, M. J. 2005. Female physical attractiveness in Britain and Malaysia: A cross-cultural study. *Body Image*, 2(2): 115–128.

Targ, S.; Almeida, D.; and Lyman, K. 2016. Resnet in Resnet: Generalizing Residual Architectures. arXiv:1603.08029.

Tiggemann, M.; and Zaccardo, M. 2015. "Exercise to be fit, not skinny": The effect of fitspiration imagery on women's body image. *Body Image*, 15: 61–67.

Wan, Y.; Subramonian, A.; Ovalle, A.; Lin, Z.; Suvarna, A.; Chance, C.; Bansal, H.; Pattichis, R.; and Chang, K.-W. 2024. Survey of Bias In Text-to-Image Generation: Definition, Evaluation, and Mitigation. arXiv:2404.01030.

Wu, Y.; Nakashima, Y.; and Garcia, N. 2025. Revealing Gender Bias from Prompt to Image in Stable Diffusion. *Journal of Imaging*, 11(2).

Xie, D.; Liang, L.; Jin, L.; Xu, J.; and Li, M. 2015. SCUT-FBP: A Benchmark Dataset for Facial Beauty Perception. In *2015 IEEE International Conference on Systems, Man, and Cybernetics*, 1821–1826.

Zhan, J.; Liu, M.; Garrod, O. G.; Daube, C.; Ince, R. A.; Jack, R. E.; and Schyns, P. G. 2021. Modeling individual preferences reveals that face beauty is not universally perceived across cultures. *Current Biology*, 31(10): 2243–2252.e6.

Reproducibility Checklist

1. General Paper Structure

1.1. Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes/partial/no/NA) NA

1.2. Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes/no) yes

1.3. Provides well-marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper (yes/no) yes

2. Theoretical Contributions

2.1. Does this paper make theoretical contributions? (yes/no) no

If yes, please address the following points:

2.2. All assumptions and restrictions are stated clearly and formally (yes/partial/no) Type your response here

2.3. All novel claims are stated formally (e.g., in theorem statements) (yes/partial/no) Type your response here

2.4. Proofs of all novel claims are included (yes/partial/no) Type your response here

2.5. Proof sketches or intuitions are given for complex and/or novel results (yes/partial/no) Type your response here

2.6. Appropriate citations to theoretical tools used are given (yes/partial/no) Type your response here

2.7. All theoretical claims are demonstrated empirically to hold (yes/partial/no/NA) Type your response here

2.8. All experimental code used to eliminate or disprove claims is included (yes/no/NA) Type your response here

3. Dataset Usage

3.1. Does this paper rely on one or more datasets? (yes/no) yes

If yes, please address the following points:

3.2. A motivation is given for why the experiments are conducted on the selected datasets (yes/partial/no/NA) yes

3.3. All novel datasets introduced in this paper are included in a data appendix (yes/partial/no/NA) NA

3.4. All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no/NA) yes

3.5. All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations (yes/no/NA) yes

3.6. All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available (yes/partial/no/NA) yes

3.7. All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing (yes/partial/no/NA) NA

4. Computational Experiments

4.1. Does this paper include computational experiments? (yes/no) yes

If yes, please address the following points:

4.2. This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting (yes/partial/no/NA) NA

4.3. Any code required for pre-processing data is included in the appendix (yes/partial/no) yes

4.4. All source code required for conducting and analyzing the experiments is included in a code appendix (yes/partial/no) yes

4.5. All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no) yes

4.6. All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes/partial/no) yes

4.7. If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results (yes/partial/no/NA) yes

4.8. This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks (yes/partial/no) yes

4.9. This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics (yes/partial/no) yes

4.10. This paper states the number of algorithm runs used to compute each reported result (yes/no) yes

4.11. Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information (yes/no) yes

4.12. The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank) (yes/partial/no) no

4.13. This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments (yes/partial/no/NA) NA