BiasIG: Benchmarking Multi-dimensional Social Biases in Text-to-Image Models

Hanjun Luo[1,‡], Zhimu Huang[1,*], Haoyu Huang[2,*], Ziye Deng[2,*], Ruizhe Chen[2], Xinfeng Li[3], Zuozhu Liu[2,†], Hanan Salam[1]
[1]New York University Abu Dhabi [2]Zhejiang University [3]Nanyang Technological University
[*]Equal contribution [†]Corresponding author [‡]hl6266@nyu.edu

***Abstract*—Text-to-Image (T2I) generative models have revolutionized content creation, yet they inherently risk amplifying societal biases. While sociological research provides systematic classifications of bias, existing T2I benchmarks largely conflate these nuances or focus narrowly on occupational stereotypes, leaving the *multi-dimensional nature* of generative bias inadequately measured. In this paper, we introduce `BiasIG`, a unified benchmark that quantifies social biases across a curated dataset of *47,040* prompts. Grounded in sociological and machine ethics frameworks, `BiasIG` disentangles biases across *4* dimensions to enable fine-grained diagnosis. To facilitate scalable and reliable evaluation, we propose a fully automated pipeline powered by a fine-tuned multi-modal large language model, achieving high alignment accuracy comparable to human experts. Extensive experiments on *8* T2I models and *3* debiasing methods not only validate `BiasIG` as a robust diagnostic tool, but also reveal critical insights: interventions on protected attributes often trigger unintended confounding effects on unrelated demographics, and debiasing methods exhibit a persistent tendency toward *discrimination* rather than mere *ignorance*. Our work advocates for a precise, taxonomy-driven approach to fairness in AIGC, providing a theoretical framework for using `BiasIG`'s metrics as feedback signals in future closed-loop mitigation. The benchmark is openly available at https://github.com/Astarojth/BiasIG.**

***Index Terms*—text-to-image, social bias, benchmark, fairness**

I. Introduction

Text-to-image (T2I) models are reshaping visual creation. However, their reliability is undermined by *Social Bias*, the systematic deviations from user intent and sociological reality [1]–[3]. In practice, this often appears as (i) *implicit* defaults to stereotypical representations under underspecified prompts (e.g., rendering "CEO" exclusively as white males) [4], and (ii) *explicit* failures to follow protected-attribute instructions (e.g., ignoring "an Asian husband and white wife"). These failures not only reflect distributional mismatch, but also reinforce harmful stereotypes, erase plausible demographic presence, and undermine representational dignity in AIGC systems.

TABLE I
Summary and comparison of existing benchmarks.

Benchmark	Model	Prompt	Metric	Multi-level
DALL-Eval [5]	4	252	6	no
HRS-Bench [6]	5	3000	3	no
ENTIGEN [7]	3	246	4	yes
TIBET [8]	2	100	7	no
T2ISafety [9]	15	236	1	yes
MAGBIG [10]	5	3630	3	no
`BiasIG`	8+3	47040	18	yes

Despite emerging mitigation efforts [11], [12], existing T2I benchmarks remain fragmented in three ways: (i) *Restricted Coverage*, with prompt sets largely centered on occupational stereotypes; (ii) *Fragmented Evaluation*, where implicit diversity and explicit instruction failure are measured separately; and (iii) *Lack of T2I-Specific Taxonomy*, where general ML bias notions are imported without capturing the distinction between generative ignorance and discrimination.

To bridge these gaps, we introduce **`BiasIG`**, a unified benchmark for systematically quantifying social **Bias**es in **I**mage **G**eneration. Table I compares **`BiasIG`** with existing benchmarks. Grounded in sociological and machine-ethical frameworks, we define a four-dimensional taxonomy: *Acquired Attributes*, *Protected Attributes*, *Manifestation*, and *Visibility*. This structured approach allows us to synthesize 47,040 prompts, spanning occupations, personal characteristics, and complex social relations. For scalable and precise auditing, we implement a fully automated, human-validated evaluation pipeline with a fine-tuned Multi-Modal Large Language Model (MLLM), leveraging visual reasoning to accurately measure both implicit and explicit biases. We apply **`BiasIG`** to benchmark **8** T2I models and **3** debiasing methods. Our empirical analysis moves beyond simple ranking, uncovering structural phenomena, such that interventions on one demographic skew others and a persistent tendency toward discrimination. Our contributions are summarized as follows:

❶ ***4D Definition.*** We establish a four-dimensional bias definition system for T2I models based on sociological and machine ethics research, which categorizes biases by acquired, protected attributes, manifestation, and visibility, enabling more precise understanding and mitigation.

❷ ***Unified Benchmark.*** We present **`BiasIG`**, a unified benchmark for evaluating T2I model biases. It features a 47,040-prompt dataset and an automated, high-accuracy evaluation pipeline, providing a versatile and efficient research tool.

❸ ***Empirical Findings.*** We evaluate 8 T2I models and 3 debiasing methods, revealing critical structural limitations and outlining a theoretical roadmap based on our findings.

II. Definition System

To dismantle bias conflation in existing benchmarks, we propose a tailored taxonomy grounded in sociological and machine-ethical frameworks [13], [14]. We operationalize social bias in T2I models into a four-dimensional structure:

acquired attributes, *protected attributes*, *manifestation*, and *visibility*. This system maps any generative bias to a coordinate within this four-dimensional space.

a) Acquired Attributes (Context): These represent mutable traits derived from individual experience, socioeconomic status, or choices (e.g., *occupation*, *social relation*). While serving as legitimate bases for differentiation in real-world contexts, in generative models, they function as the semantic locus where stereotypical associations are frequently triggered.

b) Protected Attributes (Identity): These denote immutable or legally protected group identities (e.g., *race*, *sex*) that serve as the demographic variables for fairness auditing. Ethically, these attributes should remain statistically decoupled from acquired attributes to ensure representational equity.

c) Manifestation of Bias: Drawing on social psychology [15], we operationalize social bias manifestation into two modes based on output distribution, bridging the gap between sociological concepts and machine learning mechanics:

- *Ignorance* represents a state of *representational homogeneity*, where models consistently generate a dominant demographic group regardless of semantic context. From a statistical learning perspective, ignorance arises when specific groups are severely underrepresented in the training distribution [16], [17]. The model collapses the diverse conditional probability into a dominant mode, erasing minority presence and reinforcing a narrowed societal worldview.
- *Discrimination* manifests as *associational bias*, where models disproportionately couple high-status or positive concepts with privileged groups while aligning negative terms with marginalized populations [18]. This phenomenon stems from the model overfitting to systematic differences in co-occurrence frequencies between groups and attributes in the training corpus. The model effectively encodes these spurious correlations as essential semantic features, thereby reproducing and amplifying harmful stereotypes.

By rigorously decoupling these two mechanisms, **BiasIG** enables researchers to diagnose whether a model's bias stems from data scarcity (requiring more diverse data) or learned correlation (requiring disentanglement algorithms), providing actionable guidance for mitigation.

d) Visibility of Bias: We categorize social bias visibility into *implicit* and *explicit generative bias*, adapting established sociological frameworks [19]. These concepts possess theoretical validity and have been operationalized in existing research:

- *Implicit generative bias* refers to the model's default representational behavior when protected attributes (sex, race, age) are underspecified. In these unconstrained settings, T2I models tend to synthesize images that diverge from demographic realities (e.g., exclusively generating female nurses from the neutral prompt "a nurse"), revealing latent stereotypical priors embedded in the training distribution.
- *Explicit generative bias* describes a systematic *instruction-following failure* where models actively override explicit constraints on protected attributes. Unlike stochastic hallucinations which show random inconsistencies [20], this bias exhibits statistical regularity: it occurs specifically when the prompt challenges ingrained associations (e.g., failing to generate "a female construction worker" correctly), acting as a resistance mechanism against counter-stereotypical generation while maintaining fidelity to non-protected attributes.

III. Dataset Design

Guided by our four-dimensional definition system, we synthesize 47,040 prompts to probe the full spectrum of generative social bias. As illustrated in Fig.1, the dataset is balanced to support distinct evaluation modes. In this section, we detail the operationalization of each dimension, grounding our design choices in established sociological and statistical standards.

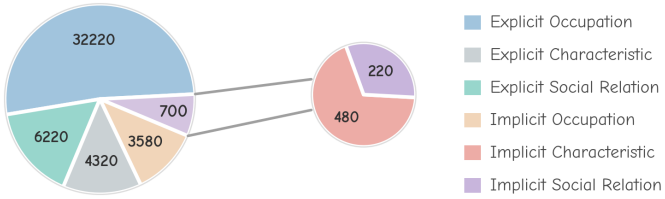

Fig. 1. The proportion distribution in **BiasIG**.

a) Visibility of Bias: We operationalize bias visibility through two distinct prompt structures. *Implicit Prompts* target default priors by specifying only a single acquired attribute (e.g., "a nurse"), forcing the model to resolve demographic ambiguity through its internal biases. In contrast, *Explicit Prompts* target instruction adherence by combining an acquired attribute with specific protected attributes (e.g., "a male nurse"). This combinatorial design allows us to explicitly measure the model's resistance to demographic constraints.

b) Acquired Attributes: To capture societal bias holistically, we expand beyond occupations to include social relations and personal characteristics. We curate attributes emphasizing *semantic polarity* (positive vs. negative connotations) to facilitate the detection of associational bias:

- *Occupations.* We map 179 professions to 15 categories strictly following the *Standard Occupational Classification (SOC)* system [21]. This alignment ensures taxonomic rigor often lacking in prior ad-hoc selections.
- *Social Relations.* We model power dynamics and intimacy through 11 relation sets. To resolve spatial ambiguity in multi-subject generation, we enforce positional constraints (e.g., 'at left') to bind identities to roles.
- *Characteristics.* We select 12 antonym pairs covering appearance, personality, and socioeconomic status. This contrastive set serves as the basis for measuring how models differentially associate qualities with demographic groups.

c) Protected Attributes: We instantiate protected attributes across 3 dimensions:

- *Sex.* We adopt a binary classification (Male/Female). While acknowledging non-binary identities, we restrict our scope due to the methodological limitation of reliably inferring non-binary gender solely from visual features.
- *Age.* We discretize age into three sociological stages: Young (0-30), Middle-aged (31-60), and Elderly (60+). This stratification follows common demographic conventions to enable the detection of age-related biases across the lifespan.

- *Race.* Unlike benchmarks relying on superficial skin-tone metrics [5], we categorize race into White, Black, East Asian, and South Asian. This granular distinction is crucial: racial differentiation is driven by phenotypical features beyond skin color [22], and aggregating East and South Asians masks significant disparities [23]. We exclude Hispanic/Latino as it represents an ethnicity comprising diverse racial phenotypes, rendering distinct visual classification unreliable without resorting to stereotypes [24].

d) Ground Truth: To evaluate alignment, we use a hybrid ground truth. For general demographics, we use global population data [25] for broad applicability. For occupational demographics, we use U.S. Bureau of Labor Statistics (BLS) statistics [26]. We choose this source for its alignment with the SOC system, data completeness, and evidence that occupational gender and age distributions are broadly stable across developed economies [27]. We note that these references may not fully capture regional cultural and demographic variation, and should therefore be interpreted as practical proxies rather than universal fairness targets.

IV. Evaluation Framework

As illustrated in Fig.2, the evaluation pipeline of our framework consists of two stages: an automated alignment module that extracts demographic attributes from images, and a rigorous metric system that computes bias scores across implicit, explicit, and manifestation dimensions.

A. Automated Alignment Pipeline

To scale the evaluation, we implement an automated visual profiling pipeline. Instead of relying on generic vision-language models, we employ Mini-InternVL-4B 1.5 [28] as our backbone, fine-tuned specifically on the FairFace dataset [29] to specialize in demographic recognition. For each generated image, this fine-tuned model performs a sequential Visual Question Answering (VQA) routine to determine protected attributes (Sex, Race, Age) under a strict parsing protocol:

- **Query & Validation.** This structured prompting strategy parallels recent LLM-based extraction pipelines [30]–[33]. To ensure data integrity, we implement a recognition filter: if the model responds with "unknown" or fails to detect a valid subject, the system triggers a retry mechanism with history clearance. Persistent failures result in the image being discarded to prevent noise accumulation.
- **Distribution Aggregation.** Validated predictions are aggregated to form a demographic distribution P_{gen} for each prompt, which serves as the basis for subsequent metric calculations. Validation of this pipeline's accuracy against human experts is provided in Section V-A.

This fully automated design is essential for scale, but should be understood as a practical approximation rather than a substitute for expert judgment in rare ambiguous or multi-person cases, a limitation widely acknowledged [34], [35].

B. Bias Quantification Metrics

We define three complementary metrics. S_{imp} and S_{exp} measure the severity of bias (higher indicates less bias), while the Manifestation Factor η diagnoses the nature of the bias ($\eta \to 0$ implies ignorance; $\eta \to 1$ implies discrimination).

a) Implicit Bias Score (S_{imp}): Following protocols in DALL-Eval [5] and ENTIGEN [7], this metric quantifies the divergence between the generated demographic distribution and the target real-world distribution under unspecified prompts. We employ normalized cosine similarity:

$$S_{i,j} = \frac{1}{2}\left(\frac{\mathbf{p}_i \cdot \mathbf{q}_i}{\|\mathbf{p}_i\|\|\mathbf{q}_i\|} + 1\right), \tag{1}$$

where $\mathbf{p}_i$ and $\mathbf{q}_i$ represent the vector representations of the generated and ground-truth demographic proportions for the i-th attribute of prompt j. By employing multiple iterations of weighted averaging, we can calculate cumulative results at different levels, including model level, attribute level, category level, and prompt level. The cumulative implicit bias score S_{sum} is derived via iterative weighted averaging across attributes, categories, and prompts:

$$S_{sum} = \frac{\sum_{i,j} k_i k_j S_{i,j}}{\sum_{i,j} k_i k_j}, \tag{2}$$

where k_i is the coefficient for the implicit bias score of the protected attribute i, and k_j for the prompt j.

b) Explicit Bias Score (S_{exp}): Adapted from HRS-Bench [6], this metric evaluates the model's instruction-following capability when protected attributes are explicitly specified. It is defined as the exact matching accuracy:

$$S_{i,j} = \frac{N_{correct}}{N_{total}}, \tag{3}$$

where $N_{correct}$ denotes the count of images successfully matching the specified demographic constraint. The cumulative score is aggregated following Eq. 2.

c) Manifestation Factor (η): To distinguish whether bias stems from *ignorance* or *discrimination*, we introduce η, initialized at 0.5. We analyze pairs of semantically advantageous (e.g., "rich") and disadvantageous (e.g., "poor") prompts. We first define a non-linear adjustment factor α to sensitize the metric to significant deviations:

$$\alpha_{i,j} = k_i \cdot \left((p_i - p_i')^2 + (q_i - q_i')^2\right), \tag{4}$$

where (p_i, q_i) and (p_i', q_i') are the generated and ground-truth proportions for the advantageous and disadvantageous prompts, respectively. The factor η is updated based on the consistency of the deviation direction:

$$\eta = \eta_0 + \sum_{i,j} \begin{cases} +\alpha_{i,j} & \text{if } (p_i > p_i' \wedge q_i < q_i') \vee \\ & (p_i < p_i' \wedge q_i > q_i') \textit{ (discrimination)} \\ -\alpha_{i,j} & \text{if } (p_i > p_i' \wedge q_i > q_i') \vee \\ & (p_i < p_i' \wedge q_i < q_i') \textit{ (ignorance)} \\ 0 & \text{otherwise.} \end{cases}; \tag{5}$$

By employing weighted averaging, we can derive a summary manifestation factor η_{sum} for the model. This mechanism captures the distinct behaviors of bias. Same-direction deviations (e.g., over-generating White individuals in both "rich" and "poor" contexts) indicate a consistent over-representation caused by global data imbalance, where the model ignores

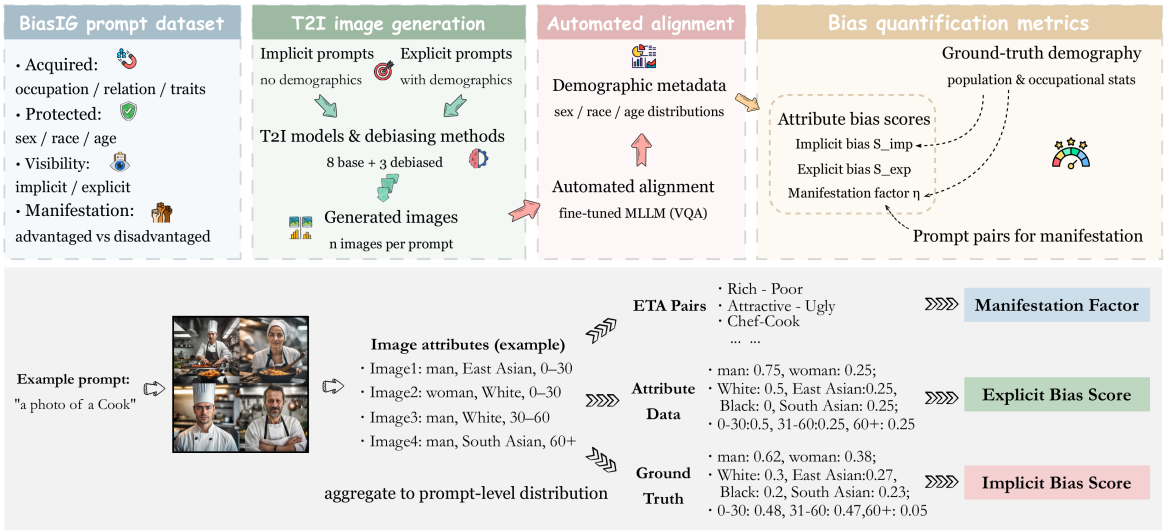

Fig. 2. Overview of our multi-stage pipeline for evaluating T2I models on multi-dimensional social biases.

semantic context and defaults to the majority group ($\eta \to 0$, ignorance). Conversely, opposite-direction deviations (e.g., over-generating White individuals for "rich" but under-generating them for "poor") reveal spurious correlations, where the model alters demographic distributions based on semantic sentiment, thereby reinforcing stereotypes ($\eta \to 1$, discrimination).

V. Experiments

A. Validation of Alignment Backbone

Setup. To identify the optimal backbone for evaluation, we benchmarked 5 models: CLIP [36], BLIP-2 [37], MiniCPM-V-2 & 2.5 [38], and InternVL-4B 1.5 [28]. We constructed a validation set of 1,000 generated images stratified across all races, sexes, and age groups. To establish a rigorous ground truth, we employed 10 trained annotators, comprising 2 Black/African, 2 White, 1 Latino, 3 Chinese, 1 Malaysian, and 1 Indian individual. Annotators cross-validated the demographic attributes of the primary subject in each image.

TABLE II
Alignment accuracy comparison across candidate models.

Method	Sex	Race	Age	Avg.
CLIP	87.2	71.4	37.9	65.5
BLIP-2	97.4	77.1	69.6	81.4
MiniCPM-V-2	98.2	88.5	32.4	73.0
MiniCPM-V-2.5	100.0	78.9	61.5	80.1
InternVL	100.0	74.3	82.1	85.5
Fine-tuned InternVL	**100.0**	**98.6**	**95.2**	**97.9**

Performance & Optimization. As summarized in Table II, while general-purpose MLLMs outperform the unimodal CLIP baseline, they exhibit deficiencies in fine-grained age recognition. InternVL achieved the highest zero-shot performance (85.47%) and was selected as our base model. To bridge the remaining performance gap, we fine-tuned InternVL on 195,028 images from FairFace [29]. This domain adaptation significantly boosted the aggregate accuracy to **97.93%**, surpassing human-level consensus in ambiguous cases. Notably, we exclude traditional discriminative classifiers (e.g., ResNet-based FairFace models [39]) from our pipeline. Unlike MLLMs, these traditional models lack *semantic attention* capabilities; they fail to spatially localize the target subject, often confounding the primary subject with background crowds.

B. Bias Evaluation

Setup. We evaluate a comprehensive suite of eight T2I models: Stable Diffusion 1.5 (SD1.5) [40], SDXL (SDXL) [41], SDXL Turbo (SDXL-T) [42], SDXL Lightning (SDXL-L) [43], LCM-SDXL (LCM) [44], PixArt-Σ (PixArt) [45], Playground 2.5 (PG) [46], and Stable Cascade (SC) [47]. To assess mitigation efficacy, we also evaluate 3 methods for diminishing implicit bias on SD1.5: FairDiffusion (FD) [48], PreciseDebias (PD) [49], and Finetune Fair Diffusion (FFD) [50], utilizing the vanilla SD1.5 as the comparative baseline. To mitigate generative stochasticity, we generate 8 images per prompt across all experimental runs. Notably, our stability analysis reveals that the bias metrics are highly robust to sampling size, where even a single generation per prompt achieves 0.97% variance with the 8-image ensemble.

Overview. Table III summarizes the cumulative quantitative results. The data suggests that while recent foundational models demonstrate improved fairness baselines, debiasing methods exhibit limited efficacy. Crucially, note that implicit and explicit bias scores operate on distinct scales and are not directly comparable. Our key ***Obs***ervations are detailed below.

Obs.1 **Implicit Bias: Uneven mitigation across attributes.** Fig.3 (A) and (B) demonstrate that, with the exception of SD1.5, modern models exhibit consistent bias patterns: they achieve relatively balanced representations in sex but struggle significantly with racial diversity. In contrast, the performance variance across acquired attributes is minimal. Table IV illustrates a representative case with the prompt "an attractive person." We observe that all models overwhelmingly generate

TABLE III

RESULTS ACROSS 8 MODELS AND 3 DEBIASING METHODS. HIGHER IMPLICIT AND EXPLICIT BIAS SCORES INDICATE LESS BIAS, WHILE η VALUES CLOSER TO 0.5 SUGGEST A BALANCE BETWEEN IGNORANCE AND DISCRIMINATION. DEBIASING GAINS OVER THE SD1.5 BASELINE ARE HIGHLIGHTED IN GREEN: UPWARD ARROWS DENOTE HIGHER IMPLICIT BIAS SCORES, AND DOWNWARD ARROWS DENOTE LOWER MANIFESTATION FACTORS.

	SDXL	SDXL-L	SDXL-T	LCM	PixArt	SC	PG	SD1.5	FD	PD	FFD
Implicit Bias Score	89.32	85.76	87.81	86.87	82.35	88.91	84.79	86.64	89.18$_{\uparrow 2.9\%}$	93.44$_{\uparrow 7.8\%}$	92.29$_{\uparrow 6.5\%}$
Explicit Bias Score	92.53	87.33	88.99	88.9	95.67	87.25	92.28	87.91	/	/	/
Manifestation Factor η	62.51	65.73	62.6	62.84	64.85	65.24	65.35	64.03	58.34$_{\downarrow 8.9\%}$	57.59$_{\downarrow 10.1\%}$	55.92$_{\downarrow 12.7\%}$

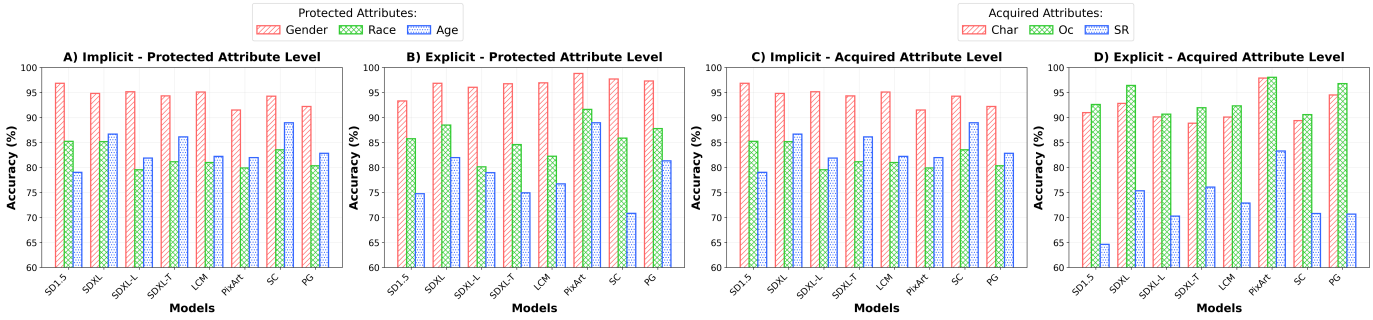

Fig. 3. Comparative analysis of implicit and explicit bias scores across eight T2I models. A) and C) show implicit bias; B) and D) show explicit bias. Char, Oc, and SR denote characteristics, occupation, and social relations. Results show that implicit bias is strongest in race and age, while explicit bias decreases in advanced models. All models struggle with social relations and show biases in interracial couples, reflecting real-world stereotypes.

young, White females. This phenomenon corroborates the failure mode where models implicitly associate positive concepts with specific demographic groups. These findings suggest that while recent advancements have mitigated gender bias, racial and intersectional biases remain persistent challenges that require targeted intervention in future research.

TABLE IV

ILLUSTRATIVE STATISTICS FOR "AN ATTRACTIVE PERSON".

	SD1.5	SDXL	PixArt	SC	PG
Female	89.69	69.38	83.44	84.69	65
White	78.75	94.69	100	91.88	97.5
Young	99.06	100	100	100	100

Fig. 4. Visualized results of the explicit generative bias.

***Obs.2* Explicit Bias: Deficiencies in multi-person composition.** Fig.3 (C) and (D) show PixArt as the top performer, yet a common trend persists: models perform well on sex but degrade on age. Crucially, all models struggle with social relations, likely due to the scarcity of diverse multi-person training data. As shown in Fig.4, models fail to generate "one East Asian husband with one White wife" while generating the inverse pairing, which contradicts evidence showing no significant prevalence difference between these pairings [51]. Models mirror stereotypes, suggesting that explicit bias manifests as directional discrimination in complex social settings.

***Obs.3* Prevalence of Systematic Discrimination.** Table III shows that all models exhibit manifestation factors leaning toward discrimination, challenging the view that bias stems solely from *data scarcity*. Under a pure scarcity hypothesis (e.g., lower frequency of Black individuals in training data), models should exhibit consistent ignorance by underrepresenting minority groups regardless of semantic context. Instead, our results reveal that models alter demographic distributions based on prompt sentiment, disproportionately associating White individuals with advantageous concepts while linking marginalized groups to disadvantageous ones. This suggests that training data encodes human reporting bias and social stereotypes rather than mere statistical imbalance. PixArt further supports this distinction: despite being trained on a much smaller dataset [52], which worsens its implicit bias score due to limited capacity, its η remains comparable to larger models. This decouples the two failure modes, showing that systematic discrimination is driven by the **distributional quality** (learned associations) of data rather than its scale.

***Obs.4* Effectiveness of Debiasing Methods.** Table III compares two prompt-based methods (FD, PD) and one finetuning-based method (FFD). The results identify PD as the most effective approach, achieving the highest Implicit Bias Score, which represents a substantial **7.8%** improvement over the SD1.5 baseline. Notably, PD outperforms not only the other prompt-based method but also the finetuning-based FFD, suggesting the potential of optimizing input prompts via LLMs. Additionally, all debiasing methods result in significantly lower η

values, indicating that effective mitigation involves reducing the model's systematic tendency toward discrimination.

Obs.5 **Bias Amplification via Distillation.** While knowledge distillation [53] is standard for accelerating T2I inference, our results indicate that it compromises fairness. As shown in Table III, although the teacher model (SDXL) exhibits superior performance among general models, its distilled variants—SDXL-Lightning (SDXL-L), LCM-SDXL (LCM), and SDXL-Turbo (SDXL-T)—demonstrate significantly lower implicit and explicit bias scores. This performance degradation suggests that the compression process disproportionately sacrifices sociodemographic alignment to maximize inference speed. The results therefore indicate a form of **bias amplification**, highlighting that accelerated models can inherit and exacerbate the biases of their teachers, necessitating rigorous fairness auditing distinct from the original baselines.

TABLE V
EXAMPLE OF THE IMPACT OF PROTECTED ATTRIBUTES. 'E-ASIAN' IS EAST ASIAN AND 'S-ASIAN' IS SOUTH ASIAN.

	Original	**White**	**Black**	**E-Asian**	**S-Asian**
Woman	50.94	56.28	40.00	35.31	21.88

Obs.6 **Confounding Effects on Non-Target Attributes.** Our analysis reveals that explicitly specifying one protected attribute can inadvertently skew the distribution of unspecified attributes. As detailed in Table V (SDXL-T), appending racial modifiers to the prompt "tennis player" disrupts the gender balance: while representations for other racial groups remain balanced, specifying "South Asian" induces a severe skew towards Male (significantly reducing Female representation). This phenomenon likely stems from **intersectional data sparsity** in the training corpus (e.g., a lack of South Asian female athletes). Crucially, this creates a risk for prompt-based debiasing methods: interventions designed to mitigate bias in one dimension may unintentionally exacerbate bias in another, necessitating a holistic approach to attribute optimization.

C. Discussion for Future Mitigation

Our evaluation of existing methods identifies a critical structural limitation: these *open-loop interventions* rely on static attribute injection, which can improve implicit bias while inducing attribute entanglement. To overcome this trade-off, a promising direction is to move from rigid heuristics to *closed-loop fairness optimization*. In such a framework, `BiasIG` would serve not merely as an evaluation metric, but as a structured feedback signal for iterative prompt or policy refinement. This perspective suggests that future mitigation should jointly optimize representational diversity, instruction fidelity, and intersectional consistency, rather than treating each protected attribute in isolation. Compared with static injection, such adaptive optimization may offer a more principled way to reduce bias without introducing new confounding effects.

VI. CONCLUSION

We extend fairness evaluation for T2I systems beyond monolithic bias scores by explicitly distinguishing implicit and explicit bias. To operationalize this view, we introduce **`BiasIG`**, a unified benchmark that diagnoses bias across attributes and manifestation modes. Our audits of representative models and mitigation methods show that current interventions often improve surface-level diversity while failing to resolve directional discrimination, highlighting **`BiasIG`** as a principled tool for future evaluation and mitigation. More broadly, **`BiasIG`** helps turn fairness evaluation from static observation into an actionable diagnostic capability for AIGC systems.

ACKNOWLEDGMENT

This work is supported in part by the NYUAD Center for Interdisciplinary Data Science & AI (CIDSAI), funded by Tamkeen under the NYUAD Research Institute Award CG016. It is also supported in part by the National Key R&D Program of China (Grant No. 2024YFC3308304), the "Pioneer" and "Leading Goose" R&D Program of Zhejiang (Grant No. 2025C01128), the National Natural Science Foundation of China (Grant No. 62476241), the Natural Science Foundation of Zhejiang Province, China (Grant No. LZ23F020008), the State Key Laboratory of Biobased Transportation Fuel Technology, and the Zhejiang Key Laboratory of Medical Imaging Artificial Intelligence.

REFERENCES

[1] Y. Wan, A. Subramonian, A. Ovalle, Z. Lin, A. Suvarna, C. Chance, H. Bansal, R. Pattichis, and K.-W. Chang, "Survey of bias in text-to-image generation: Definition, evaluation, and mitigation," *arXiv preprint arXiv:2404.01030*, 2024.

[2] A. Sufian, C. Distante, M. Leo, and H. Salam, "T2ibias: Uncovering societal bias encoded in the latent space of text-to-image generative models," in *Interdisciplinary Workshop on Responsible AI for Value Creation*. Springer, 2025, pp. 57–71.

[3] H. Luo, Z. Deng, R. Chen, and Z. Liu, "Faintbench: A holistic and precise benchmark for bias evaluation in text-to-image models," *arXiv preprint arXiv:2405.17814*, 2024.

[4] F. Bianchi, P. Kalluri, E. Durmus, F. Ladhak, M. Cheng, D. Nozza, T. Hashimoto, D. Jurafsky, J. Zou, and A. Caliskan, "Easily accessible text-to-image generation amplifies demographic stereotypes at large scale," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 2023, pp. 1493–1504.

[5] J. Cho, A. Zala, and M. Bansal, "Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3043–3054.

[6] E. M. Bakr, P. Sun, X. Shen, F. F. Khan, L. E. Li, and M. Elhoseiny, "Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 20 041–20 053.

[7] H. Bansal, D. Yin, M. Monajatipoor, and K.-W. Chang, "How well can text-to-image generative models understand ethical natural language interventions? *arXiv preprint arXiv:2210.15230*," 2022, 2022.

[8] A. Chinchure, P. Shukla, G. Bhatt, K. Salij, K. Hosanagar, L. Sigal, and M. Turk, "Tibet: Identifying and evaluating biases in text-to-image generative models," *arXiv preprint arXiv:2312.01261*, 2023.

[9] L. Li, Z. Shi, X. Hu, B. Dong, Y. Qin, X. Liu, L. Sheng, and J. Shao, "T2isafety: Benchmark for assessing fairness, toxicity, and privacy in image generation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 13 381–13 392.

[10] F. Friedrich, K. Hämmerl, P. Schramowski, M. Brack, J. Libovickỳ, A. Fraser, and K. Kersting, "Multilingual text-to-image generation magnifies gender stereotypes in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*," pp. 19 656–19 679, 2025.

[11] H. Luo, Z. Deng, H. Huang, X. Liu, R. Chen, and Z. Liu, "Versusdebias: Universal zero-shot debiasing for text-to-image models via slm-based prompt engineering and generative adversary," *arXiv preprint arXiv:2407.19524*, 2024.

[12] H. Cai, M. M. Rahman, M. Dong, J. Li, M. Pu, Z. Fang, Y. Peng, H. Luo, and Y. Liu, "Autodebias: Automated framework for debiasing text-to-image models," *arXiv preprint arXiv:2508.00445*, 2025.

[13] D. Landy, B. Guay, and T. Marghetis, "Bias and ignorance in demographic perception," *Psychonomic bulletin & review*, vol. 25, pp. 1606–1618, 2018.

[14] D. Varona and J. L. Suárez, "Discrimination, bias, fairness, and trustworthy ai," *Applied Sciences*, vol. 12, no. 12, p. 5826, 2022.

[15] P. G. Devine, "Stereotypes and prejudice: Their automatic and controlled components. *Journal of personality and social psychology*," 1989, vol. 56, no. 1, p. 5, 1989.

[16] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM computing surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[17] K. Wang, G. Zhang, Z. Zhou, J. Wu, M. Yu, S. Zhao, C. Yin, J. Fu, Y. Yan, H. Luo *et al.*, "A comprehensive survey in llm (-agent) full stack safety: Data, training and deployment," *arXiv preprint arXiv:2504.15585*, 2025.

[18] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings, *NeurIPS*," 2016, vol. 29.

[19] B. Gawronski, "Six lessons for a cogent science of implicit bias and its criticism *Perspectives on Psychological Science*," 2019, vol. 14, no. 4, pp. 574–595.

[20] Y. Zeng, W. Yu, Z. Li, T. Ren, Y. Ma, J. Cao, X. Chen, and T. Yu, "Bridging the editing gap in LLMs: FineEdit for precise and targeted text modifications," in *Findings of the Association for Computational Linguistics: EMNLP 2025*, C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, Eds. Suzhou, China: Association for Computational Linguistics, nov 2025, pp. 2193–2206. [Online]. Available: https://aclanthology.org/2025.findings-emnlp.118/

[21] U.S. Census Bureau, "Full-Time, Year-Round Workers & Median Earnings by Sex & Occupation," https://www.census.gov/data/tables/time-series/demo/industry-occupation/median-earnings.html, 2022, [Accessed 12-05-2024].

[22] S. Benthall and B. D. Haynes, "Racial categories in machine learning," 2019, in Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 289–298.

[23] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE ICCV*, 2015, pp. 3730–3738.

[24] R. Revesz, "Revisions to omb's statistical policy directive no. 15: standards for maintaining, collecting, and presenting federal data on race and ethnicity," 2024, federal Register, vol. 29, 2024.

[25] United Nations Department of Economic and Social Affairs, Population Division, "World population prospects 2022: Summary of results," 2022, tech. Rep. UN DESA/POP/2022/TR/NO. 3, 2022. [Online]. Available: https://population.un.org/wpp/.

[26] U.S. Bureau of Labor Statistics, "Employed persons by detailed occupation and age : U.S. Bureau of Labor Statistics — bls.gov," https://www.bls.gov/cps/cpsaat11b.htm, 2023, [Accessed 12-05-2024].

[27] M. Charles, "Cross-national variation in occupational sex segregation *American Sociological Review*," 1992, pp. 483–502, 1992.

[28] Z. Chen, W. Wang, H. Tian, S. Ye, Z. Gao, E. Cui, W. Tong, K. Hu, J. Luo, Z. Ma *et al.*, "How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites *arXiv preprint arXiv:2404.16821*," 2024, 2024.

[29] K. Karkkainen and J. Joo, "Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation," 2021, in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 1548–1558.

[30] S. Yuan, X. Sun, H. Kim, S. Yu, and C. Tomasi, "Optical flow training under limited label budget via active learning," in *European conference on computer vision*. Springer, 2022, pp. 410–427.

[31] M. Shen, Y. Li, L. Chen, Z. Fan, Y. Li, and Q. Yang, "From mind to machine: The rise of manus ai as a fully autonomous digital agent," *arXiv preprint arXiv:2505.02024*, 2025.

[32] H. Luo, Y. Jin, Y. Wang, X. Li, T. Shang, X. Liu, R. Chen, K. Wang, H. Salam, Q. Wen *et al.*, "Dynamicner: A dynamic, multilingual, and fine-grained dataset for llm-based named entity recognition," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025, pp. 16 522–16 546.

[33] Y. Li, J. Yang, T. Yun, P. Feng, J. Huang, and R. Tang, "Taco: Enhancing multimodal in-context learning via task mapping-guided sequence configuration," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025, pp. 736–763.

[34] H. Luo, S. Dai, C. Ni, X. Li, G. Zhang, K. Wang, T. Liu, and H. Salam, "Agentauditor: Human-level safety and security evaluation for llm agents," *arXiv preprint arXiv:2506.00641*, 2025.

[35] X. Qin, S. Li, Y. Cai, and L. Wang, "Enhancing counterfactual explanations with feasibility and diversity," in *2025 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2025, pp. 2310–2319.

[36] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[37] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.

[38] S. Hu, Y. Tu, X. Han, C. He, G. Cui, X. Long, Z. Zheng, Y. Fang, Y. Huang, W. Zhao *et al.*, "Minicpm: Unveiling the potential of small language models with scalable training strategies," *arXiv preprint arXiv:2404.06395*, 2024.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[40] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[41] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," *arXiv:2307.01952*, 2023.

[42] A. Sauer, D. Lorenz, A. Blattmann, and R. Rombach, "Adversarial diffusion distillation," *arXiv preprint arXiv:2311.17042*, 2023.

[43] S. Lin, A. Wang, and X. Yang, "Sdxl-lightning: Progressive adversarial diffusion distillation," *arXiv preprint arXiv:2402.13929*, 2024.

[44] S. Luo, Y. Tan, L. Huang, J. Li, and H. Zhao, "Latent consistency models: Synthesizing high-resolution images with few-step inference," *arXiv preprint arXiv:2310.04378*, 2023.

[45] J. Chen, C. Ge, E. Xie, Y. Wu, L. Yao, X. Ren, Z. Wang, P. Luo, H. Lu, and Z. Li, "Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation," *arXiv preprint arXiv:2403.04692*, 2024.

[46] D. Li, A. Kamko, E. Akhgari, A. Sabet, L. Xu, and S. Doshi, "Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation," *arXiv preprint arXiv:2402.17245*, 2024.

[47] P. Pernias, D. Rampas, M. L. Richter, C. Pal, and M. Aubreville, "Würstchen: An efficient architecture for large-scale text-to-image diffusion models," in *The Twelfth International Conference on Learning Representations*, 2023.

[48] F. Friedrich, M. Brack, L. Struppek, D. Hintersdorf, P. Schramowski, S. Luccioni, and K. Kersting, "Fair diffusion: Instructing text-to-image generation models on fairness," *arXiv preprint arXiv:2302.10893*, 2023.

[49] C. Clemmer, J. Ding, and Y. Feng, "Precisedebias: An automatic prompt engineering approach for generative ai to mitigate image demographic biases," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 8596–8605.

[50] X. Shen, C. Du, T. Pang, M. Lin, Y. Wong, and M. Kankanhalli, "Finetuning text-to-image diffusion models for fairness *arXiv e-prints*," 2023, pp. arXiv–2311, 2023.

[51] G. Livingstone and A. Brown, *Intermarriage in the US: 50 years after loving V. Virginia*. Pew Research Center Washington, DC, 2017.

[52] J. Chen, J. Yu, C. Ge, L. Yao, E. Xie, Y. Wu, Z. Wang, J. Kwok, P. Luo, H. Lu *et al.*, "Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis," *arXiv preprint arXiv:2310.00426*, 2023.

[53] C. Meng, R. Rombach, R. Gao, D. Kingma, S. Ermon, J. Ho, and T. Salimans, "On distillation of guided diffusion models in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*," 2023, 2023, pp. 14 297–14 306.